Lecture Notes in Information Technology

Volume 10

Lecture Notes in Information Technology

2012 International Conference on Affective Computing and Intelligent Interaction (ICACII 2012)

Taipei, Taiwan, February 27-28, 2012

Edited by

JiaLuoluo

INFORMATION ENGINEERING RESEARCH INSTITUTE, USA

Copyright © 2012 Information Engineering Research Institute, USA

All rights reserved. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the Information Engineering Research Institute.

Information Engineering Research Institute
100 Continental Dr, Newark, DELAWARE 19713, Unite State, USA
http://www.ier-institute.org

ISBN: 978-1-61275-004-0
Lecture Notes in Information Technology Vol.10
ISSN: 2070-1918

Distributed worldwide by
Information Engineering Research Institute
100 Continental Dr, Newark, DELAWARE 19713, Unite State, USA

E-mail: admin@ier-institute.org

Message from the ICACII 2012 Chair

2012 International Conference on Affective Computing and Intelligent Interaction (ICACII 2012) will be held in Taipei, Taiwan, February 27-28, 2012.

Affective computing is the study and development of systems and devices that can recognize, interpret, process, and simulate human affects. It is an interdisciplinary field spanning computer sciences, psychology, and cognitive science. While the origins of the field may be traced as far back as to early philosophical enquiries into emotion, the more modern branch of computer science originated with Rosalind Picard's 1995 paper on affective computing. A motivation for the research is the ability to simulate empathy. The machine should interpret the emotional state of humans and adapt its behavior to them, giving an appropriate response for those emotions.

Intelligent Interaction research theme seeks to enhance human-machine interface design through the optimization of state-of-the-art technology development and engineering of multimodal interface design concepts. II also seeks to explicate the mechanisms of human perception, cognition, and action that are relevant to industrial, military, and consumer products. Projects in HCII involve the close collaboration of computer scientists, electrical engineers, neuroscientists, linguists and others in pursuit of knowledge relevant to the design of interfaces for human-computer systems.

ICACII 2012 will be the most comprehensive conference focused on the various aspects of advances in Affective Computing and Intelligent Interaction. The conference is intended to bring together the researchers and scientists working in different aspects of affective computing and intelligent interaction. In addition to the contributed papers, the conference committee has invited papers by active researchers from various countries in relevant topic areas. Internationally known experts are invited to give keynote speeches.

Welcome to ICACII 2012; welcome to Taipei. Taipei is the capital of Taiwan. It is in the northern part of the island in a basin between the Yangming Mountains and the Central Mountains. It is, with 2.6 million inhabitants, the fourth largest administrative area of Taiwan, after New Taipei, Kaohsiung and Taichung. However, the Greater Taipei metropolitan area, which encompasses the central Taipei City along with the surrounding New Taipei City and Keelung, represents the largest urban cluster in Taiwan with nearly 7 million people. Taipei serves as the island's financial, cultural, and governmental center.

It is hoped that the present book will be useful to scientists, both specialists and graduate students.

I express my deep gratitude to all contributors of this book who worked very hard to make this project successful. If our efforts came to the present result, it was not without ideas, encouragement and support from Publishing Manager that must be gratefully acknowledged.

JiaLuoluo, Kinmen County Ningxiang University Avenue, Taiwan

ICACII 2012 Organizing Committee

Honorary Chair and Keynote Speakers

Chin Chen Chang	Feng Chia University, Taiwan
Wei Lee	Melbourne ACM Chapter Chair, Australia

General Chairs

Zhenghong Wu	East China Normal University, China
Minli Dai	Suzhou University, China

Publication Chair

JiaLuoluo	Kinmen County Ningxiang University Avenue, Taiwan

Organizing Chairs

Jia-chin Chang	National Chung Hsing University, Taiwan
Kaikai Yang	National Chiayi University, Taiwan

Program Chair

DehuaiYang	Huazhong Normal University, China

Program Committee

Zhou William	Belize University, Beliza
Anne Wing	Yangon University, Myanmar
Khing Mg Win Fay	University of Guyana, Guyana
Qihai Zhou	Southwestern University of Finance and Economics, China
Zhenghong Wu	East China Normal University, China
Tatsuya Akutsu	ACM NUS Singapore Chapter, Singapore
AijunAn	National University of Singapore, Singapore
Yuanzhi Wang	Anqing Teachers' University, China
YiyiZhouzhou	Azerbaijan State Oil Academy, Azerbaijan

Table of Contents

Volume 10

Detecting and Resolving Constraint Conflicts in Role-Based Access Control

QIU Jiong [1,a], MA Chen-hua [2,b]

[1] Department of Computer Science & Technology, Hangzhou Dianzi University, Hangzhou 310018, China

[2] Engineering & Computer Graphics Institute, Zhejiang University, Hangzhou 310027, China

[a,b] mchma@zju.edu.cn

Keywords: Role-based Access Control; Constraint Conflict; Conflict detection and Resolution

Abstract. The detection and resolution of constraint conflicts in RBAC have been overlooked and remain a significant research challenge. To address these concerns, in this paper, we classify constraint conflicts into two categories: internal constraint conflicts that occur when two or more constraints are deemed incompatible with each other and external constraint conflicts that occur when the configuration of a RBAC system violates the defined constraints, and propose a set of detection rules for these conflicts. Furthermore, we introduce the notions of resolution value and valid resolution value, and show how they are useful in guiding external constraint conflict resolution.

Introduction

Role-based access control (RBAC) is known as the most suitable access control model for enterprise organizations. The importance of the constraints in RBAC has been recognized for a long time[1]. In the past decade, a considerable amount of work [2-5] has been done on RBAC constraints. However, the focus of these researches has been predominantly on the specification of RBAC constraints. Effective conflict detection and resolution methods used to maintain the consistency between constraints have been overlooked and remain a significant research challenge.

Strembeck [6] discussed the conflict checking of separation of duty constraints in RBAC and presented conflict checking methods as implemented in the xoRBAC software component. Moon [7] addressed the issue of conflict detection to maintain the consistency of permission assignment constraints in RBAC. However, the conflict detection and resolution methods used to maintain the consistency between constraints have not been addressed. Janpitak [8] proposes a simple but effective model to solve the problem of the dynamic separation of duties.The conflict of interest can be verified at run time. But the model cannot support role hierarchies in RBAC. To address the problem, we propose in this paper the following approaches:

- We classify constraint conflicts into two categories: internal constraint conflicts and external constraint conflicts. Internal constraint conflicts refer to the conflicts exhibited by two or more incompatible constraints. External constraint conflicts refer to the conflicts exhibited by the configuration of a RBAC system and the constraints defined in the system.
- We give several conflict detection rules for internal constraint conflicts by the definition of

Foundation Item: Project (No. 2009C03015-1) supported by the Large Science and Technology Special Program of Zhejiang Province.
Corresponding author: QIU Jiong, Associate Professor; Tel: +86-13805742886; E-mail: mchma@zju.edu.cn

constraint conflict graph. An effective detection rule for external constraint conflicts is proposed by defining two concepts, RBAC configuration graph and conflict pattern. An approach to external constraint conflict resolution is proposed.

RBAC model

We will base our discussion on RBAC96. In this section, we provide an overview of the central concepts within the model. A role hierarchy is a partial order on roles called the inheritance relation, written as $\geq$, where $r_i \geq r_j$ only if all permissions of r_j are also permissions of r_i. Users are associated with roles using the user–role assignment relation $UA \subseteq U \times R$. If there exists a pair $(u, r) \subseteq UA$, then role r is explicitly assigned to user u. Permissions are associated with roles using the permission-role assignment relation $PA \subseteq P \times R$. If there exists a pair $(p, r) \subseteq PA$, then permission p is explicitly assigned to role r. Constraints are a powerful mechanism for laying out higher-level organizational policy.

The specification of constraints in RBAC

Constraints proposed in RBAC models can be classified into three broad categories:

1. Separation of Duty (SoD) constraints: SoD constraints aim at reducing the risk of fraud by not allowing any individual to have sufficient authority within the system to perpetrate a fraud on his/her own. In this paper, we focus on static SoD constraints. Three varieties of static SoD constraints have been proposed so far:

- Conflicting role constraints: Let CR represent the collection of conflicting role sets, $CR=\{cr_1,cr_2,\ldots,cr_n\}$,where $cr_i\,(i=1,\ldots,n)$ denotes a conflicting role set. Two or more roles of a conflicting role set cannot be assigned to the same user.
- Conflicting permission constraints: Let CP represent the collection of conflicting permission sets, $CP=\{cp_1,cp_2,\ldots,cp_m\}$,where $cp_i\,(i=1,\ldots,m)$ denotes a conflicting permission set. Two or more permissions belonging to a conflicting permission set cannot be assigned to the same role.
- Conflicting user constraints: Let CU represent the collection of conflicting user sets, $CU=\{cu_1,cu_2,\ldots,cu_t\}$,where $cu_i\,(i=1,\ldots,t)$ denotes a conflicting user set. Two conflicting users cannot have roles in the same conflicting role set.

2. Cardinality constraints: A cardinality constraint can be formally defined as (r, n). Where r is the role associated with the constraint; n denotes the numerical limitation for the role.

3. Prerequisite constraints: Prerequisite constraints are defined based on competency and appropriateness whereby a user can be assigned role r_1 only if the user is already a member of role r_2, or a permission p_1 can be assigned to a role only if the role already possesses permission p_2.

Constraint conflict detection and resolution

Definition 1 (Internal Constraint Conflict). Internal constraint conflicts occur when two or more constraints are deemed incompatible with each other. For example, there exists a prerequisite role constraint in which role r_1 is defined as a prerequisite role of role r_2; whereas, there exists another prerequisite role constraint in which role r_2 is defined as a prerequisite role of role r_1. In this case, the two constraints are contradictory and exhibit an internal constraint conflict.

Definition 2 (External Constraint Conflict). External constraint conflicts occur when the configuration of a RBAC system doesn't satisfy the constraints defined in the system. For example, if a new conflicting role set consisting of two roles r_1 and r_2 is created, and there is an existing inheritance relationship between r_1 and r_2 in the role hierarchy, then an external constraint conflict occurs.

Internal constraint conflict detection. Definition 3 (Constraint Conflict Graph). Constraint conflict graph is used by the security administrator to understand internal constraint conflicts easily and to detect internal constraint conflicts effectively. Constraint conflict graph is a multi graph,

denoted as $G(CC)=(V, E)$. Where V is the set of vertices in $G(CC)$, composed of the users, roles and permissions associated with constraints. An edge in E can be denoted as e:(v_1, v_2, *label*, *id*), where v_1 and v_2 represent the source vertex and the target vertex of e respectively; *label* denotes a possible constraint relation between v_1 and v_2; *id* is the identifier of the constraint corresponding to the edge. For an edge e, it may be one of the following cases:

- e:(v_1, v_2, *conflicting*, *id*), where v_1 and v_2 are two members of a conflicting element set (e.g., a conflicting role set, a conflicting permission set, and a conflicting user set), and *id* is the identifier of the SoD constraint associated with the conflicting element set.
- e:(v_1, v_2, *prerequisite*, *id*), where v_1 is the target element of a prerequisite constraint (e.g., the target role of a prerequisite role constraint, and the target permission of a prerequisite permission constraint), v_2 is a member of the prerequisite element set of the constraint (e.g., the prerequisite role set of a prerequisite role constraint), and *id* is the identifier of the prerequisite constraint. For instance, there exist five constraints as defined below.

cr_1: (r_1, r_2, r_3); prc_1: (r_1, r_2); ppc_1: (p_1, p_2); ppc_2: (p_2, p_3); ppc_3: (p_3, p_1)

Where cr_1 is a conflicting role set in which r_1, r_2 and r_3 are defined as conflicting; prc_1 is a prerequisite role constraint in which r_2 is defined as the prerequisite role of r_1; ppc_1, ppc_2 and ppc_3 are three prerequisite permission constraints that define p_2, p_3 and p_1 as the prerequisite permissions of p_1, p_2 and p_3 respectively. The constraint conflict graph constructed for these example constraints is as shown in Figure 2. Internal constraint conflicts can be detected by the construction of $G(CC)$.

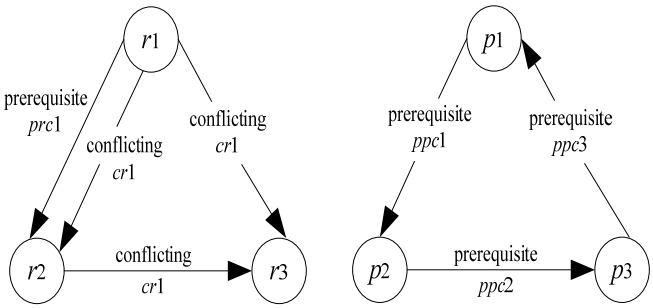


Fig. 1.. Example constraint conflict graph

Rule 1. If there exist two edges between two vertices v_1 to v_2 in $G(CC)$, one of the edge is labeled as conflicting and the other is labeled as prerequisite, then the two constraints corresponding to the two edges are inconsistent and internal constraint conflicts occur.

Rule 2. If there is a loop in $G(CC)$, consisting of a sequence of edges labeled as prerequisite, then the prerequisite constraints corresponding to these edges are inconsistent and internal constraint conflicts occur. Figure 1 shows an example of such conflicting situation. There exists a loop composed of three edges labeled as prerequisite, and the three constraints ppc_1, ppc_2 and ppc_3 are inconsistent.

Rule 3. For two cardinality constraints (r_1, n_1) and (r_2, n_2), if ($r_1 = r_2$) $\wedge$ ($n_1 \neq n_2$), then they are inconsistent and exhibit an internal constraint conflict.

External Constraint Conflict Detection. External constraint conflicts arise when the configuration of a RBAC system doesn't satisfy the constraints defined in the system. For example, if a new conflicting permission constraint is created in which two permissions p_1 and p_2 are defined as conflicting, and there exists a role r that has been assigned both p_1 and p_2 in the current RBAC configuration, then in this case, the current configuration violates the new constraint and external constraint conflicts occur.

Definition 6 (RBAC Configuration Graph). RBAC configuration graph is a directed one, denoted as $G(RCG)=(V, E)$, where V is the set of vertices in $G(RCG)$, composed of all users, roles and permissions in the RBAC system, i.e., $N=U \cup R \cup P$. E is the set of edges in $G(RCG)$ defined

by existing explicit assignment relations in the system. If vertex v_1 is explicitly assigned to vertex v_2 (e.g., a role is explicitly assigned to a user), then an edge e starting from v_1 to v_2 is created, denoted as e:(v_1, v_2), where v_1 and v_2 represent the source vertex and the target vertex of e respectively.

For two vertices v_i and v_j in the graph, if there exists a path starting from v_i to v_j, then there is an assignment relation between v_i and v_j, denoted as *assign*(v_i, v_j). An assignment relation between two vertices may be an explicit assignment relation or an implicit assignment relation, which is a chain of explicit assignment relations. For two vertices v_i and v_j in the graph, if there exist no path starting from v_i to v_j, then there is no assignment relation between v_i and v_j, denoted as $\neg$*assign*(v_i, v_j). An example RBAC configuration graph is as shown in Figure 2.

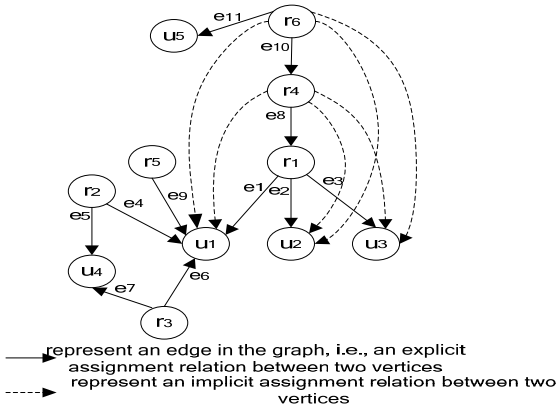


Fig. 2.. Example RBAC configuration graph

There are six roles (r_1, r_2, r_3, r_4, r_5 and r_6), and five users (u_1, u_2, u_3, u_4 and u_5) in the graph. For example, role r_6 is explicitly assigned to role r_4. Note that there are six implicit assignment relations *assign*(r_4, u_1), *assign*(r_4, u_2), *assign*(r_4, u_3), *assign*(r_6, u_1), *assign*(r_6, u_2) and *assign*(r_6, u_3) in the graph, and each of them is a chain of explicit assignment relations. For instance, *assign*(r_6, u_1) is a chain of three explicit assignment relations, *assign*(r_1, u_1), *assign*(r_4, r_1) and *assign*(r_6, r_4), i.e., r_6 is implicitly assigned to u_1 in the presence of the role hierarchy.

Definition 7 (Conflict Pattern). For a given constraint c, if the combination of certain assignment relations *assign*$_1$, …, *assign*$_n$ results in the violation of the constraint, it can be said that (*assign*$_1$, …, *assign*$_n$) is the conflict pattern of c, denoted as: *Conflict_Pattern* (c): (*assign*$_1$, …, *assign*$_n$). For instance, for a conflicting role constraint cr:(r_1, r_2, …, r_n), we can identify the conflict pattern for the constraint as described below.

Conflict_Pattern (cr): {*assign* (r_i, u), *assign* (r_j, u)}, r_i , $r_j \in cr$, $i \neq j$

The conflict pattern states that the assignments of both two conflicting roles r_i and r_j to the same user u can lead to the violation of the constraint. According to the pattern, if there exist two assignment relations in the RBAC configuration graph that match the specified relations defined in the pattern, then an external constraint conflict occurs. The conflict patterns for different types of constraints proposed in RBAC are as shown in Table 1.

Table 1. Conflict patterns for constraints proposed in RBAC

Constraint	Conflict pattern	description
cr:(r_1, r_2, …, r_n), cr is a conflicting role constraint	*Conflict_Pattern*(cr):{*assign* (r_i , u), *assign* (r_j , u)}, r_i , $r_j \in cr$, $i \neq j$	The conflict pattern states that the assignments of both two conflicting roles r_i and r_j to the same user u may lead to the violation of the constraint.
cp:(p_1, p_2, …, p_n), cp is a conflicting permission constraint	*Conflict_Pattern*(cp):{*assign* (p_i , r), *assign* (p_j , r)}, p_i , $p_j \in$ cp, $i \neq j$	The conflict pattern states that the assignments of both two conflicting permissions p_i and p_j to the same role r may lead to the violation of the constraint.

cu:(u_1, u_2, …, u_n), cu is a conflicting user constraint	*Conflict_Pattern*(cu):{*assign* (r, u_i), *assign* (r, u_i)}, u_i, $u_j \in cu$, $i \neq j$, $r \in cr$	The conflict pattern states that the assignments of both two conflicting users u_i and u_j to roles in the same conflicting role set may lead to the violation of the constraint.
c:(r, n), c is a cardinality constraint	*Conflict_Pattern*(c):{*assign*(r, u_1), *assign* (r , u_2),…, *assign*(r, u_t)}, $t > n$	The conflict pattern states that if the number of users owned by role r exceeds the maximum number n, then the cardinality constraint is violated.
c: (tr, *PreRoleSet*), c is a prerequisite role constraint	*Conflict_Pattern*(c):{*assign*(t r, u), ¬ *assign* (r_i , u)}, $r_i \in$ *PreRoleSet*	The conflict pattern states that if role tr is assigned to user u without all roles in *PreRoleSet* have been assigned to u, then the prerequisite role constraint is violated.
c: (tp, *PrePermSet*), c is a prerequisite permission constraint	*Conflict_Pattern*(c):{*assign*(t p, r), ¬ *assign* (p_i , r)}, $p_i \in$ *PrePermSet*	The conflict pattern states that if permission tp is assigned to role r without all permissions in *PrePermSet* have been assigned to r, then the prerequisite permission constraint is violated.

Rule 4 (External Constraint Conflict Detection). In G(RCG), if there exist assignment relations that match the specified relations defined in a conflict pattern, then external constraint conflicts occur. Each external constraint conflict can be denoted as a combination of certain existing assignment relations in G(RCG). For example, if there is a conflicting role constraint in which r_1, r_2 and r_4 in Figure 2 are defined as conflicting, then the analysis of Figure 2 reveals five external constraint conflicts due to the violation of the constraint. Each conflict is a combination of two assignment relations.

$conflict_1$: {*assign* (r_1,u_1), *assign* (r_2,u_1)}; $conflict_2$: {*assign* (r_1,u_1), *assign* (r_4,u_1)}
$conflict_3$: {*assign* (r_2,u_1), *assign* (r_4,u_1)}; $conflict_4$: {*assign* (r_1,u_2), *assign* (r_4,u_2)}
$conflict_5$: {*assign* (r_1,u_3), *assign* (r_4,u_3)}

External Constraint Conflict Resolution. Since each external constraint conflict is a combination of some assignment relations in G(RCG), the removal of at least one of the assignment relations may lead to the resolution of the conflict. For example, $conflict_3$ in Figure 2 can be resolved by the removal or either *assign* (r_2,u_1) or *assign* (r_4,u_1). Since the removal of different edges may have different impact on conflict resolution, we would like to remove the edge that has the greatest impact, such that many conflicts can be resolved at the same time with little effort.

For instance, to resolve $conflict_3$, we can remove e_1, e_4 or e_8. The removal of e_1 results in the resolution of three conflicts $conflict_1$, $conflict_2$ and $conflict_3$. While the removal of e_8 results in the removal of four conflicts $conflict_2$, $conflict_3$, $conflict_4$, and $conflict_5$.

Therefore, in order to resolve conflicts effectively, we must compute the number of conflicts covered by each edge involved in conflict situations.

Definition 8 (Edge Cover Set). For an external constraint conflict $conflict_i$, let *EdgeCoverSet*($conflict_i$) represent the set of edges covered by the conflict, composed of all edges in G(RCG) that are associated with the assignment relations in the conflict. For example, the edge cover sets of the five conflicts shown above are listed as follows:

EdgeCoverSet($conflict_1$): (e_1,e_4); *EdgeCoverSet*($conflict_2$): (e_1,e_8);
EdgeCoverSet($conflict_3$): (e_1,e_4,e_8); *EdgeCoverSet*($conflict_4$): (e_2,e_8);
EdgeCoverSet($conflict_5$): (e_3,e_8)

Definition 9 (Resolution Value). For an edge e in edge cover sets, we can identify the number of conflicts resolved by the removal of the edge, which is defined as the resolution value of the edge, denoted as *ResolveValue*(e).

ResolveValue(e) = $\sum$*EdgeCoverSet*($conflict_i$), $e \in$ *EdgeCoverSet*($conflict_i$), that is, *ResolveValue*(e) is the number of edge cover sets of which edge e is a member. For example, the resolution value of each edge in the above edge cover sets is calculated and shown as below.

ResolveValue(e_1) = 3; *ResolveValue*(e_2) = 1; *ResolveValue*(e_3) = 1;
ResolveValue(e_4) = 2; *ResolveValue*(e_8) = 4

Another important issue to note is that the removal of an edge involved in conflicting situations may lead to new conflicts. Let us further assume that there is an existing prerequisite role constraint

in which role r_1 is defined as a prerequisite role of r_5 in Figure 2. In this case, although the removal of edge e_1 can resolve three existing conflicts, $conflict_1$, $conflict_2$ and $conflict_3$, it may result in the violation of the prerequisite role constraint and a new conflict $conflict_6$: {*assign* (r_5,u_1) , $\neg$*assign* (r_1,u_1)} may occur. Given this informal analysis, we introduce a new concept of valid resolution value to identify the real impact of the removal of an edge on conflict resolution.

Definition 10 (Valid Resolution Value). For an edge e in edge cover sets, the valid resolution value of the edge is its resolution value less the number of new conflicts caused by the removal of the edge. For example, the valid resolution value of e_1 is 2. Given the analysis above, we define the following approach to external constraint conflict resolution.

Step1. Identify the edge cover set for each detected external constraint conflict.

Step2. Calculate the resolution value and valid resolution value for each edge in edge cover sets.

Step3. Identify the set of edges in which each edge's resolution value is equal to its valid resolution value. If the set is not empty, identify and select an edge with the maximal resolution value and remove the edge; if the set is empty, identify and select an edge with the maximal valid resolution value and remove the edge. If the removed edge is associated with an immediate inheritance relationship, then new immediate inheritance relations need to be added between the immediate descendants of its source vertex and target vertex in the role hierarchy, and corresponding edges should be created in $G(RCG)$.

Step4. Calculate the resolution value and valid resolution value for each edge involved in the remaining conflicts.

Step5. Repeat step 3 and step 4 until no conflict remains.

Conclusions and future work

In this paper, we provide approaches to help security administrators in RBAC systems to construct a consistent constraint schema. Constraint conflicts are classified into two categories: internal constraint conflicts and external constraint conflicts. A set of internal constraint conflict detection rules is defined to guarantee the exemption of inconsistency and ambiguities within constraints. When a new constraint is created, the current configuration of a RBAC system may be in conflict with the constraint and external constraint conflicts occur. By the identification of conflict patterns and the construction of RBAC configuration graph, we can detect external constraint conflicts effectively. To guide the resolution of external constraint conflicts, we present new concepts of resolution value and valid resolution value that are useful to guide the resolution process since they represent the effect that the removal of an explicit assignment will have on the resolution of conflicts.

References

[1] JAEGER T. On the increasing importance of constraints[C]// Proceedings of 4th ACM Workshop on Role-Based Access Control, Fairfax, Virginia: ACM, 1999: 33–42.

[2] AHN G J. The RCL 2000 language for specifying role-based authorization constraints[D]. George Mason University, 2000.

[3] LI N H, TRIPUNITARA M V, BRIZI Z. On mutually exclusive roles and separation of duty[C]// Proceedings of the 11th ACM Conference on Computer and Communications Security, New York: ACM, 2004: 42-51.

[4] SOHR K, AHN G J, GOGOLLA M. Specification and validation of authorization constraints using UML and OCL[C]// Proceedings of 10th European Symposium on Research in Computer Security, Milan, Italy: Springer Verlag, 2005: 64-79.

[5] HELIL N, RAHMAN K. RBAC constraints specification and enforcement in extended XACML [C]// Proceedings of 2010 International Conference on Multimedia Information Networking and Security (MINES), Nanjing, China: IEEE, 2010: 546-550.

[6] STREMBECK M. Conflict checking of separation of duty constraints in RBAC implementation experiences[C]// Proceedings of the Conference on Software Engineering, Innsbruck, Austria: ACTA, 2004: 1-6.

[7] MOON C J, PAIK W J, KIM Y G, KWON J H. The conflict detection between permission assignment constraints in role-based access control[C]// The 1st SKLOIS Conference on Information Security and Cryptology, Beijing, China: Springer Verlag, 2005: 265-278.

[8] JANPITAK N, SATHITWIRIYAWONG C. Run-time enforcement model for dynamic separation of duty [C]// Proceedings of 2010 International Symposium on Communications and Information Technologies (ISCIT), Tokyo, Japan: IEEE, 2010: 115-120.

The Research and Simulation of Multilayered Satellite Routing Algorithms

Liu Tong[a], Yang Chunxiu [b] , Zhang Linbo [c]

Harbin Engineering University, China, Harbin

[a]liutong@hrbeu.edu.cn, [b]ycx05081@126.com, [c]zhanglinbo@hrbeu.edu.cn

Keywords: satellite network; routing algorithm; MLSR; delay report

Abstract. Based on the deep discussion and the research of existing satellite routing algorithms, this paper puts forward a suitable routing algorithm (NMLSR) for LEO&MEO&GEO multilayer satellite network. This algorithm gives full play to the ground gateway in communication based on MLSR. Using the regularity and predictability of the satellite communication network, let the ground gateway store and transmit part of link information. This method can reduce routing computation cost and shorten the router update time. At the same time, to enhance capacity of the the network to adapt to the sudden traffic, NMLSR increases accesses of monitoring the satellite network traffic flow and chosing satellite.

Introduction

As an important part of the third generation mobile communication, the satellite communication has become a powerful means of modern communication, because of its prominent advantages, such as the global coverage, simply accessing, extensible, and variable bandwidth according to the needs. Compared with single-layer LEO satellite network, the multilayer satellite network has advantages of high utility of space spectrum, network flexibility, strong survivability, and diversified functions etc [1-3]. Satellite communication also gradually changes from the previous work mode of the single heavenly body, pure forwarding to the one that consist of many stars and has the ability to deal with things [4,5] . Facing the complex satellite communication network, the satellite routing technology as the core of the satellite network technology, determines the performance of the whole network system.

In recent years, people began to research the multilayer satellite network routing algorithm and multi-layer satellite routing algorithm MLSR put forward by Akyildiz, is one of the most representative of satellite routing algorithms. The simulation results show that the performance of the MLSR algorithm is the same as the shortest path routing except for a short oscillatory phase when the hops are switched to a higher satellite layer [4].

But, MLSR algorithm still has several problems following:

(1) Dose not make full use of the regularity and forecasting of the satellite network. Satellites have to send heavy ping messages to obtain the link delay in the network, which leads to high routing computation cost and increases the system burden.

(2) Negative the effect of ground gateways in the routing computation. All the routing computation and management tasks is assigned to GEO and MEO satellites, which reduces the life of the whole system and anti-destroying ability.

Supported by the Fundamental Research Funds for the Central Universities: HEUCF110802.

(3) Lack of the adaptability to sudden changes of network flow, this will lead a larger packet loss rate in the case of unbalanced traffic flow.

Compared with the above questions of the MLSR algorithm, this paper designs a kind of improved multilayer satellite network routing algorithm NMLSR. The idiographic description of the algorithm can be seen in next section.

Definitions

In the satellite network, the links are associated with delays. The total delay is composed of propagation, processing, and queuing delays. Since processing delay is very little and almost is for different satellites. NMLSR algorithm selects propagation and queuing delays as the measurement of links between satellites. And the total link delay is equal to the sum of propagation and queuing delays. Their specific definitions are given as follows:

Definition 1 (Propagation Delay Function) : Let $l_{x\to y}$ be a direct ISL from node x to node y. The delay function $PD\left(l_{x\to y}\right)$ is defined as follows:

$$PD\left(l_{x\to y}\right)=\begin{cases} Prop_{x\to y} & , \quad \exists l_{x\to y} \\ \infty & , \quad \text{otherwise} \end{cases}, \tag{1}$$

Where $Prop_{x\to y}$ is the propagation delay from one to another.

Definition 2 (Queuing Delay Function): Queuing delay function $QD\left(l_{x\to y}\right)$ is defined as follows:

$$QD\left(l_{x\to y}\right)=\begin{cases} Queue_{x\to y} & , \quad \exists l_{x\to y} \\ \infty & , \quad \text{otherwise} \end{cases}, \tag{2}$$

Definition 3 (Total Delay Function): the total delay function $TD\left(l_{x\to y}\right)$ is defined as follows:

$$TD\left(l_{x\to y}\right)=PD\left(l_{x\to y}\right)+QD\left(l_{x\to y}\right). \tag{3}$$

Definition 4 (Propagation Delay Report): propagation delay report PDR is a set of tuples $\left\{x,y,PD\left(l_{x\to y}\right)\right\}$, where x and y are satellites in the network such that $ISL_{x\to y}$ exists between x and y.

Definition 5 (Queuing Delay Report): Queuing delay report $QDR(x)$ of satellite x is a set of tuples $\left\{y,QD\left(l_{x\to y}\right)\right\}$, where x and y are satellites in the network such that $ISL_{x\to y}$ exists between x and y. Queuing delay reports of LEO, MEO and GEO satellites are respectively described as follows:

In this paper, LEO, MEO, and GEO satellites are respectively expressed as $LS_{i,j,k}$, $MS_{i.j}$ and GS_i, and the ground gateway and ground users are respectively expressed as GR and GU.

(1) Queuing delay report $QDR\left(LS_{i,j,k}\right)$ of $LS_{i,j,k}$ satellite is a set of tuples $\left\{y,QD\left(LS_{i,j,k}\to y\right)\right\}$, where y is a adjacent satellite of $LS_{i,j,k}$ such that $ISL_{x\to y}$ exists between them.

(2) Queuing delay report $QDR\left(MS_{i.j}\right)$ of $MS_{i.j}$ satellite is a set of the queuing delay reports of itself and its members,

(3) Queuing delay report $QDR(GS_i)$ of GS_i satellite is a set of the queuing delay reports of itself and its members.

GS_i will exchange their respective $QDR(GS_i)$ in their orbit, then get queuing delay report of the entire network.

Definition 6 (Path) $P_{x\to y}$ is defined as the minimum delay path associated with source x and destination y. It is a sequential list of the satellites on the path.

Definition 7 (Detailed Routing Table) Detailed routing table $DRT(x)$ is calculated and kept by the manager. It provides an entry for each of its care-of members, and registers paths from them to all destinations.

Definition 8 (Simplified Routing Table) Simplified routing table $SRT(x)$ of the satellite x is created by and sent from its manager. Each entry of this routing table has a destination field and a next-hop field.

Implementation

A. The initial process of the satellite network

During the initial stage of creating the satellite network, NMLSR need to accomplish some initial work to ensure the normal communication network, which are described below:

(1) The ground gateway to calculates and stores the information of time division and satellite grouping of the whole network, using STK, and then, transmit it up to the top GEO satellite.

(2) GEO satellite stores the whole information of time division and satellite grouping, and distributes it down to the MEO satellites in its coverage.

(3) MEO satellite stores part of the information of time division and satellite grouping, and distributes it down to the LEO satellites in its coverage.

(4) LEO satellite store the information received from MEO satellite.

In this algorithm, the satellite grouping is based on the standard of the longest coving time, which can decrease the number of the ISL switches.

B. The route update of the whole satellite network

The route update of the whole satellite network is described as follows:

(1) The ground gateway GR calculates and stores the propagation delay of all the ISLs, using STK. When the routing update timer arrives, GR creates the propagation delay report and sends it up to GEO satellite GS_i whose coverage it is in.

(2) After receiving the Propagation delay report, GS_i in transmits it in the same layer.

(3) Once the satellite route update timers arrive, the satellite in the network collects and calculates the queuing delay of its outlinks at the moment. The LEO satellite $LS_{i,j,k}$ generates the queuing delay report $QDR\left(LS_{i,j,k}\right)$, and sent it to the top management o satellite $MS_{i.j}$.

(4) After receiving all the queuing delay report from all team members in LEO layer, $MS_{i.j}$ generate the report $QDR\left(MS_{i.j}\right)$ and its manager GS_i.

(5) Having received all the queuing delay report from all team members in MEO layer, GS_i generates the report $QDR(GS_i)$, and transmits it in its layer.

(6) After receiving all the queuing delay of the network, the GS_i satellite stores the total delay information, generates the total delay report, and transmits it down to MEO satellites in its coverage.

(7) Using the total delay information, GS_i generates $PRT\left(MS_{i.j}\right)$ and $SRT\left(MS_{i.j}\right)$ for its members by the SPF algorithm, stores $PRT\left(MS_{i.j}\right)$ and sends $SRT\left(MS_{i.j}\right)$ to $MS_{i.j}$.

(8) After receiving $SRT\left(MS_{i.j}\right)$ from GS_i, $MS_{i.j}$ generates $PRT\left(LS_{i,j,k}\right)$ and $SRT\left(LS_{i,j,k}\right)$ for its members $LS_{i,j,k}$ by the SPF algorithm, stores $PRT\left(LS_{i,j,k}\right)$ and sends $SRT\left(LS_{i,j,k}\right)$ to $LS_{i,j,k}$.

Thus, one route updating in the whole satellite network has finished.

C. The congestion avoidance mechanism

(1) Generate the congestion report. The satellite uses real-time monitoring of the link direction of the number of cached data packets in the network interface layer. When the number exceeds a pre-set threshold, which is seen as congestion, the satellite will generate congestion report.

(2) Transmitting the congestion report. The LEO satellite sends the congestion report to its manager MEO satellite, and MEO will send the congestion report of itself or its member to the GEO satellite. The GEO satellite not only transmits the report in the GEO layer, but also distributes it to the lower MEO satellites.

(3) Selective re-routing. MEO and GEO satellite receives the congestion message, check the detailed routing table of itself and its members, and the path through the congested link with the SPF algorithm re-routing. MEO and LEO satellite receives the routing update packet, the routing information stored prior to and update the routing table.

(4) The congestion recovery. MEO and LEO satellites schedule a timer for a certain period, and recover the initial route table when the timer arrives.

Simulation

The simulation tools using in this paper are STK and VRNET, where STK is used to obtain orbit information of satellites and covering relationship between two satellite layers, and main simulation work is done using VRNET which is a discrete event network emulator and can very applicable for wired and wireless network simulation .

Table 1 The simulation parameters of the satellite network

Parameters	Satellite layer		
	GEO	MEO	LEO
Satellite number	3	12	48
Orbit altitude[km]	35786	20000	1400
Orbit angle [^{o}]	0.0	55.0	52.0
Orbit number	1	3	8

In the simulation, we chose LEO&MEO & GEO three-layered network structure, the parameters of the simulation network structure is shown in Table 1. In the network, satellites in the same orbit distribute with the equivalent interval. In VRNET the simulation scene and satellite node internal structure are shown in Fig. 1, where the ground gateway is located at (45 N, 120 E).

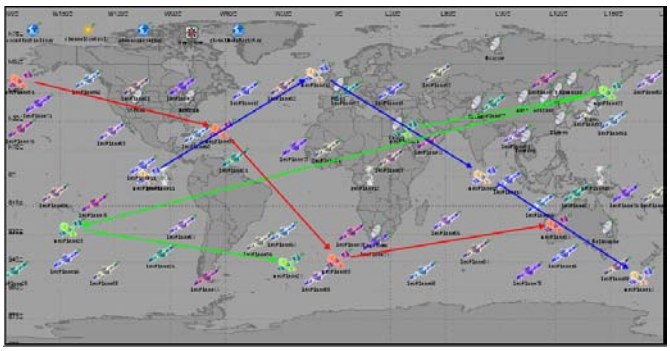


Fig. 1 The running scenario of satellite network under VRNET

In all simulations, The capacities of all UDLs, ISLs, and IOLs are chosen as 200 Mb/s, and each outgoing link has been allocated a buffer space of 5 Mb. If we assume an average packet size of 1000

B, the link capacity becomes 25 000 packets/s and the buffer space becomes 5000 packets. In addition, the period of route updating is 60s, and the congestion recovery interval is 15s.

For performance evaluation of NMLSR and MLSR, we conducted two sets of experiments, show the route overhead, the average end-to-end delay and the packet loss probability difference between NMLSR and MLSR.

A. Routing overhead

To accurately compare the average routing overhead of the two algorithm, we let the traffic flow in the first scene equal to zero, and doesn't add the ground user business. In the ideal state, statistical three layer satellites respectively statistic the average single routing overhead of satellites in different layers. the simulation time is 120 min.

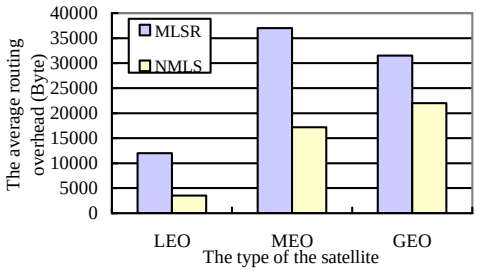


Fig. 2 The average routing overhead of MLSR and NMLSR

As shown in Fig. 2 above, the average single routing overhead of NMLSR is obviously less than MLSR's. It may be result by the following several aspects: first of all, when the route-updating timers arrive, the satellite in arbitrary layer doesn't need to send a ping message to their adjacent satellites in order to obtain link delay information. The LEO/MEO satellite is only required to send queuing delay report to its upper manager. Second, MEO satellite can obtain link information of the entire network from GEO, instead of their exchange local link information with its adjacent satellites. What's more, due to GEO satellite still has responsibility for sending the total delay report and route table to MEO members, the difference between the two algorithms is not obvious.

B. The average end-to-end delay and packet loss rate

To evaluate the performance of the two algorithms, the path metrics (i.e., average end-to-end delay, packet loss rate) of a connection between a source-destination pair are monitored. The source is located at (37.5 N, 112.5 E) in Beijing, China, and the destination is at (12 E, 41 N) in New York, USA. The sender generates a communication business with average rate 8 Mb/s, which will last 120 min.

We set two scenario configurations to compare the performance of the two algorithms under different traffic load. The first scene is used to verify the performance under different LEO ISL utilizations, where the ISL utilizations of GEO and MEO satellites are set as 0. The LEO ISL utilization takes in 17 values between 92% and 100% with an equal interval. In the second scene configuration, the MEO ISL utilization takes in 17 values between 92% and 100% with an equal interval, where the ISL utilizations of GEO and LEO satellites are set as 0. Among them, the ISL utilization is set by changing the background traffic. Each simulation lasts 120 min, and the results are shown in Fig. 3 and Fig. 4.

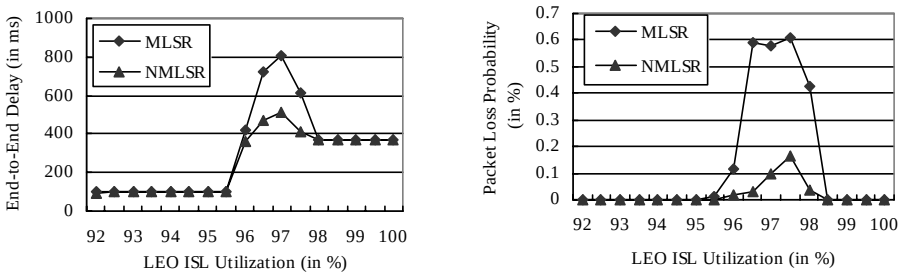
As shown in Fig. 3 and Fig. 4 followed, the performance of the two algorithms is the same except for a short oscillatory phase when the hops are switched to a higher satellite layer. In the oscillatory phase, the end-to-end delay and packet loss rate of NMLSR are much small than MLSR's, the reasons are:

(1)In the oscillatory phase, there are frequent re-routings which can cause much routing overhead. Luckily, since the ground gateway answers for some information transfer, NMLSR doesn't have to

transmit lots of ping packets when rerouting. This greatly reduces the rerouting overhead and shortens the rerouting time.

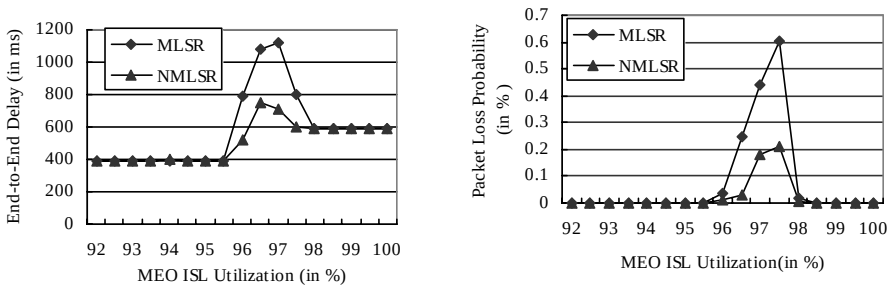
(2) In NMLSR, each satellite can monitor the caches of its outgoing links all the time, and achieves congestion control and recovery. This mechanism not only decreases packet loss caused by the congestion, but also shortens the end-to-end delay.

(3) When choosing the access satellite for ground user, NMLSR selects the satellite with longer cover time and more obligate resource such as frequency band, which in a certain degree reduces end-to-end delay caused by the UDL switch.



(a) The end-to-end delay (b) The packet loss probability

Fig. 3 The performance comparation based on LEO ISL utilization



(a) The end-to-end delay (b) The packet loss probability

Fig. 4 The performance comparation based on MEO ISL utilization

Conclusion

Based on studies and analyses of the existing satellite routing algorithms, the paper proposes and designs an improved multilayer satellite network algorithm NMLSR. This algorithm makes full use of the regularity and forecasting of the satellite communication network, using the ground gateway node transmitting part of link information, which greatly reduces the routing overhead and route update time. NMLSR also increases the flow monitoring mechanism, which can turn away congestion and reduce packet loss rate .Based on the same simulation environment, this paper compares the performance of NMLSR and MLSR routing algorithm. Our simulation results show that whether in routing overhead or end-to-end delay and packet loss rate, the performance of NMLSR is much better than MLSR.

References

[1]Elbert B R: The satellite communication application handbook (2004), p. 8

[2]Chen C: A QoS based routing algorithm in multimedia satellite networks (2003), p.2703-2707

[3]Chen C: Advanced routing protocol for satellite and space network (2005)

[4]Akyildiz I F: MLSR: A novel routing algorithm for multilayered satellite IP networks (2002)

[5]Long F and Sun F C: A QoS routing based on heuristic algorithm for double-layered satellite network (2008), p.1866-1872

A Multiple Attribute Decision Making-Based Access Selection for Heterogeneous WCDMA and WLAN Networks

Fan Ning [a], Pinjing Zhang [b]

Beijing University of Posts and Telecommunications, Beijing, China

[a] ningfan@ccbupt.cn, [b] zhangpinjing@gmail.com

Keywords: heterogeneous wireless networks; access selection; MADM; WCDMA; WLAN;

Abstract: High-speed and mobility are the main features for future traffic, WCDMA/WLAN interworking networks are considered to support this very well. The studies of this issue are focused on multi-radio access selection (MRAS). For MRAS methods, they can be generally parted into several types: traditional single-objective decision, police-based and utility function-based strategies. In this paper, we propose a multiple attribute decision making-based access selection (M-AS) to determine which network is most suitable for a mobile node's service request in heterogeneous WCDMA and WLAN networks. The M-AS method takes five factors into account, network load, date rate, delay, packet error rate (PER) and mobility. Especially the M-AS method considers the different requirements of real-time services and non-real-time services, so it designs different weights for these factors respectively in different occasions in order to obtain better performance. Simulation results show that the M-AS method achieves smaller PER and delay, larger system throughput than the UFAS method [1] which takes the same factors into account.

1. Introduction

The next generation wireless network is envisioned as a convergence of different wireless access technologies, such as GSM/EDGE Radio Access Network (GERAN), UMTS Terrestrial Radio Access Network (UTRAN), Wireless Local Area Network (WLAN), Worldwide Interoperability for Microwave Access (WiMAX), etc. The independent work of these networks results in a huge waste of radio resources, therefore, how to integrate and use a variety of wireless network resources efficiently has been a hot spot in the research field of radio communication [2].

Supporting high-speed multimedia services is one main characteristic in today's wireless communication system and the set of applications running on the mobile devices is increasingly diversifying. While some of these applications are very well suited to run over 3GPP access systems (e.g. VoIP over EPS) for their stringent QoS requirement, some other applications may also be well suited to run over some other complementary access system (e.g. best effort high bit rate FTP transfer via WLAN). 3GPP has already presented an interworking model between 3GPP access system and WLAN and IP flow mobility and Seamless Offload (IFOM) which enable applications to choose different access networks according to their types [3].

For multi-radio access selection (MRAS), besides traditional single-objective decision [4], police-based [5] and utility function-based strategies [1], recent research has focused on Multi-Attributes Decision Making (MADM). However, the single-objective decision often has the optimal result on one point, not the whole networks, and police-based method which uses WLAN networks as soon as it is available has high handoff frequency when the MN is at a high speed or roaming in the cell boundary. A utility function-access selection (UFAS) method can efficiently use the network resources [1], but it doesn't consider the characteristics of different services and can't offer the best users' experience. In this paper, we present a multiple attribute decision making-based

access selection (M-RS) for heterogeneous WCDMA and WLAN networks. Compared to UFAS, the M-RS has a better performance in packet error rate (PER), delay and throughout.

The paper is organized as follows. Section Ⅱ describes the heterogeneous WCDMA and WLAN networks. Section Ⅲ describes the M-RS method. After that, performance evaluation and the simulation analysis are presented in the Section Ⅳ. Finally conclusion is given in Section Ⅴ.

2. System Architecture

A heterogeneous network of WCDMA and IEEE 802.11g networks is considered. The WLAN networks with small coverage underlie the fully deployed multi-cell WCDMA networks. The WLAN networks act as hotspots to support high data rate transmission to serve as a complementary technology to WCDMA.

In the heterogeneous networks, an interworking layer is performed at radio network controller (RNC), which records the QoS requirements of MNs and collects the link states of MNs and channel qualities of WLAN and WCDMA cells. According to the information, the network access selection is determined by the proposed M-AS method. The MNs, moving around in the heterogeneous networks, are equipped with two different RF modules of WCDMA and WLAN radio interfaces so as to connect with the two networks simultaneously. Each MN has an MAC layer to combine multiple services into a single data stream so that they can transmit over a single radio interface.

3GPP [6] defines four traffic classes, which involves conversational class, streaming class, interactive class, and background class. These four traffic classes are clustered into two groups, the real-time group, and the non-real-time group. The real time group includes most delay sensitive applications, which involves conversational and streaming classes. The non-real-time group cares PER more, which includes interactive and background classes.

3. M-AS Method

To illustrate the proposed M-AS method more clearly, we divide it into three parts. Firstly, we describe the cells classification that can speed up the cell selection and provides a suitable and reasonable access decision. Then, we analyze the main factors which play an important role in the access selection. Finally, based on the MADM theory, we formulate our access selection algorithm named M-AS, which is used to improve the QoS and the throughput of whole network.

A. Cells Classification

Two constraints are considered to select cells according to the service characteristics to obtain a candidate cell group N for the service request.

1) Signal Strength Constraint

When an MN moves to the areas which cover by multiple networks, firstly the MN measures the signal strength around in order to eliminate the cell with deficient signal strength. If the received signal strength from cell $n \in N$, denoted by PWn, exceeds a given power threshold, the cell is feasible.

2) QoS constraint

Besides the signal strength, we also need to compare the QoS requirement of services to the QoS provided by candidate network. If the network can satisfy the requirement, then it can be chosen.

B. Main Factors

In the proposed M-AS method, five factors are taken into account: network load, data rate, delay, PER and the mobility. The load balancing index η denotes the loading of cell n in WCDMA or WLAN networks, which has been obtained in [7, 8].The data rate, delay and PER are important indexes of QoS. And the mobility is defined as:

$$f_{v,n}^{(s)} = \frac{2l_n/v}{\sum_{n\in N} 2l_n/v} = \frac{l_n}{\sum_{n\in N} l_n} \tag{1}$$

Where l_n is the radius of cell n and v is the MN's current velocity. The dwell time for the MN in cell n will be less than $2l_n/v$ if the velocity and moving direction are unchanged. When the MN is in high mobility, $f_{v,n}^{(s)}$ will influence the selection result that can avoid frequent handoffs. A smaller cell that

will incur more handoffs have a less chance to be selected.

According to the features of real-time services and non-real-time services, we give different weights to these factors. The non-real-time services prefer to a lower packet loss but can receive a certain delay. So in M-AS method, we will choose appropriate network according to services' characteristics.

C. Access Selection Processing: a MADM problem

TOPSIS (Technique for Order Preference by Similarity to Ideal Solution), as one of the widely used method to solve MADM problem[9], is based on the principle that the chosen candidates should have the shortest distance from the ideal solution and the farthest distance from the negative-ideal solution. It is relative simple and easy to understand. Thus, we choose TOPSIS as the basic method to carry out the access selection processing.

Using TOPSIS, the first step is to normalize the decision matrix. Here we use the method based on the average level to realize it [10]. According to the character of different factors, we divide them into two terms: benefit and cost. Those belong to benefit should have the common feature that the larger, the better, and the cost ones, just the opposite character. To eliminate the difference between them, we make some conversions to them, and the formulas are given in (2) and (3).

$$x_{ij} = \frac{a_{ij} - m_j}{m_j}, (i = 1,2,\cdots,K; j = 1,2,\cdots,M) \tag{2}$$

$$x_{ij} = \frac{m_j - a_{ij}}{m_j}, (i = 1,2,\cdots,K; j = 1,2,\cdots,M) \tag{3}$$

Considered the factor value can't maintain a certain proportion with the normalized one sometimes. We use the following equation (4) to cope with it based on the reason that it denotes the deviating degree between factor and its average value.

$$b_{ij} = \frac{1-a^{-x_{ij}}}{1+a^{-x_{ij}}}, a > 1 \tag{4}$$

Here, we take the general choice makes $a=e$, thus, the above equation can be expressed as:

$$b_{ij} = \frac{1-e^{-x_{ij}}}{1+e^{-x_{ij}}} \tag{5}$$

Thus, with the equation (5), we can get the normalized matrix B by converting the x_{ij} got by (2) and (3):

$$B = \left(b_{ij}\right)_{K\times M} \tag{6}$$

In the second step, we will get the weighting vector ω, used to illustrate the weight of different factors. According to the feature of services, we can easily get matrix $C_{M\times M}$ of the real-time services which presents the relations among them. Here, p_1~p_5 represents *Network load*, *Data rate*, *Delay*, *PER* and *Mobility* respectively.

$$C_{5\times5} = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & p_4 & p_5 \end{array} \\ \begin{bmatrix} 1 & 2 & 2 & 4 & 2 \\ 1/2 & 1 & 1 & 3 & 3 \\ 1/2 & 1 & 1 & 3 & 3 \\ 1/4 & 1/3 & 1/3 & 1 & 1/2 \\ 1/2 & 1/3 & 1/3 & 2 & 1 \end{bmatrix} \end{array} \tag{7}$$

Based on the following eigenvector method, we can get the weight vector ω according to the matrix *C*. The first is to get the ω_i^*, defined as:

$$\omega_m^* = \sqrt[M]{p_{m1} \times p_{m2} \times \cdots \times p_{mM}}, 1 \le m \le M \tag{8}$$

Then, normalize the ω_i as:

$$\omega_m = \frac{\omega_m^*}{\omega_1^* + \omega_2^* + \cdots + \omega_M^*}, 1 \le m \le M \tag{9}$$

Thus, we can get the weighting vector:

$$W = (\omega_1, \omega_2, \cdots, \omega_M)^T = (0.347, 0.234, 0.234, 0.074, 0.116)^T \tag{10}$$

Therefore, this step has not been completed. As the matrix $C_{M\times M}$ is the foundation of weighting vector *W*, its rationality will play a crucial role in a successful process of access selection. Thus, a

consistency check is necessary. Here, we use the Saaty method to get the dominant eigenvalue. Firstly, get summation of every row in matrix$C_{M\times M}$:

$$S_j = \sum_i^M p_{ij}\,, 1 \le j \le M \tag{11}$$

The dominant eigenvalue λ_{max} can be expressed as:

$$\lambda_{max} = \sum_i^M \omega_i S_i \tag{12}$$

With the equation (10) (11) (12), we can get the dominant eigenvalue of matrix$C_{5\times5}$: λ_{max} =5.20. Refer to the ultimate eigenvalue of five order matrix λ'_{max}=5.45, that is to say:

$$\lambda_{max} < \lambda'_{max} \tag{13}$$

It meets the request of consistency check, so we can draw the conclusion that the given matrix$C_{5\times5}$ is available in our selection scheme.

As the same way, we can obtain matrix$C'_{5\times5}$ of the non-real-time services:

$$C'_{5\times5} = \begin{bmatrix} 1 & 2 & 4 & 2 & 2 \\ 1/2 & 1 & 3 & 1 & 3 \\ 1/4 & 1/3 & 1 & 1/3 & 1/2 \\ 1/2 & 1 & 3 & 1 & 3 \\ 1/2 & 1/3 & 2 & 1/3 & 1 \end{bmatrix} \tag{14}$$

With the above results, we can get the integrate decision matrix:

$$V = A_{N\times M} \cdot \omega \tag{15}$$

The third step is to determine the ideal network $A^*$and the negative-ideal network A^-. They are shown in (16) and (17), where J is associated with the benefit criteria and J' is associated with the cost criteria.

$$A^* = [v_1^*, v_2^*, v_3^*, v_4^*, v_5^*] = \left\{ \left(max_i\, v_{ij} \,|j \in J\right), \left(min_i\, v_{ij} \,|j \in J'\right) \middle| \begin{matrix} i = 1, \cdots, N \\ j = 1, \cdots, 5 \end{matrix} \right\} \tag{16}$$

$$A^- = [v_1^-, v_2^-, v_3^-, v_4^-, v_5^-] = \left\{ \left(min_i\, v_{ij} \,|j \in J\right), \left(max_i\, v_{ij} \,|j \in J'\right) \middle| \begin{matrix} i = 1, \cdots, N \\ j = 1, \cdots, 5 \end{matrix} \right\} \tag{17}$$

Next, the fourth step is to calculate the separation of each alternative from the candidate networks, and the negative candidate network, using the formula given in (18) and (19)

$$S_{i^*} = \sqrt{\sum_{j=1}^5 \left(v_{ij} - v_j^*\right)^2}\,, 1 \le i \le N \tag{18}$$

$$S_{i^-} = \sqrt{\sum_{j=1}^5 \left(v_{ij} - v_j^-\right)^2}\,, 1 \le i \le N \tag{19}$$

The relative closeness to the ideal network can be calculated in the fifth step using the formula:

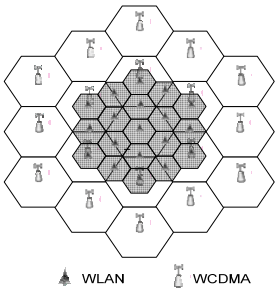
$$C_{i^*} = S_{i^-}/(S_{i^-} + S_{i^*})\,, 1 \le i \le N \tag{20}$$

Thus, according to the C_{i^*}, we can pick out the suitable network as the access network of service.

4. Simulation Results and Discussion

A. Simulation Environment

The simulation environment and parameters is shown in Figure 1, where WLAN system (with radius of 1.5km) and WCDMA system (with radius of 3km) are fully overlaid and the APs of center cells of the two systems are overlaid. At the beginning of simulation, MNs are randomly distributed in the overlap area and generate the motion direction randomly. When a MN moves to the border of a cell, handover is executed. The MNs are respectively at the velocity of 3*km/hr*, 60*km/hr* or*120km/hr* with equal probability. The moving direction of a MN changes $\pm45°$ with probability of 0.2 and the location of a MN is updated once every 100*ms*. Every MN generates the real-time traffic and the non-real-time traffic with equal probability independently. The threshold PW_n of WCDMA and WLAN systems are assigned by -102*dBm* and -75*dBm*.



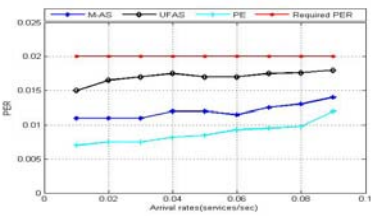
Voice Traffic Type	Traffic Parameters
Mean talkspurt/silence duration	19/14 seconds
Mean call holding time	180 seconds
Required delay	300 milliseconds
Required PER	0.05
Web-browsing Traffic Type	**Traffic Parameters**
Mean rate	32 Kbps
Maximum file size	1858 bits
Mean reading time	30 seconds
Mean interarrival time	0.125 seconds
Mean call holding time	300 seconds
Required delay	1 second
Required PER	0.005

Figures.1simulation scenario and parameters

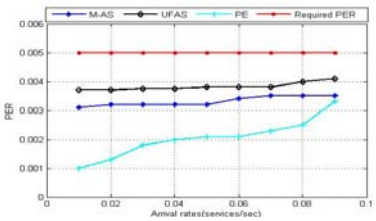
System performances of the M-AS method are compared with the UFAS method and PE handoff decision method [5]. The UFAS method which is based on utility function took the same five factors into account, but it doesn't consider the different requirements of real-time services and non-real-time services. The PE method used a cost function, which considered cost, power consumption and bandwidth. The strategy of the PE method selects WLANs mostly because WLANs have lower cost and larger bandwidth than WCDMA networks.

B. Simulation Results

The performances of the two systems' center cells are collected in simulation. Figure 2(a) and (b) show the PERs of voice service and web-browsing service in WCDMA networks, respectively, where the red line is the required PER value of voice service or web-browsing service. It can be found that the PERs of both services with M-AS are lower than that of UFAS, it indicates M-AS has a better improvement. However, the PERs of both services are larger than that of PE method, it is because the M-AS selects the suitable network according to not only PER but also network load, data rate, delay, and mobility. It is possible that a network with mildly lower channel quality (higher PER) has better system throughput or less number of handoffs.

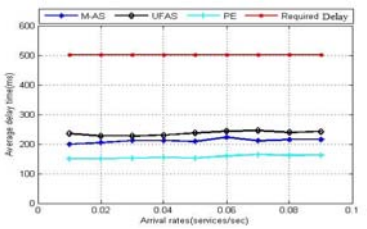


(a)

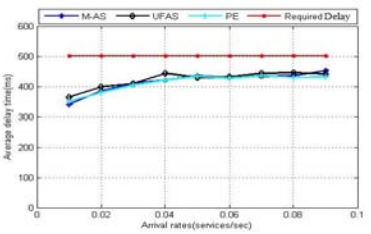


(b)

Figures.2 (a) The PER of voice service in WCDMA networks, (b) the PER of web-browsing service in WCDMA network



(a)

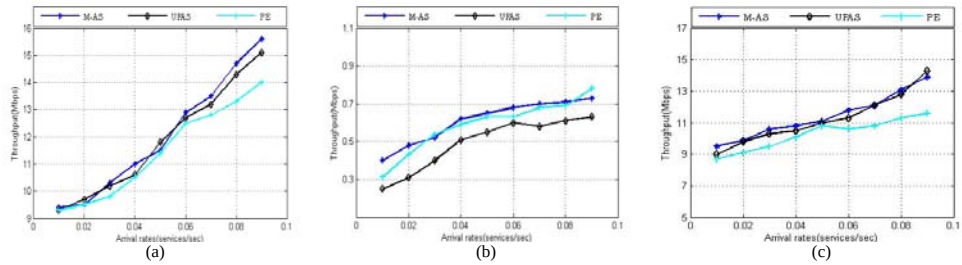


(b)

Figures.3 (a) the average delay time of web-browsing service in WCDMA networks, (b) the average delay time of web-browsing service in WLANs

Figure 3(a) and (b) show the average delay time of web-browsing service in WCDMA networks and WLANs, respectively. Although the average delay time of web-browsing service with the M-AS method in WCDMA network is larger than that with the PE method, it is still far under the required delay time. Because WCDMA networks tend to accommodate the MNs with high-mobility and low

transmission rate, the M-AS method takes a conservative strategy in WCDMA networks for web-browsing service that demands higher transmission rate. But the delay with M-AS is still smaller than that of UFAS. In WLANs, the average delay time of web-browsing service with the three methods is almost the same.



(a) (b) (c)

Figures.4 (a) the system throughput of the heterogeneous WCDMA and WLAN network, (b) the individual throughput in WCDMA, (c) the individual throughput in WLAN

Figure 4(a) depicts the system throughput. The system throughput is defined as the sum of the transmission bits per second in WCDMA and WLAN networks air interface. Figure 4(b) and (c) show the individual throughput in a WCDMA cell and in a WLAN cell, respectively. It can be found that the M-AS method achieves a little higher system throughput than UFAS method and obviously higher than PE method, where the individual throughput in a WCDMA cell with the M-AS method is smaller than that with PE, but the individual throughput in WLAN networks with the M-AS method is larger than that with the PE method. This phenomenon results from the design philosophy of the M-AS, which adopts service required date rate and mobility as the factors for network access selection. WLANs tend to accommodate low-mobility, high transmission rate MNs, while WCDMA networks tend to accommodate high-mobility, low transmission rate MNs.

5. Conclusion

This paper presents an M-AS method to determine the network access selection in heterogeneous WCDMA/WLAN networks. Besides system QoS requirements, MNs' mobility, and networks' load balancing, the M-AS method also takes different requirements of real-time services and non-real-time services. Simulation results show that the proposed M-AS method has smaller PER and delay, and higher system throughput than UFAS method. These result from the design philosophy of the M-AS method, which tends to select WLANs to accommodate low-mobility MNs requiring high transmission rate and WCDMA networks to accommodate high-mobility MNs requiring low transmission rate. Also, the M-AS method tries to balance the traffic load among networks and maintains the QoS requirements for both WCDMA and WLAN networks.

References

[1] Y. Chen, N. Yang, and C. Chang, *A utility function-based access selection method for heterogeneous WCDMA and WLAN networks,* in *Proc.* IEEE PIMRC, 2007, pp. 1–5.

[2] Stuchmann P, Zimmermann R. *Toward ubiquitous and unlimited capacity communication network*, European research in framework program 7. IEEE Communications Magazine，2007, 45(5) : 148-157

[3] 3GPP TS 23.261 v10.1.0, "IP flow mobility and seamless Wireless Local Area Network (WLAN) offload; Stage 2 (Release 10)", Sep. 2010.

[4] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. P. Makela, R. Pichna, and J. Vallstrom, "Handoff in hybrid mobile data networks," *IEEE Personal Communications*, vol.7, no.2, pp,34-47, Apr. 2000.

[5] H.J. Wang, R.H. Katz, and J. Giese, *Policy-enabled handoffs across heterogeneous wireless networks*, IEEE Workshop on Mobile Computing Systems and Applications, pp, 51-60, 1999.

[6] 3GPP, QoS Concept and Architecture, Release 6, TR 23.107, Mar. 2004.

[7] H. Holma and A. Toskala, *WCDMA for UMTS, 2nd Edition.* John Wileyand Sons, LTD., 2002.

[8] R.-G. Cheng, C.-J. Chang, C.-Y. Shih, and Y.-S. Chen, *A new scheme to achieve weighted fairness for WLAN supporting multimedia services*, IEEE Transactions on Wireless Communications, vol. 5, no. 5, pp. 1095–1102, May 2006.

[9] Azibi R, Vanderpooten D, *Construction of rule-based assignment models*. European Journal of Operational Research, 2002, 138:274-293.

[10] Ching-Lai Hwang and Kwangsun Yoon, *Multiple Attribute Decision Making*, Springer-Verlag, 1981.

A Review of Ensemble Method

Hui-Lan LUO[1, a], Zhong-Ping LIU[2,b]

[1,2] School of Information Engineering ,Jiangxi University of Science and Technology ,Ganzhou, Jiangxi, China

[a]3972988@qq.com, [b] lzp_llz@163.com

Keywords: Ensemble Method, Ensemble Classifier, Combination Rules

Abstract. This paper aims to provide a review of ensemble method which is a machine learning algorithm by use a series of learning classifier and combine the outcomes of its to get the better result than any single one. In this paper, the background of ensemble method is presented firstly, secondly the typical approach of ensemble method are introduced ,at last, some of the latest developments and research focus in order to make further research are presented.

1. Introduction

Ensemble method could get the better performance than a single classifier when learning new data through combining a series of single base classifier. The history of ensemble method dates back to as early as 1977 with Tukey Twicing[1]. By 1979, Dasarathy and Sheela proposed using an ensemble system in a divide-and-conquer fashion, partitioning the feature space using two or more classifiers[2]. Over a decade later (1990), the main progress was achieved by Hansen and Salamon [3]. They showed that from ensembles of similarly configured neural networks can improve the predicative performance than from a single one. The same year, it was Schapire that put ensemble method at the center of machine learning research. He showed that a strong classifier in probably approximately correct sense be got through boosting a multiply "base classifier" and these classifier performance is required only slightly better than random guess[4]. Boosting algorithms requires to knowing the error bound of the base classifier , while that is usually impossible to get. The Adaboost algorithm, experimented by Freund and Schapire(1996),which does not require this unavailable information. Owing to their sound theoretical foundation, very accurate prediction, and great simplicity, Adaboost and other variants have been applied to diverse fields with great success[5].

2. Popular Ensemble Theory

2.1 Structure of ensemble method.

Krogh & Vedelsby proved that the generalization ability of ensemble classifier was better than a single classifier only if both there exists differences among predicted result from different classifier involved and that each classifier error rate is less than 0.5[6]. Empirically, the more diversity among the classifiers, the better the performance of algorithm is than the single one. Typically, an ensemble is constructed in two steps. First, a number of base classifiers are produced in a parallel or sequential style. Then the base classifiers are generalized in certain combination scheme to make more accurate as possible than a single base classifier[7]. Figure 1 illustrates the basic structure of ensemble system.

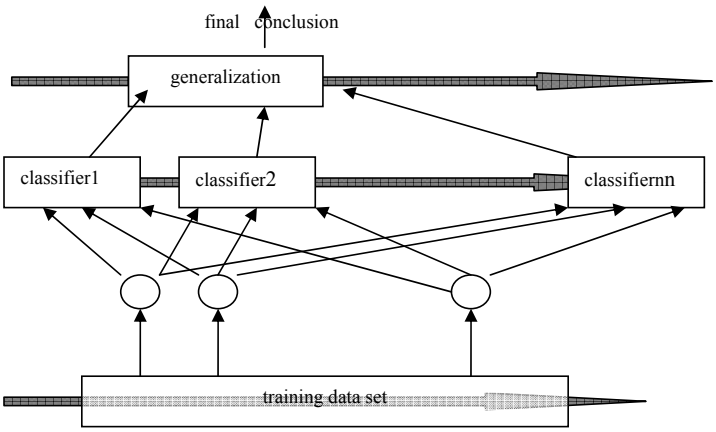


Fig.1 Ensemble Structure

2.2 Representative classifier.

Following will briefly introduce two representative methods, Bagging and Adaboost. For simple,considering the binary classification , we have the data pair $D=\{X,Y\}=\{(x_i,y_i)|\ (i=1,2\ldots m),y_i=\{1;+1\}\}$ and X denote the instance space (feature space) while Y denote the value of class labels which are to be learn . After training the D, algorithm will output a hypothesis, which is a mapping from X to Y, or called classifier.

2.2.1 Bagging

The algorithm of bagging consolidates the outputs of various base classifiers into a single classification by voting, class label value is predicted. The predication accuracy in a multi-classifier system is higher than any individual classifier. Specifically, each classifier in the ensemble system is trained on the sample of instances taken with replacement strategy from the training data. One of the main advantages of bagging is that it can be easily implemented in a parallel mode on different processors simultaneously [9]. The pseudo-code of Bagging is shown in Figure 2.

Input:
Data Set $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$;
Process:
for t=1,…,T;%reperat T round
Dt=Bootstrap(D); % Generate a bootstrap sample from D
$h_t=L(D_t)$; % Train a base learner *ht* from the bootstrap sample%
end.
Output: $H(x)=\arg\max_{y\in Y}\sum_{t=1}^{T}[h_t(x)=y]$

Fig.2 The Bagging algorithm

2.2 .2 Adaboost[7,9,10,11]

Contrary to bagging, boosting make much focus on the missed classified data for next round sampling[10]. In fact, there are many variants in boosting family. Let the most famous implementation Adaboost as an example, the pseudo-code of it is summarized in Fig.3. First, it assigns equal weights to all the training instances. D_t is the weight distribution of the round t . From the training data set and D_t, the algorithm generates a base classifier h_t :X →Y . By then, it uses the training examples to test h_t, and the weights of the incorrectly classified examples will be increased. Thus, an updated weight distribution D_{t+1} are got. From the training data set and D_{t+1} Adaboost generates another base classifier again. Such a process is repeated for T round, and the final classifier is derived by weighted majority voting of the T base classifier, on which the weights of the learners are determined during the training process.

Input:
Data Set D={$(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)$};
Process:
$D_1(i)=1/m$
for t=1,…,T;

$h_t=L(D,D_t)$; %Train a base learner ht from D using distribution Dt%

$e_t=Pr_{i\sim D_t}[h_t(x_t)\neq y_i]$; % Measure the error of h_t %t

$$\alpha_t=\frac{1}{2}\ln\frac{1-e_t}{e_t};\qquad D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}\exp(-\alpha_t);if\,(h_t(x_i=y_i))\\ \exp(\alpha_t);if\,(h_t(x_i\neq y_i))\end{cases}$$ % Update the distribution,

$$=\frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$ % Z_t is a normalization factor which enables D_{t+1} to be a distribution

end

Output: $H(x)=sign\sum_{t=1}^{T}\alpha_t\ h_t(x_t)$

Fig.3 The boosting algorithm

2.3 Ensemble combination rules

Combining multiple base classifiers and make final conclusion is this section content. Three approaches are introduced: voting, stacked generalization and cascade generalization.

2.3.1 Voting

Considering voting methods operated on binary labels, $d_{t,j}$ is 1 or 0 depending on whether classifier t chooses j or not. The ensemble system chooses the class j which has the largest sum of vote [12]:

$$\sum_{t=1}^{T}d_{t,j}(X)=\max\sum_{t=1}^{T}d_{t,j}(j=1,...C) \tag{1}$$

If there are T classifiers for a two-class problem, the decision maked by the system will be correct if at least $\lfloor T/2+1\rfloor$ classifiers choose the correct class. Now assume that each classifier has a probability *p* of making a correct decision. then, the probability of making a correct decision has a binomial distribution, specifically, the probability of correct out of T while k>(T/2) is:

$$Pens=\sum_{k=(T/2)+1}^{T}\left(\begin{matrix}T\\k\end{matrix}p^k(1-p)^{T-k}\right) \tag{2}$$

Note the requirement p>0.5 is necessary and sufficient for a two class problem, whereas it is sufficient, but not necessary for multi-class problems. The optimized algorithm weighted majority voting is shown in Figure 4:

$$\sum_{t=1}^{T}w_t d_{t,j}(X)=\max\sum_{t=1}^{T}w_t d_{t,j}(j=1,...C) \tag{3}$$

The optimal weights for the weighted majority voting rule can be shown to be: $w_t\infty\frac{p_t}{1-p_t}$ if the *T* classifiers are class-conditionally independent with accuracies $p_1,p_2,\ldots,p_T$.

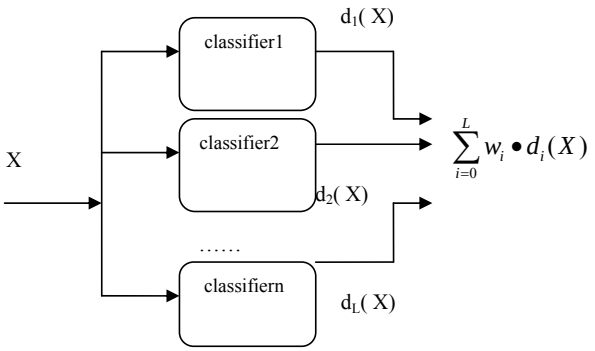


Fig.4 Weighted majorityvoting

2.3.2 Stacked generalization(stacking)

Stacking is a rule of combining multiple classifiers by Wolpert in 1992[13,14]. Typically, the stacking consists of multilevel heterogeneous classifiers, respectively called base classifier and top classifier. Figure 5 is the model of stacking. The top classifier learns on the output of base classifier, estimated via cross-validation as follows:

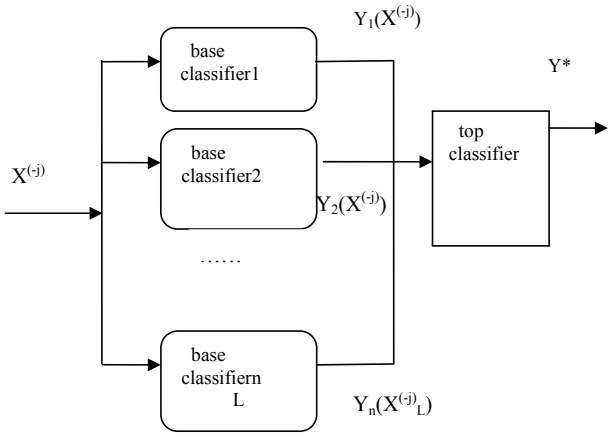


Fig.5 Stacking generalization

Given a data set D={X,Y}={(x_1,y_1), (x_2,y_2)…(x_n,y_n)},where X is a value of attribute vector and Y is the class label value. Split randomly X into J equals parts X_1,X_2,…,X_j and define $X^{(-j)}$=X- X_j as training data set for the jth round J-fold cross-validation. First trains the data $X(^{-j})$ on the given N base classifier, and output the result $Y_n(X(^{-j})$. Then the $Y_n(X(^{-j})$ are input into the top classifier and the output Y* is the final conclusion.

2.3.3 Cascade generalization

The basic idea of cascade generalization, presented by Gama, is to use sequentially a series of classifiers, at each step performing an extension of the original data by the insertion of new attributes[15]. The new attributes are derived from the probability class distribution given by a base classifier. Figure 6 presents a N-pipeline cascade generalization model, which is the most simple form that just only a classifier at each level.

Given a learning set D={X,Y}={(x_1,y_1), (x_2,y_2)…(x_n,y_n)},where X is the input value vector , and Y is the output variable. On classification case, y_n is the class label values predefined. Let M(D) denote the classifier learned by applying M on the data set D, then M(x_i,D) denotes assigning a class label to an example x_i. In each classifier M(X,D) output a probability distribution, [p1,p2,…pc],which represents the probability class label the X belongs to .

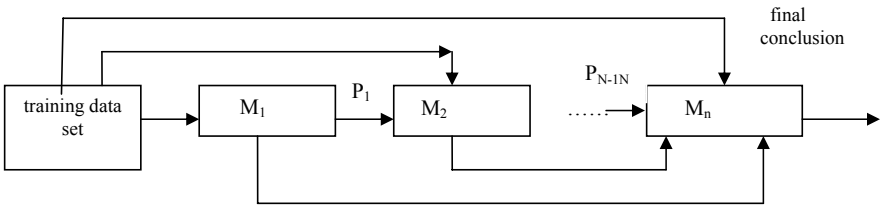


Fig.6 Cascade generalization

4. Summary

Ensemble learning is a powerful method in machine learning field in deal with diverse application such as optical character recognition, text categorization, face recognition, computer-aided medical diagnosis, gene expression analysis, etc. By using multiple learners, the generalization ability of an ensemble can be much better than any of a single learner. The current main deficiency ensemble methods is following topics: combinations of different sources of diversity; understanding and interpretation of ensembles; understanding and explaining in more basic terms why ensembles perform better that individual model. If those issues can be addressed well, ensemble learning will be able to contribute more to more applications.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) project (No. 61105042), Shanghai Key Laboratory of Intelligent Information Processing, China (Grant No. IIPL-09-009), Natural Science Foundation of Jiangxi Province, China (Project No. 2010GZS0075), and Educational Commission of Jiangxi Province, China (Project No. GJJ11464).

References

[1] Tukey J.W., Exploratory Data Analysis，Addison Wesley, 1977.
[2] B.V.Dasarathy, and B.V. Sheela, A composite classifier system design: concepts and methodology, Proceedings of the IEEE, 67, 5, May 1979
[3] L.K.Hansen, and P.Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, No. 10, October 1990
[4] R.E.Schapire. The strength of weak learnability. Machine Learning, 5(2) 1990.
[5]Yoav Freund & Robert E.Schapire, Experiments with a New Boosting Algorithm , Machine Learning: Proceedings of the Thirteenth International Conference, 1996.
[6] Krogh,A.& Vedelsby,J. ,Neural Network Ensembles, Cross Validation and Active Learning, in Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994] 1995
[7] Z.-H. Zhou. Ensemble learning. In: S. Z. Li ed. Encyclopedia of Biometrics, Berlin: Springer, 2009
[8] BREIMAN, Leo, Bagging predictors. Machine Learning, 24(2) 1996
[9] Lior Rokach, Pattern Classification Using Ensemble Methods, World Scientific PubCo Inc, 2010
[10] Z.-H. Zhou & Yang Yu. Adaboost., The Top Ten Algorithms in Data Mining, Boca Raton, FL: Chapman & Hall, 2009
[11] http://en.wikipedia.org/wiki/Adaboost
[12] http://www.scholarpedia.org/article/Ensemble_learning
[13] David H. Wolpert ,Stacked generalisation, Neural Networks 5, 241-259, 1992
[14] Kai Ming Ting & Ian H. Witten,Stacked generalization: when does it work?, Proceedings of the 15th international joint conference on AI 1997
[15] João Gama & Pavel Brazdil,cascade generalization,Machine Learning , Vol 41 Issue 3, Kluwer Academic Publishers ,December 2000

Research on Innovative Practice of Informatization in China Rural Areas

Ruyuan Li, Zhi'an Wang, Weihua Zheng, Junxia Yan
Han Dan College
461662080@qq.com

Keywords: Rural Information Service Station; Rural Informatization; Twelfth Five Year Plan; Unattended Children; Empty Nester; Innovative Model

Abstract: The realization of informatization in China rural areas is a strategic mission for national development. In 2006, former Ministry of Information Industry issued that: the task for establishing pilot cities of rural comprehensive information services. Called upon by the nation, the three communication giants and relevant big-sized enterprises all make great efforts to realize rural informatization. The key lies in setting up rural information service stations which would undoubtedly contribute to understanding peasant and rural areas as well as applying tools of information society. But it is difficult for majority of the service stations to survive if government support is unavailable. As a solution to realize long-term survival for service station, it is suggested that on the basis of fully awareness towards social development, incorporation with co-operation pattern as well as new approaches to invest and finance should be adopted.

1 Case of Propagating Rural Informatization Assisted by Rural Information Service Station of "Educational Training Pattern"

Characteristic of multi-media classroom operating point, the pilot program of rural informatization is called as "Educational Training Pattern" Program. It covers establishing regional center cities as bases, gaining guidance from government Information Office and support from all levels of departments of rural government, fixing the enterprises with informational and technological experiences on education application as the implementing body as well as setting the platform of website information distribution and management as the link.

A. Program.

Selecting suitable public areas in each village and decorating a piece of activity room simply sized between 40~60 square meters where a rope of internet line is linked, a projection and a computer with multi-terminals are installed and an information officer in charge of security, education and operation is provided. Works that need to be done: training on spreading internet knowledge; providing information technological support and service; organizing advanced peasants to explain technology and experience for acquiring wealth; helping the peasants with information search and release; helping the peasants with maintaining family contact through videos; broadcasting science education and entertainment programs; publicizing national or governmental policies, regulations and so on.

B. Implementation in Details

According to the total population and the situation of public areas, the amount of investment to equipments of different levels is different. In principle, the total investment in each point should be no greater that 30,000 RMB. Hardware investment projects are listed as follows:

Under the help of village committee (direct access into peasant household is considerable under

special conditions), a convenient public area is selected where a piece of room is decorated simply, electricity is equipped (security should be taken into consideration), a fan, 20~30 special seats and two pairs of office equipments are provided. The total investment is about 2500 RMB.

Under the support of government, a rope of internet line should be built with Netcom, Telecom or TieTong with which ADSL (Asymmetrical Digital Subscriber Loop) with comparatively low prices can be realized. The total investment is about 500 RMB a year.

A computer of moderate cost performance with one main engine and two terminals should be provided and be equipped with two operating terminals and a camera. The total investment is about 8,000 RMB a year. A durable projecting apparatus with bulb of long life span that can resist voltage and be turned on or down in time should be selected. In addition, another set of twenty meters long video cable and patch board should be provided for outdoor usage during night times. And a set of power amplifier acoustic equipment for outdoor usage should be provided. The total investment is about 8,000~15,000 RMB.

C. Operation

Catered to peasants who are engaged in agriculture, the system provides service to adult peasant with certain study capability rather than young teenagers. The key for operating staff in the first three months is propagation with certain intensity and density which can avail the peasants to know that asking for advice, studying, releasing and searching information is available in the activity room. Specific operating approach is as follows:

Broadcasting science and education film with interludes to introduce the function of operating points to attract peasants idling at home.

Teaching peasants to contact with their children through internet telephone and video.

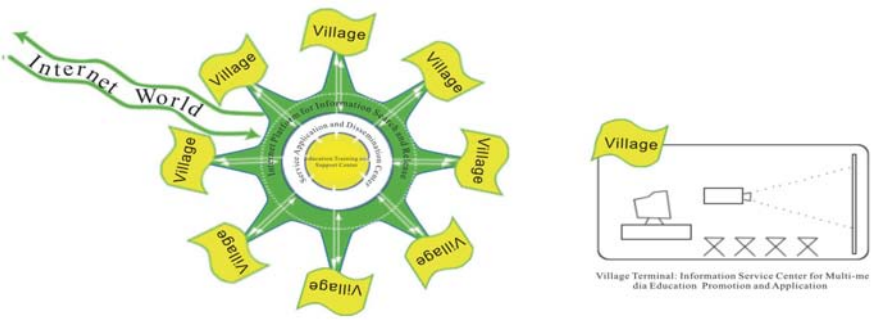
Teaching peasants to release supply and demand information and accumulate local farming resources. Joint efforts should be made for establishing 《Han Dan Online》 channel with farming information.

After three months' of propagation, all kinds of training classes with low price can be started. One kind is instruction by the operating staff, and the other is distance teaching via video and radio by online headquarters.

After three months, local residents with special capacities should be kept in touch with and their experience can be propagated through internet.

D. Payments Estimation

Staff salary, bulb consuming of the projection, electric charge, internet costs are main costs in the one-time equipment investment. Suppose that monthly salary of each staff is 800 RMB, costs on the projection bulb is 360 RMB in thirty days within a month with six hours a day and the costs being two RBM an hour, the total electric charge is 120 RMB a month within thirty days with 0.8 RMB per day if five degrees are consumed a day, the total cost for internet within a month is 40 RMB a month, the total monthly expenses is 1320 RMB with no regard to the situation of equipment depreciation.



Schematic Picture for Rural Informatization in "Educational Training Pattern"

E. Conclusion

The below conclusion can be reached after thorough introduction to rural comprehensive service station of information: the total hardware investment is about 30,000 RMB. Monthly expense of the service station is about 1500 RMB. According to nearly two years' experience of pilot experiment in rural areas, service station cannot survive without government subsidies.

2 Design on the Survival of Rural Comprehensive Information Service Center

The design for the survival of rural comprehensive information service center is an integrated innovative survival based on internet, cooperation and capital. Dependence on internet doesn't simply mean the net, but an omnipresent network in one region that has unified mark in unified service pattern. In such case can the network gain momentum and influence, consequently the foundation for survival has been laid; Dependence on cooperation suggests that the service station should benefits all the local residents. Based on the idea of cooperation, hundreds of households should make joint efforts for establishing the station. The station itself which plays the role of mutual assistance should serve the local residents through cooperation. Furthermore, the other layer of the meaning of dependence on cooperation is that service station should cooperate with innovation and technology investment enterprises in regional center; Dependence on capital suggests that under the guideline of mutual cooperation, partners should make joint contribution to make benefits and in turn maintain the survival of service station.

According to the above-mentioned design, a more detailed process is described as follows: to establish rural comprehensive information service stations with unified labels and service contents in a comparatively wider rural region with innovation and technology investment enterprises. Featured the same properties like co-operations, each service station should be established on the joint capital of the local households. According to the unified requirements of the nation for rural co-operations establishment, regulation should be made, membership should be absorbed and shares should be enlarged. In principle, the share for each person is 10,000 RMB and members for the co-operations should be restricted within 100. It is agreed that capital of the service station (co-operation) should be controlled integrally by innovation and technology investment enterprises (parts of the capital should be reserved for urgent situation for some members). Other benefits are disposable for purchasing life necessities and maintaining the operation of service station on the basis that interests for the members are no less than those of the banks'.

The things innovation and technology investment enterprises should be taken into consideration is whether there are projects and enterprises locally for investment as well as projects and enterprises for investment in other regions.

3 Feasibility on Operating the Survival Design for Comprehensive Service Station of Information in Rural Areas

The key to the success of the design lies in whether technology investment enterprise can transform investment assets into benefits. It should be noted that national economic growth rate is above 8%. 15% of annual investment benefit is accessible for an enterprise with good projects. For the moment, it is agreed that rate of return is 15 %(yields that surpass 15% are preserved in investment enterprise) among which 3% is saved for operating investment enterprise, the remaining 12% is saved for returning membership benefits. That is to say annual disposable income is 1%. It is estimated that share for each member is 10,000 RMB, in that way, monthly disposable expenditure is 100 RMB. Putting aside 350 RMB which comes from the one-year deposit income, 35 RBM is used for replacement out of convenience. In such case, 65 RMB is left among which 15 RMB is for operating service station and 50 RMB is for group purchasing necessities of members (oil, salt, sauce and vinegar). As shown in Table1, after gaining momentum, 62.5 RMB of retailing commodity quality is brought to the service station out of 50 RMB. As shown in Table2, another simpler approach is utilizing 80 RMB for group purchasing necessities (oil, salt, sauce, vinegar and

energy) and 20 RMB for operating without sharing interests to members.

TABLE I. BUDGET OF ECONOMIC DISTRIBUTION IN RURAL COMPREHENSIVE INFORMATION SERVICE STATION MADE UP OF 100 MEMBERS

Joint Funds for Every Individual	Monthly Investment Income	Benefits for Members in Return	Group Purchasing fee for Members in Return (Estimated Income: 50*125%=62.5)	Money at the Service Station's Dispose (15*100)
10,000RMB	100RMB	35RMB	50RMB	15RMB

TABLE II. BUDGET OF ECONOMIC DISTRIBUTION IN RURAL COMPREHENSIVE INFORMATION SERVICE STATION MADE UP OF 100 MEMBERS

Joint Funds for Every Individual	Monthly Investment Income	Group Purchasing fee for Members in Return (Estimated Income: 80*125%=100)	Money at the Service Station's Dispose (20*100)
10,000RMB	100RMB	80RMB	20RMB

In practical survey, expenses on life necessities are 50 RBM to 70 RMB for an average family with four members; and energy consumption costs between 100 RMB to 120 RMB. That is to say both the two approaches can satisfy basic life requirements of members with different needs in different areas. Through this way, an information service station is established and operated with the money saved from group purchasing life necessities. The cooperation between different service stations and cooperation between service station and central station would bring huge changes to local education, enjoyment of the aged and information propagation. Families with more than 10,000 RMB gain large proportion. They gain profits out of joint funds and contribute to others by operating service station.

4 Sustainability on Operating the Survival Design for Comprehensive Service Station of Information in Rural Areas

The sustainability of the survival pattern should be taken into account. There would be risks if investment enterprises seek economic profits solely by investing the funds to projects with large interests. Possessed with the most advanced financing and investment technology both domestic and abroad, innovative enterprises should conduct projects in rural areas and around every service station. Aiming at rural development and construction, the target design originates from and takes root in rural areas on the bases that peasant needs are satisfied and membership incomes increased.

Basic service function of the information service station is providing informatization education, purchasing commodities online on behalf of members, helping the members with online payment, offering communication services for free and providing some paid service to nonmembers according to market situation. After gaining momentum, service station can win profit support from industries like banking, marketing and communicating with which service contents covering the physical examination organization as well as wedding and funeral operation can be extended and strengthened. Being the station for surrogate purchasing, group purchasing and problem settlement, the comprehensive service station is educational and cultural with modern innovative productivity.

Accomplishing informatization in rural areas is a national strategic policy. Advancing the pace of urbanization doesn't equal to abandon the countryside. It is time for providing life care, psychological counseling, emergency assistance, health care and legal aid to empty nesters bothered with mental crisis accompanied by the aggravating phenomenon of empty nest; It goes the same for rural unattended children. It is of avail to enhance mutual understanding between family members apart in two places through video communication service provided by information service station. Once becoming mature, service station can recruit university students to rural areas for internship

program with the cooperation of universities (Work Replacement System for students to local schools has been established by Ministry of Education in Heibei province. It is of great significance if students are sent to work in service station as well).

In all, union is strength. Service station is established through joint efforts of members, connection between different villages is realized through service stations and all the service stations are united together into a group by technology investment enterprises. Being capable of increasing income and reducing expenditures, rural investment and progress will become the new pattern for modernized rural technological development.

5 Conclusions

Foundation for designing the survival pattern of rural comprehensive service station of information lies in joint funds, cooperative establishment, rational financing and profit development. As a result, this may spread doubts over the legitimacy of money collection.

As for this aspect, it is agreed that the establishment of rural comprehensive service station of information accords with national strategic policy. Service station establishment under the principles that guided co-operation accords with national policy on developing agriculture, countryside and peasant. The service station involves with all the residents in the neighboring villages and towns rather than general public. Capital venture investment enterprises are selected for taking charge of financing and increasing profits on the basis of absorbing modern financing concept towards joint funds when no mature projects are available; in the second place, by means of increasing income and reducing expenditure on the basis of group purchasing life necessities, meager profits can be increased; Last but not least important, local villagers should not invest too much money, and 10,000 RMB is enough. Clauses that benefit members should be made in cooperative regulations.

In conclusion, it is deemed that the design of service station is innovative which accords with national policy and benefits rural informatization, rural education and new countryside service mode development.

References

[1] Discussion on the Development of Agricultural Informatization, Zhangli and Tan Xigui, the fourth issue of Journal of Anhui Agricultural Sciences, 2003.

[2] The Development and Suggestion of Technology Supporting System of Agricultural Informatization in China, Guo Shupu , the fourth issue of Journal of Anhui Agricultural Sciences, 2003.

[3] Walking out Two Error Cognitions in Construction of Agricultural Informatization in China, Lin Tao and Liang Xian, the fourth issue of Guangxi Academy of Agriculture Science, 2003.

[4] Discussion on ‘the First One Hundred Meters’ of Agricultural Informatization in China, Lin Tao, the fifth issue of Journal of Agriculture Southwest China, 2006.

[5] Study on the Agricultural-Informationlization Organization System, Cui Yan, doctoral dissertation of Northwest Agriculture and Forestry University, 2007.

[6] Research on the Strategies in the Development of the County Economic Informatization, Zhang honggang, doctoral dissertation of Agricultural University of Hebei, 2008.

Dynamic Test for Elastic Modulus and Damp Ratio of Three Layers Plywood

Wang Zheng[1, a], Yang Xiaojun[1], Wang Xiwei[2], Li Jie[1], and Fan Wubo[1]

[1] College of wood Science and Technology
Nanjing Forestry University, Nanjing, 210037, China
[2] College of Mechanical and Electronic Engineering
Nanjing Forestry University, Nanjing, 210037, China
[a]wangzheng63258@163.com

Keywords: plywood, dynamic elastic modulus, damp ratio, computer science

Abstract. The dynamic elastic modulus and damp ratio of Plywood were measured by adopting the mode of "cantilever beam" freeness end, and the static elastic modulus of plywood were measured to test the correctness of the dynamic testing method by means of computer science. The results showed that, the dynamic elastic modulus and static elastic modulus were consistent, which show the merit of dynamic measure, such as speediness, handy and good in repetition etc.

Introduction

The modulus of elasticity E and damp ratio ξ are two important indexes for weighing the physical property of wood and non-wood composite material. In practical use, the modulus of elasticity and loss of material is very important for the study of the use, modification and microstructure of material. In addition, in the non-destructive check for material, the dynamic elastic modulus and the change of loss are also the important indexes for judging if there is the defect in the material. Thus, more and more studies have put attention to the dynamic elastic modulus of material in the production and practice. Except the standard static force method, study also adopt the dynamic method of measuring its natural frequency to measure the dynamic elastic modulus of material. Among them, the vibration test by means of computer science can be used for measuring the modulus of elasticity and logarithm attenuation ratio of vibration damp, which has the application value.

In the view of the plywood owning the characteristic such as increasing the width of mats, carrying on the advantages of natural wood and developing the utilization ratio and so on, the plywood owns the longest history and widest use, so in this paper the method of the free end of the cantilever beam is adopted, and mainly do the vibration test for measuring the dynamic elastic modulus Ed and damp ratio ξ of the three layers plywood.

Material and instrument

Material. In the thesis, the vernier caliper and analytical balance are used for measuring its magnitude and weight, the detailed result is in table 1.

Table 1 Three layers Plywood density

length l （mm）	width b（mm）	height h（mm）	quality m（g）	density ρ（g/cm3）	M.C.（%）
310	52	28	390	0.86	10.8

Instrument. Vernier caliper (0～500mm), Electrooptical analytical balance(0.01mg) and Dial gauge(0.001mm); Vibration and dynamitic signal collecting analysis system including control box, collecting box and software SSCRAS, Made in Nanjing AnZheng Sftware Engineering Co,Ltd. Piezoelectric accelerate sensor, model CA-YD-107.

Measuring method and process

Measuring method and principle. In the way of the free end of the cantilever beam (Fig.1), the static elastic modulus Ee of the samples of plywood can be calculated according to static force increment method, as shown by Eq.1.

$$E_e = \frac{pl^3}{3I(\triangle l)}, I = \frac{1}{12}bh^3, E_e = \frac{4\triangle pl^3}{bh^3\triangle(\triangle l)} \tag{1}$$

Among them, I-Inertia moment, p-Static force, Δl-Deflection of free end of cantilever beam.

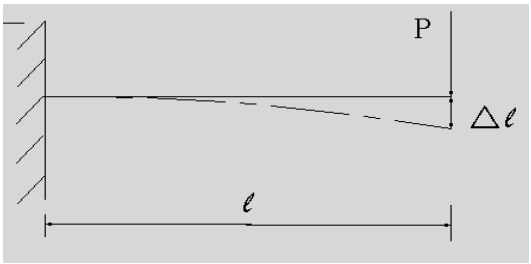


Fig.1 Static forces test figure on the free end of cantilever girder

In the way of the free end of the cantilever beam, the dynamic elastic modulus Ed of the plywood can be calculated according to stochastic exciting power spectral method, as shown in Eq. 2.

$$E_d = \frac{16\pi^2(M + \frac{33}{140}m)l^3 f_1^2}{bh^3} \times 10^{-9}(GPa) \tag{2}$$

Among them, M —accelerate mass (g).

The metrical principium is that install the accelerometer with the mass 29g on the fee end of the cantilever beam plywood with the mass 390g, through the collector and computer, vibrating and dynamic signal collecting analysis system software and data collecting and spectral analysis, the first natural frequency of three layer plywood f1 can be calculated and the Ed also can be reckoned, in addition, the natural vibration cycle or frequency value can be gotten by the time that the ten time domain waveforms pass.

In the way of the free end of the cantilever beam, the damp ratio ξ of samples can be calculated by the free attenuation tested by the dynamic signal, its simplified formula is shown in Eq. 3.

$$\delta = \frac{\ln(A_1 / A_5)}{5} \quad \varepsilon = \frac{\delta}{2\pi} \tag{3}$$

In the view of the rich damp information in the free vibration of the cantilever beam, so that the thesis synthesizes the ξ of the system by the time domain method of amplitude logarithm attenuation ratio and frequency domain method of half power belt width. When the structure of the three layer

plywood vibrate freely evolved, because of the damp, its amplitude assumes the logarithm attenuation waveform, and the values of amplitude A1, A2 ,…An can be read from the vibration wave curve through time domain attenuation, which can calculate the logarithm range of decrease coefficient δ and damp ratio ξ.

Measuring process

Measuring process of static elastic modulus. Upon the free end of the cantilever beam, the weight can be increased by $\Delta p =$ 0.98N every time, which is shown in Fig.1. At the action of the static force P, the deflection Δl can be measured at the lower of the plywood.

Measuring process of dynamic elastic modulus. In fig.2, install the plywood well to realize the cantilever beam, on the set collecting the parameter of the vibration signal, adopt the negative spring method (the spring is 10%,the spring delayed is -20), the range of the voltage is from -5000Mv to 5000Mv, and set the average of four times. The analysis frequency is 500Hz,the accuracy can be reached to 1.25H or more. In order to prevent the frequency mixing together, the experiment choose the upper limit of the lowpass, the frequency is set for 500Hz. Before the measurement, go into the condition of the showing wave, knocking on the samples continuously to check whether the instrument is switched on or not and whether the waveform is reasonable or not, if not reasonable, set it again, when measuring, knock on the samples with rubber hammer, touch off to collect data, and do the dynamic signal frequency, and numerate the first natural frequency value or the average cycle or frequency that ten waveform pass from the first pink value on the power drawing.

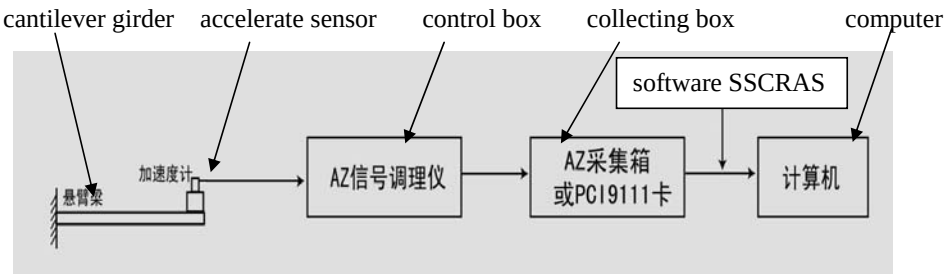


Fig.2 Flow chart of measuring dynamic elastic modulus and damp ratio

Calculation results

Static elastic modulus and result calculation. The static elasticity modulus condition of the plywood calculated by static force P can be seen in Fig.2. in the Fig.1, every sample can be measured through ten times loading, every load is $\Delta p =$0.98, and deflection Δl of the sample at the free end of the cantilever beam by the dial gauge, and also, calculate the every practical deflection $\Delta(\Delta l)_i$ and the average deflection, thus the static elasticity modulus can be gotten that is Es=1808.92MPa, which is shown in Table.2.

Table 2 static elastic modulus of Plywood

P （N）										
Measuring parameters	0.98	1.96	2.94	3.92	4.90	5.88	6.86	7.84	8.82	9.80
Δl （mm）	0.031	0.088	0.14	0.188	0.240	0.298	0.375	0.43	0.485	0.540
$\Delta(\Delta l)_i$ （mm）	0.057	0.052	0.048	0.052	0.058	0.077	0.055	0.055	0.055	
$\Delta(\Delta l)_i$ （mm）	0.057									
E_e (MPa)	$E_e = 1808.92$									

Dynamic elastic modulus measurement and damp ratio. At the condition of satisfying the structure dynamics theory, on the one hand, the thesis adopts the cantilever beam free end method to measure the first step natural frequency though the random prompting power method and also calculate its dynamic elasticity modulus value, on the other hand, the thesis adopts the dynamic test free attenuation method by the cantilever beam free end method to ensure the damp ratio of three player plywood through the method of amplitude logarithm attenuation ratio and frequency domain method of half power belt width. The time domain, power chart and data of the three layer plywood can be seen in Fig.3, Fig.4 and Table 3. The Fig.3 is the time domain chart of three player plywood, its x-coordinate denotes the time (ms) and ordinate denotes the oscillation amplitude (EU). Obviously, the samples of three layer plywood have the obvious characteristic of high damp ratio, its damp ratio is higher than the elasticity material such as steel and so on. This is based on the hypothesis the force without outer cycle acting on vibrating three player plywood, its original energy can be diminished because of the part radiation production and inter friction energy. Because the higher damp capability of the sample of the three layer plywood weaken its free vibration, so, the continuous oscillation amplitude is smaller and smaller, until the samples of three layer plywood come back to the static condition. Especially in the frequency chart on the Fig.4, its x-coordinate denotes time(ms) and ordinate denotes frequency (Hz), which reflects the characteristic of vibration energy of the plywood material diminishing fast and bigger damp.

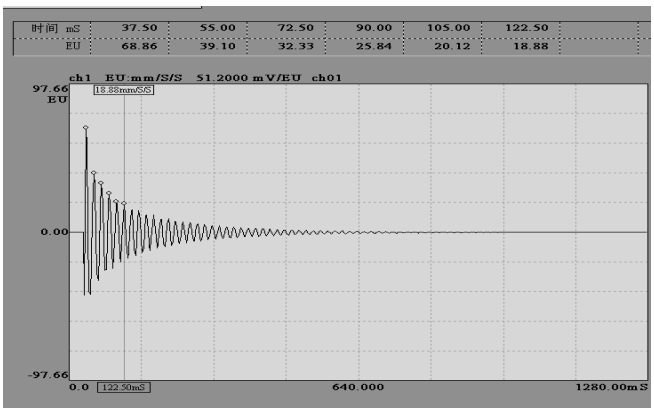


Fig.3 Wave figure

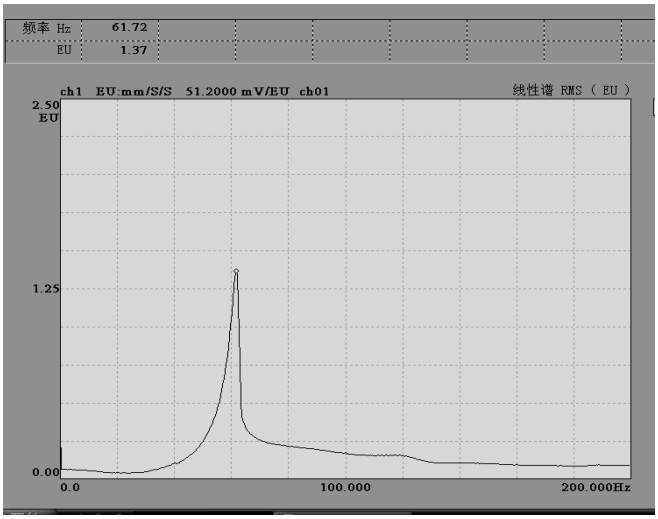


Fig.4 Frequency chart

In Fig.4, the first step natural frequency f1 of the sample of the three layer plywood is 61.72Hz, so the dynamic elasticity modulus value can be got by Eq. 4

$$E_{\mathrm{d}} = \frac{16\pi^2 (M + \frac{33}{140} m) l^3 f_1^2}{bh^3} \times 10^{-9} \tag{4}$$

The result is $E_{\mathrm{d}} = 1896.57$ (MPa), which is shown in table 3.

Table 3 Dynamic elastic modulus and damp ratio of Plywood

f_1 (Hz)	A_1 (EU)	A_5 (EU)	ξ (%)	E_{d} (MPa)	E_{e} (MPa)
61.72	39.10	18.88	2.31	1896.57	1808.92

Summary. The follow main conclusions can be drawn from the present study.

In Table 3, the dynamic elastic modulus value is close to the static elasticity modulus value, which reflects the dynamic and static test results are the same.

The Fig.3 and Fig.4 show that the plywood material has the obvious characteristic with higher damp ratio, its damp property is higher than the elastic property of steel and so on. Especially the frequency chart on the Fig.4 reflects the characteristic of vibration energy of the plywood material diminishing fast and bigger damp.

According to the analysis of relativity of recursive function of plywood test specimen and dynamic elastic modulus value, the plywood has good linear characteristic.

In order to bring the reducing vibration and diminishing energy function into play, it is suggested that take the plywood like the three layer plywood as the damp material under consideration.

The method of dynamic signal used for measuring vibration has the obvious advantage, the accuracy of the dynamic elasticity modulus of the three layer plywood is higher than the static test value, and also can reflect the damp ratio information timely, in addition, the test time of the dynamic measurement is shorter.

Acknowledgements

This study was funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

[1] Ruan Xigen, Yu Guanxia: *Wood physics* (China Forestry Publishing House, Beijing, August 2005).
[2] Sheng Zhaoshun, Yin yiling: *The technology applying of equipment inspect and fault diagnose* (Metallurgical Industry Press, Beijing, 1995).
[3] Wu Zhengyi: *Testing technology and signal processing* (Qinghua University Press, Beijing, 1991).
[4] Mei Changtong, Zhou Xiaoyan, Jin Juwan: *Man-made board* (China Forestry Publishing House, Beijing, December 2005).
[5] Wang Quanzhong: Study on dynamic characterictic of the bamboo plywood. Journal of Nanjing Forestry University Vol. 21(1997), p. 80～82.
[6] Chi Deru: *Dynamic detection for elastic modules and bend strength of fiberboard* (Wood Processing Machinery, Nanjing, 2006).
[7] Hu chengying, Wang Fengshan, Liu Yixing, Tetsuya Nakao: Lossless detection of the dynamic deflection modulus of elasticity for flakeboard. Journal of Northeast Forestry University Vol. 29 (2001), p. 9～11.
[8] Dai Chengyue, Liu Yixing, Ding Hanxi: Study on wood strength detection with ultrasonic, Journal of Northeast Forestry University Vol. 15(1987), p. 82～95.

Automated Accompaniment System Based on Bayesian Mining of Score Context

Ryosuke Yamanishi[a], Ryogo Okamura and Shohei Kato[b]

Dept. of Computer Science and Engineering , Graduate School of Engineering,

Nagoya Institute of Technology

[a]ryama@juno.ics.nitech.ac.jp, [b]shohey@juno.ics.nitech.ac.jp

Keywords: Music, Automated accompaniment, Bayesian mining, Score context

Abstract. Recently, composing has received greater attention as one of the forms of enjoying music, and most multimedia contents should be created with an original music. However, the composition of music is difficult for people who do not have sufficient musical background or experience. We therefore developed an automated accompaniment system focusing on bass and drums. In this paper, the musical temporal variates were considered the score context and the nuances of the music in a database were learned using a Bayesian network. In composing experiments, we observed various accompaniments depending on the used database. Results suggested that this system could effectively learned the nuance of the database. We tested the effectiveness of the proposed system with subjective evaluation experiment, and result showed that this system can enable users to compose music with accompaniments that has the nuance of the music in the database, even if they do not have much musical experience.

Introduction

Recently, there are diverse and numerous forms of entertainment to enjoy, which has been known as *Multimedia entertainment*, and almost all of them are composed with music because music provides people with more affective effects.

To create more affective multimedia contents, an original dramatic music should be enclosed. Thus the research in the field of Information Science, especially in entertainment computing, has focused on sound and music recognition [1-4] and generation [5, 6] systems, and composing music has become more accessible for ordinary people. However, it is still difficult for people who do not have sufficient musical experience to compose an original music because of that the high level of knowledge and experience are needed. We believe that user-friendly music composing support systems are needed, and focus on the music accompaniment composing as an element of such systems in this study.

Accompaniment, which is one of the key factors involved in influencing emotions [7], is nevertheless the most difficult element of a musical composition because it is necessary to know every detail about the interaction between multiple musical instruments. We therefore developed an accompaniment composing support system that can learn the structure of music, and the corresponding interaction between musical instruments by using a Bayesian network [8]. Fig. 1 shows an overview of the proposed system. In this paper, we focus on bass and drums as key factors for making the mood of the music. The system needs three pieces of input information: melody, chord, and section. It then outputs the melody with bass and drums, and the system composes music considering the correlation between the melody and each instrument in the learning database. Users do not need any knowledge or experience to compose, since they only have to previously teach the

978-1-61275-004-0/10/$25.00 ©2012 IERI

system the target songs.

Score Context

Each section is made up of a group of bars, and is defined and classified considering similarities of both the melodies and the chord progression patterns on each bar [9, 10], Some well-known section types include the *bridge*, *motif*, and *chorus*, which are typically considered key concepts for music composition. There is no universal definition of "section," so we assume a group made of the four bars to be a section in this paper. The music for learning in our system is all 4 by 4 and includes melody, bass, drums, chords, and section information. The sixteenth note and rests are the minimum unit and are assumed to be one beat.

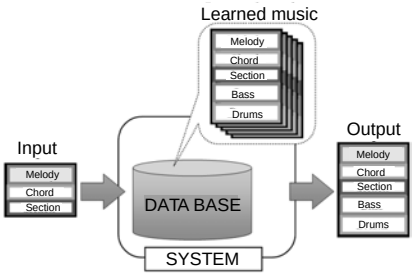


Fig. 1 Overview of the proposed system

Table 1 The reference tone of each chord

Chord	C	C♯	D	D♯	E	F	G	G♯	A	A♯	B
The reference tone	60	61	62	63	64	65	66	67	68	69	70

In this paper, the score context is made from a seven-piece set: chord, rhythm of the melody, pitch of the melody, rhythm of the bass, pitch of the bass, drums, and bar counter. The score context on beat p included in bar t, sc_p^t is shown by following equation, which will be described in detail in the following sections).

$$sc_p^t = \left(ch_p^t, mr_p^t, mp_p^t, br_p^t, bp_p^t, dr_p^t, bc_p^t\right). \tag{1}$$

Modeling Chord. The chord on beat p included in bar t, ch_p^t is shown by the chord name. In this paper, the chord name is made up of the root of the chord and sign that discriminates the tonality of the chord: major or minor [11]. The pitch of the melody and bass are both based on the chord.

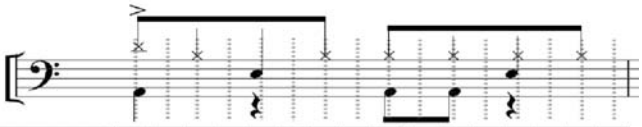
Modeling Melody and Bass. Melody and bass are each modeled by dividing them into two parts: rhythm and pitch [12].

Modeling Rhythm. The rhythm of the melody and the bass on beat p included in bar t are shown as mr_t^p and br_t^p, respectively. Rhythm, which is composed of notes and rests, is a series that has the following three states: RS1 (a note that produces a sound), RS2 (a rest), and RS3 (continuance of the previous state).

Modeling Pitch. The pitch of the melody and the bass are described by whichever numbers between 0 and 127 conforms to the note number of the Musical Instrument Digital Interface (MIDI). If the rhythm state in the beat is RS2, then the pitch of the melody and the bass are φ. The pitch of the melody and the bass on beat p included in bar t are shown as mp_t^p and bp_t^p, respectively. mp_t^p and bp_t^p are shown by relative pitch to ch_t^p. Relative pitch is then calculated by the difference between

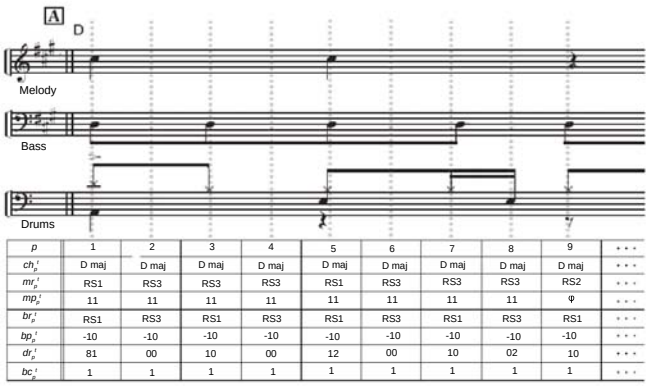
the pitch and the reference tone of the corresponding chord in the beat (see Table. 1).

Modeling Drums. Drums on beat p included in bar t, dr_t^p are described by a combination of the pronunciation states of the percussion instruments that make up the drums. In this study we used eight percussion instruments: crash cymbal, ride cymbal, open hi-hat, closed hi-hat, tom, floor tom, snare, and bass drum. Percussion instruments do not follow any rules of length, so each instrument has only



Crash-cymbal	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ride-cymbal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Open-hihat	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Closed-hihat	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
Tom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Floor-tom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Snare	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Bass-drum	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0
p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
dr_p^t	81	00	10	00	12	00	10	00	11	00	11	00	12	00	10	00

Fig. 2 Example of a drum score and the corresponding drums state



p	1	2	3	4	5	6	7	8	9	...
ch_p^t	D maj	D maj	D maj	D maj	D maj	D maj	D maj	D maj	D maj	...
mr_p^t	RS1	RS3	RS3	RS3	RS1	RS3	RS3	RS3	RS2	...
mp_p^t	11	11	11	11	11	11	11	11	φ	...
br_p^t	RS1	RS3	RS1	RS3	RS1	RS3	RS1	RS3	RS1	...
bp_p^t	-10	-10	-10	-10	-10	-10	-10	-10	-10	...
dr_p^t	81	00	10	00	12	00	10	02	10	...
bc_p^t	1	1	1	1	1	1	1	1	1	...

Fig. 3 Example of score context

two states, RS1 and RS2. An example of drums score and the corresponding drum state dr_t^p is shown in Fig. 2.

Bar Counter. The index of the bar in a section is considered the bar counter, and the bar counter on beat p included in bar t is shown as bc_t^p. When it transitions to the next section, the bar counter is reset.

Fig. 3 shows an example of the score context in a bar. When an entire note rest is observed on the bass or drums in a bar, the score context in the bar is removed from the learning database.

Accompaniment Model

The proposed system stores user's target songs in a learning database, and composes accompaniment models for the rhythm of bass, pitch of bass, and drums. These models are each constructed by the parameters calculated from the score context by using a Bayesian network.

Accompaniment Model for Rhythm of Bass. The accompaniment model for the rhythm of bass is composed by creating a correlation between the rhythm of bass and the following three parameters, mr : the rhythm of the melody, $partition$: the index of the beat in the bar, and bc : the bar counter.

Accompaniment Model for Pitch of Bass. The accompaniment model for the pitch of bass is composed by creating a correlation between the pitch of bass and the following seven parameters, $length$: the length referring to rhythm of the bass, $cadence$: the relative difference between the corresponding chord route and the key of the music, $pre\_cadence$: the cadence of the previous note, $next\_cadence$: the cadence of the next note, $chord\_key$: distinction between major and minor for the key of the chord, $partition$: the index of the beat in the bar, and bc : the bar counter. When the generated rhythm of the bass is RS1, the pitch of the bass on the beat is generated.

Accompany Model for Drums. Accompaniment model for drums is composed by creating a correlation between the drums in the learning database and the following four parameters, mr : the rhythm of the melody, br : the rhythm of the bass, $partition$: the index of the beat in the bar, and bc : the bar counter.

Composition and Subjecting Experiments

We evaluated the effectiveness of the proposed system through composition experiments using various databases and subjective evaluation experiment.

We prepared three databases, each of which each learned five jazz, dance, and rock songs. In the composition experiment, three different accompaniments with respect to each database were generated with common input information: melody, chords, and section information. Fig. 5, 6, and 7 each show an example of accompaniment generated by the proposed system with each database. Despite the use of only one piece of input information, the three generated accompaniments were all different, and it seemed that each had the nuance of their respective genre of the learned database.

Results of the subjective evaluation experiment were conducted to verify whether the composed music had the nuance of the learned songs or not. In the experiment, 25 subjects, 12 of whom had some musical experiences, were asked to listen to the music composed by the proposed system. We also prepared music samples of each genre, which were not contained in the learning music databases, for reference. We did not show the genre names of the music samples but rather labeled them “Genre A,” “Genre B,” and “Genre C.” Subjects chose the most appropriate genre for each piece of music composed by the proposed system from four options: “Genre A,” “Genre B,” “Genre C,” and “Unknown.” The experimental results are shown in Table. 2. We calculated the kappa coefficient from the results listed in Table. 2 and evaluated the correlation between the music genres and the subjects’ answers. Generally, a kappa coefficient of $\kappa \geq 6.0$ means a significantly high correlation rate, and we confirmed that $\kappa = 0.65$ for the results. This suggests that the genre of the composed music conforms very well to each genre in the learning music database.

Conclusion

In this paper, we proposed *score context* as an expressive method of musical score with time-variation and an accompaniment composing system based on Bayesian mining of score context.

Three learning music databases were prepared for the composing experiment, each of which stored jazz, dance or rock music. Despite the fact that only one common input information piece was used, the three composed accompaniments corresponding to each database were all different. This suggests that the composed accompaniment achieved the genre-specific nuance of the learned songs in each database. The subjective evaluation experiment demonstrated the effectiveness of the proposed system. Results of listener evaluations showed that each genre of composed music conformed to that

of the learned music.

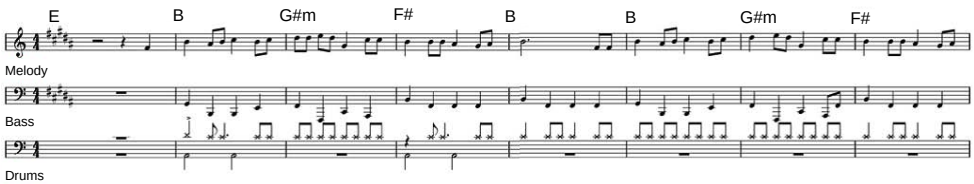


Fig. 4 Output example composed by the proposed system with Jazz database

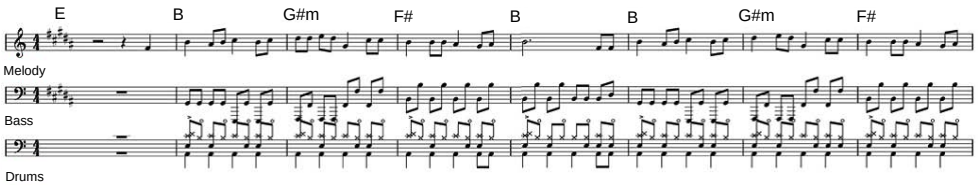


Fig. 5 Output example composed by the proposed system with Dance database

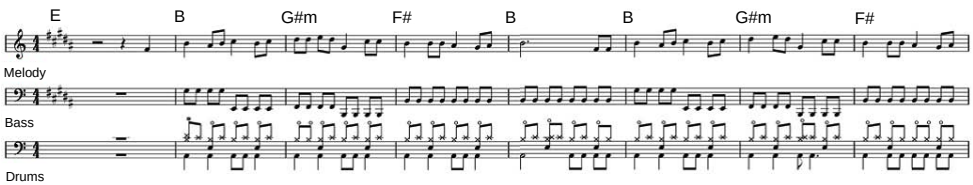


Fig. 6 Output example composed by proposed system with Rock database

Table 2 The Result of the Concordance Rate

		Subjects' answers				
		A	B	C	Unknown	Concordance rate
The genres of the used database for the learning	A (Jazz)	**16**	1	5	3	0.64
	B (Dance)	0	**22**	2	1	0.88
	C (Rock)	0	5	**19**	1	0.76

In the future, we intend to focus our research on the following.

1. Advancement of usability

The proposed system needs to have melody, chord, and section information as input. In the future it will automatically obtain the chord and section information by using pattern detection [13, 14], because the system will only need the melody.

2. Including other musical instruments

More instruments will be covered, e.g., strings, piano, and guitars.

3. Make the composed accompaniment be playable

In this paper, we had an assumption that the composed accompaniment was played by DAW software and the whether playable or not was not taken into consideration. In our future, we improve the proposed system to be able to compose the accompaniment that human being can play

with real musical instruments considering fingerings.

And we will also verify the usability of the proposed system to compose the original and affectivbe enclosed music for multimedia contents.

Acknowledgment

This work was supported in part by the Ministry of Education, Science, Sports and Culture, Grant. in. Aid for Scientific Research under grant #20700199, and HORI SCIENCE AND ART FOUNDATION.

References

[1] M. Balaban, K. Ebciogle, O. Laske: Understanding Music with AI: Perspectives on Music Cognition (AAAI Press, United States of America 1992)

[2] D. Cope: Computers and musical style (Computer Music and Digital Audio Series) (A-R Editions, Inc., United States of America 1991)

[3] D. Hörnel: A multi-scale neural-network model for learning and reproducing chorale variations, Computing in Musicology, volume 11, pp.141-158 (1998)

[4] N. Nettheim: Melodic pattern-detection using musearch in schubert's die schöe mülerin, Computing in Musicology 11, pp.159-168 (1998)

[5] M. Alfonseca, M. Cebrian, A. Ortega: A simple genetic algorithm for music generation by means of algorithmic information theory, Proceedings of 2007 IEEE Congress on Evolutionary Computation, pp. 3035-3042 (2007)

[6] S. Fukayama, K. Nakatsuma, S. Sako, Y. Yonebayashi, T. H. Kim, Q. S. Wei, T. Nishimoto, S. Sagayama: Orpheus: Automatic composition system considering prosody of Japanese lyrics, Entertainment Computing - ICEC 2009, pp.309-310 (2009)

[7] L. B. Meyer: Emotion and Meaning in music (University of Chicago Press, Chicago 1956)

[8] G. F. Cooper, E. Herskovits: A bayesian method for the induction of probabilistic networks from data. MACHINE LEARNING, volume 9, pp.309-347 (1991)

[9] J. J. Nattiez, C. Abbate: Music and Discourse: Toward a Semiology of Music/Translated from French (Princeton University Press, United States of America 1990)

[10] D. Stein: Engaging Music: Essays in Music Analysis (Oxford University Press, United States of America 2004)

[11] G. Ewer: Essential Chord Progressions (Pantomime Music Publications, United States of America 2006)

[12] R. Yamanishi, K. Akita, S. Kato: Automated composing system for sub-melody using hmm: A support system for composing music. Lecture notes in Computer Science, volume 6243**,** pp.425-427 (2010)

[13] S. Gulati, P. Rao: Rhythm pattern representations for tempo detection in music, Proceedings of International Conference on Intelligent Interactive Technologies and Multimedia, pp. 241-244 (2010)

[14] V. Zenz, A. Rauber: Automatic chord detection incorporating beat and key detection, Proceedings of 2007 IEEE International Conference on Signal Processing and Communications, pp. 1175-1178 (2007)

The Decision Model of Customer Segmentation Censoring

Hui-Hsin Huang[1,a]

[1] Department of Business Administration, Aletheia University

32Chen-li St., Tamsui, Departement of Business Administration, Aletheia University, Taipei, Taiwan

[a.]hoyasophia@gmail.com

Keyword: Bayesian model, Customer Segmentation Censoring, Light Buyer

Abstract. This paper provides a Bayesian sampling plane of intelligent computing on considering the exponential distribution as the monetary light buyers spend. To give the maintenance cost and profits obtainment in the loss function, the author demonstrates a decision rule to suggest marketing manager to accept or give up these light buyer segmentations.

Introduction

The intelligent computing techniques have been applied in marketing Management many years ago. But most of these researches are discussed the issue of predicting the customer lifetime value and finding the high contributions customer ([1]; [2]; [3]). This paper focuses on light buyer topics and uses intelligent computing technique to consider the segmentation of customer as a batch of product and bases on the industrial management concept to construct a decision model for the company to consider light buyer customers should be eliminated or maintained([4]).

The Models

The light buyers segmentation are the consumer whose purchase monetary are less([5]).Thus the author consider the monetary the light buyer spent is followed exponential distribution with the parameter λ.

$$f(x|\lambda) = \lambda \exp(-\lambda x) \tag{1}$$

And the parameter λ is followed the gamma distribution with the given parameters α、β

$$g(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda\beta) \tag{2}$$

Our decision function is

$$L(a, \lambda, n) = ah(\lambda) + (1-a)C_1 + nC_2 \tag{3}$$

In this equation, if a=1, the light buyer segmentation is accepted; and if a=0 , that means we will reject the light buyer segmentation.

Where the C_1 and C_2 are positive constants, C_1 is the cost if we reject the light buyer segmentation. That means C_1 is losing customer cost which is include losing the profit of customer and acquisition customer cost.

C_2 is the cost when we conduct survey to obtain the data of per sample. $h(\lambda)$ is the loss of accepting this light buyer segmentation and n is the sample size.

Using the loss $L(a, \lambda, n)$, the Bayes risk of a sampling plan (n) is

$$r(\delta|n) = E_{\Lambda} E_{\sum_{i=1}^{n} X_i|\Lambda} \left\{ C_1 + nC_2 + \delta\left(\sum_{i=1}^{n} X_i | n \right) [h(\Lambda) - C_1] \right\} = C_1 + nC_2 + r_1(\delta|n) \tag{4}$$

The Bayes decision function $\delta_B(|n)$ which minimizes $r_1(\delta|n)$ among all decision function $\delta_B(|n)$ is given by:

$$\delta_B\left(\sum_{i=1}^{n} x_i | n \right) = \begin{cases} 1 & if\ \varphi_g\left(n, \sum_{i=1}^{n} x_i \right) \le C_1 \\ 0 & otherwise \end{cases} \tag{5}$$

This paper considers rule of $h(\lambda)$ is following

$$h(\lambda) = a_0 + a_1\lambda + a_2\lambda^2 \tag{6}$$

The Bayes decision function

$$\delta_B\left(n, \sum_{i=1}^{n} x_i | n \right) = \begin{cases} 1 & if\ \varphi_g\left(n, \sum_{i=1}^{n} x_i \right) \le C_1 \\ 0 & otherwise \end{cases}$$

Note that if $C_1 \le a_0$ then $\varphi_g\left(n, \sum_{i=1}^{n} x_i \right) > C_1$ for all $\left(n, \sum_{i=1}^{n} x_i \right)$. Therefore $\delta_g\left(n, \sum_{i=1}^{n} x_i \right) \equiv 0$ .To avoid this extreme case, we assume that $C_1 > a_0$. It follows that $\delta_B\left(n, \sum_{i=1}^{n} x_i | n \right)$=1 if an only if

$$\sum_{i=1}^{n} x_i + \beta \ge \frac{a_1(n+\alpha) + \sqrt{a_1^{\,2}(n+\alpha)^2 + 4(C_1 - a_0)a_2(n+\alpha)(n+\alpha+1)}}{2(C_1 - a_0)} \tag{7}$$

Where x_i is the cumulative monetary amount which are spend by each customer in this light buyer segmentation. n is the sample size which is conduct by survey.
Then given the α、β, we can decision the best n based on Baye loss.

The data collection and model measurement

The paper obtains the data set of customer monetary transactions from a credit card company in April 2009 to May 2009. The total samples are 2,566 customers who make 10,804 transactions. The total monetary spending of these customer, $\sum_{i=1}^{n} x_i$ are 40,643,412 NT. dollars. $\bar{x}$ is 15,839.21 NT. dollars. Then we can estimate $p_0 = P(X > \lambda)$ =0.27. If we give the stander m_0 to be 2162 NT. dollars which is calculated by the meatiness cost of the credit card company, we can compute $R(m_0 = 2162)$ =0.754. According equation 14, a_0=-0.73, a_1=2162 and a_2=-2,337,122. In the case ofα=0, β=0 , then $D_n(n)$ =406.434. Because of

$$\sum_{i=1}^{n} x_i = 406.434 > D_n(n) = 0.231$$

Then the decision is to accept this segmentation.

Conclusion

In this paper the author proposes the customer censoring model for light buyer but this model not only be used for monetary estimation of light customer. It can be extent to calculate customer alive time and compute their lifetime value for marketing strategy.

Acknowledgments

The author thanks the supported by the Grant NSC NSC 99-2410-H-156 -013 of National Science Council of Taiwan, ROC.

Reference

[1] R. Colombo, W. Jiang: Journal of Interactive Marketing, 13(3), 2-12(1999)

[2] P. S, Fader, B. G. S. Hardie, K. L. Lee: Journal of Marketing Research, 42(4), 415-430 (2005)

[3] A. Weber: Target Marketing, 20(3), 72-75(1997)

[4] S. Knox: European Management Journal, 16(6), 729–737(1998)

[5] G. D. Morrison: Journal of Marketing Research, 3, 289-291(1966)

EEG-Based Brain Computer Interface for Game Control

Xing Song[1,a], S. Q. Xie[1,b], and K. C. Aw[1,c]

[1] Mechanical Department, the University of Auckland, Auckland, New Zealand 1010

[a] xson026@aucklanduni.ac.nz, [b] s.xie@auckland.ac.nz, [c] k.aw@auckland.ac.nz

Keywords: brain computer interface; BCI; electroencephalography; EEG; Hanning Window

Abstract. This paper describes the implementation of an EEG-based Brain Computer Interface (BCI) for game control using electroencephalography (EEG) rhythms. The signals were collected using a headset and sent to a laptop by employing Nordic 2.4 GHz wireless proprietary transceiver. After being denoised and classified using Fast Fourier Transform (FFT) and adaptive thresholds, these signals were translated to game control commands. These control commands are used to control actions of game characters in the video game "Super Street Fighter". The system is developed under the Emotiv platform and has been tested by five healthy participants. The results show that this platform can provide EEG-based BCI users more dimensions of control and bring them greater immersion into the game.

1. Introduction

A Brain Computer Interface (BCI) is a communication and control system in which human mind can be translated to the external world without the help of the normal output pathways of muscles and nerves [1]. For example, a BCI user can turn on/off a light or change TV channels by imaging a specific motion (e.g. imagining moving left/right arm) without any physical movement [2]. Recent advances in brain and BCI research reveal that BCIs can play a significant role in the future. They are found helpful in function restoring tools for supporting people with disabilities (e.g. to use BCI to control a prosthetic arm [3]) and game controller for serving game player with high interactivity (e.g. controlling game character by EEG-based BCI instead of joy stickers [4]). This field is still in its infant stage and most of the applications are proofs of concepts**.**

To access this interaction, various paradigms have been employed to collect the electrophysiological signals of human brain and to translate them into control commands. They are usually categorized into invasive [5], less invasive [6] and non-invasive [7]. The first method records signal patterns of a small group of neurons by implanting tiny electrodes into the brain [8]. The second one collects signals by putting sensors outside the skull under the scalp [9]. The third method avoids the risks of surgery by placing external sensors on the user's scalp [10]. Among various non-invasive BCIs, EEG-based BCI attracts most of the researchers' interest stemming from its cost-efficient and relatively short response time. Intensive training is important for the success of EEG-based BCI application [11], [12], and the biggest obstacle for effective training is the fatigue caused by the repeated boring operation during intensive training session [13].

An essential issue in EEG-based BCI development is how to solve the problem of the subjects' fatigue and help them to master the BCI operation quickly in training session. There are two adaptors in EEG-based BCI: user and system [1]. Both of them will adjust themselves to be better adapted to each other during the training session. The subject must modify his/her brain activities to maintain good correlation between his/her intent and the signal features used by the BCI. At the same time, the BCI must extract and classify the EEG signal features and must translate those features to device commands correctly. Repeating a simple operation with a great chance of failures

is a really boring task and usually causes the participants to feel fatigue. As with the acquisition of conventional skills, anxiety, frustration, or fatigue can degrade performance, particularly in the beginning stage [14], [15]. Previous researches reported that the keyboard variant was boring, while the BCI controlled games provided a more challenging and immersing experience, which helps the participants to enjoy the training sessions [16]. Thus various EEG-based games have been developed to improve the efficiency of BCI training, ranging from simple one dimension cursor movement to complex three dimensions robot control [17][18],[19].

The current state of the art of EEG-based game control can be divided into the following three major groups based on the features of the input signals. The first one is the feedback paradigm which uses the broadband frequency patterns of the brain as input for game control [20]. Usually, a function of band powers is introduced to calculate the relaxation of the player, which then be classified into a binary decisions. At the same time, Event Related Potentials (ERPs) such as P300 and Visually Evoked Potentials (VEPs) are also used as signal sources of game control [21]. Depending on the presentation of external stimuli (e.g. flashing), the ERPs can be used for game control with high transfer rates. The third one is imagined movement which totally employs motor activities as the input of game control [22]. According to previous researches, the brain activities during imagining specific movement are similar to that during actually doing the same movement [23]. Employing motor imagination as the control strategy, players will feel that the controlled game character is acting as the same as themselves. Hence, BCI games based on imagined movement can provide players greater immersion than other kinds of BCI games.

Pineda et al. [22] steered a first person shooter game using the *mu* rhythm power on motor cortices with 3 electrodes. The input of this game control was the difference in power between left and right cortices. Four participants, training for over 10 hours, learned to control the game quickly. Krepki et al. [24] used Lateralized Readiness Potential (LRP), which is a slow negative EEG shift over the activated motor cortex during a period of about 1 second before the actual movement onset, to control a Packman game. It is reported that the game player had the feeling that Packman moved in the correct direction though the player was not aware of his/her decision. Lehtonen et al. [25] also employed real finger movement for a simple BCI game: moving a ball toward the left or right side of the screen as indicated. By using actual finger movement, 3 trials were able to reach the target, as reported.

This paper presents an EEG-based BCI that can be used to steer game characters in “Super Street Fighter” to execute complex activities. The paper is organized as follows. Section II throws light on the hardware and software of the BCI, together with the brief introduction of the employed video game. The experiments and results are reported in section III. Finally, in section IV the main conclusions are summarized and possible future works are discussed.

2. BCI Setup

A BCI consists of input, signal processing, output and feedback. Fig. 1 shows the elements and principal workflow of a general EEG-based BCI. Electrical signals deriving from the brain are collected by electrodes on the scalp followed by amplifying and digitizing to get the raw EEG signals. The raw EEG signals are processed to extract specific signal features (e.g. amplitudes of evoked potentials or sensorimotor cortex rhythms) which reflect the subject’s intent. These features are translated into commands that operate a device (e.g. a word processing program, a game controller etc.). Feedback, in the form of auditory or vision, helps the user to modify his/her mind so as to operate the devices properly.

2.1 Signal Recording

The primary hardware used in this BCI is based on the Emotiv devices [26], consisting of a headset, a wireless dongle and a laptop. The headset integrates a bio-amplifier, 14 signal electrodes, 2 reference electrodes and a wireless transmitter. The wireless dongle sends digital signals, received from the headset, to the laptop via Nordic 2.4 GHz wireless proprietary transceiver, as shown in Fig. 2. Post-processing software runs on the laptop and transfers detection results to applications using the Application Programming Interface (API).

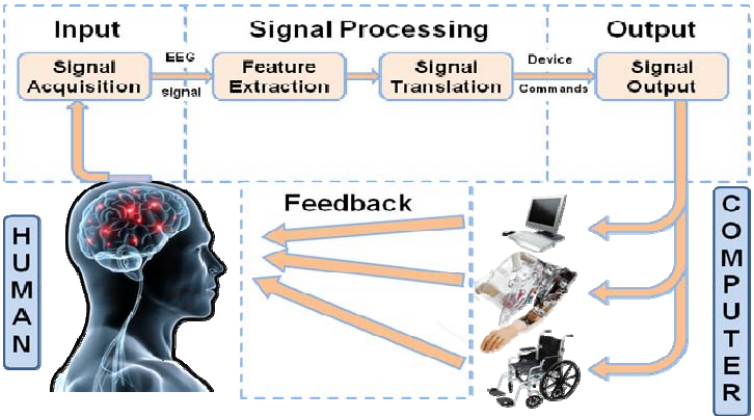


Fig. 1. Basic framework and operation flow of a general EEG-based BCI system.

The electrodes were placed according to the international 10-20 system with little modification, as shown in

Fig. **3**, where AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1, and O2 represent the signal electrodes, and the reference electrodes were represented by CMS and DRL, which are located on the bony bump just behind each ear lobe. Good contact of the electrodes to the scalp is essential for a good signal, thus self-test is used to indicate the status of electrode contact. Square waves are sent out periodically, and received by adjacent electrodes. The less distortion in received signals, the better contact quality will be got, which is displayed by showing different color of the corresponding electrodes.

Fig. 2. Primary hardware used in EEG-based BCI for entertainment.

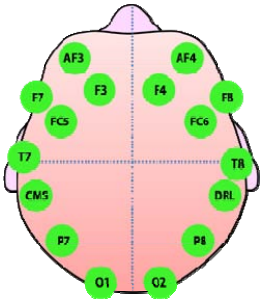


Fig. 3. The placement of electrodes in Emotiv headset.

2.2 Signal Processing

Signals recorded from scalp should be processed to provide meaningful messages or device commands. Such signal processing includes two stages: feature extraction and signal translation (or signal classification). Feature extraction (the value calculation of specific features of the signals) plays an important role in BCIs. As the first stage of signal processing, its quality has a significant

effect on the subsequent process, and consequently on the system as a whole. The second stage of signal processing is signal translation that classifies the features from the feature extraction stage into different groups represented various messages or device commands such as cursor movements, the typing of the words, or game control commands etc.

In this research case, digital EEG data is received via dongle which connects laptop with USB. The algorithm filters out the unrelated components using 2-40 Hz band-pass filter, then analyzes the filtered data with Fast Fourier Transform (FFT). In order to eliminate the picket fence effect (PFE) and to improve the side-lob rejection of the frequency response [27], the Hanning window is employed and the time-shifted forms of the windows are as follows:

$$\omega(n) = 0.5\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right)$$

and

$$\omega_0(n) = 0.5\left(1 + \cos\left(\frac{2\pi n}{N-1}\right)\right) \quad (1)$$

where N represents the width, in samples, of a discrete-time, symmetrical window function; n is an integer, with values $0 \le n \le N-1$; $\omega_0(n)$ is maximum at $n = 0$.

The frequency patterns of signals from electrodes F3 and F4 are taken for examples, as shown in Fig. 4. The variation of *delta* rhythm band (<4 Hz) is found during some emotion expression such as happiness, sadness and fear. The amplitude changes of *mu* rhythm (7~13 Hz) are related to the kind of imagined movement such as lifting and pushing. The brain activities are distinguished by adaptive thresholds for *delta* and *mu* rhythms respectively. The thresholds can be trained in training session or can be changed manually according to the performance of the user. According to the adaptive thresholds, the brain activities are labeled with different states and extent ranging from 0.0 to 1.0.

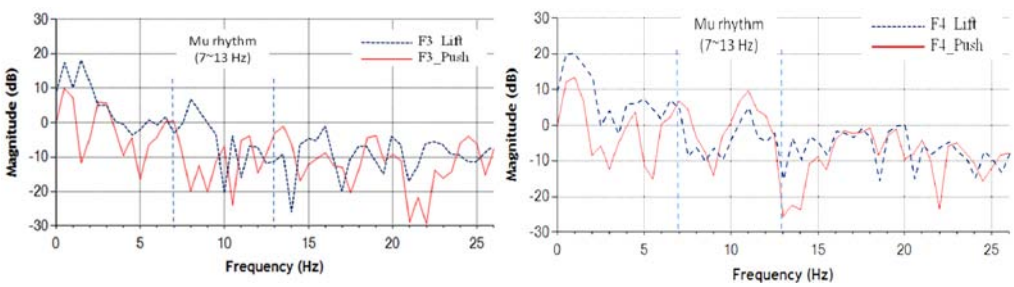


Fig. 4. The frequency pattern of signals from sites F3 (left) and F4 (right) during imagining the movement of lifting and pushing

2.3 Super Street Fighter Video Game

The game used in this paper as an example is “Super Street Fighter” which is a very popular electronic game among young people [28]. For the sake of the research, this game is adapted to a simpler version in flash format using the flash develop platform “Macromedia Studio 8” [29]. The number of game characters is reduced to three but all of its functions of fighting are reserved, such as moving forward, fisting, kicking and casting spells. Set in the easy mode, the computer-controlled game character will primarily defend, so the player has enough time to exercise the control skills of BCI. In a medium mode, this character will become more offensive, so the game is closer to a real game.

In traditional joystick control strategy, a game player should press several keys combination according to specific sequence in a short time (typically less than 1 second) to cast spells. It is impossible for current BCI to output several commands in this short time. Hence, a specially designed API is used to reach the target of casting spells. Once the API detects the command of casting spells, it will automatically translate this command to the specific keys combination in

demanded sequence, and then transfer them to the interface of “Super Street Fighter” one by one with 0.3 second intervals.

3. Experimental Test

3.1 Participants

The EEG signals were recorded from five subjects (3 males and 2 females) at the University of Auckland. The selected participants were between 26 - 28 years old with little experience using EEG-based BCI and 2 of them have never played this game before. Using the Edinburgh Handedness Inventory [30], all of the participants were right handed. All procedures were approved by a local ethics committee and informed consent was given by all participants.

3.2 Pre-training Session

Pre-training session is designed to help participants to familiarize themselves with and to master the basic operation of the EEG-based BCI. This session includes two following stages.

Thresholds calibration: Participants sat in front of a laptop, wearing the Emotiv headset while resting their arms on a table, as shown in Fig. 2. At the beginning of the thresholds calibration, participants were instructed to relax thoroughly and remain in this status for about 30 seconds during which the system recorded the background brain activities. This status was labeled “neutral”, which would be used as the reference for other brain activities. Then participants were instructed by an animation to imagine lifting, pushing and moving left/right hands without actual movement. If the participant was satisfied with the performance of his/her imagination, the frequency pattern would be recorded to train the Artificial Neural Network (ANN). The outputs of the ANN were the trained thresholds for imagined movement. The above process was repeated to record the thresholds for the emotion expression.

Imagined Writing: This sub-session was designed to test the previous training and to improve the participants’ skills on operation of BCI. Several characters were showing on the screen of a laptop. Each of them represented a specific movement or emotion expression. Participants were asked to rewrite the provided characters by imagining movement or expressing emotion, which was represented by that character.

3.3 Game Strategy

In this game, both emotion expression and motor imagery are employed as the inputs of game control. The imaginations of moving left/right hands and lifting an object up/down are translated to the commands of moving left, moving right, jumping or creeping respectively. The changes of player’s emotion will trig other more complex commands. For example, “angry” will steer the BCI control character to attack continuously and cast super offensive spells, but “fear” will cause the character to act more defensively and try to run away from the enemy.

Participants firstly played the game in easy mode. In this mode, the computer controlled game character is weak and the time for each round is unlimited. A round will be ended if one player’s blood indicator is zero. In medium mode, the computer controlled game character is stronger and the time for each round is limited to 99 seconds. If one player is defeated or the time runs out, a round will be ended and the player with more blood will be the winner.

3.4 Results

After 40 minutes of pre-training session, all of the five participants can successfully master the control of this BCI. In easy mode, all of them could win this game and there was no significant change in the results, except the lasting time for each round. A screenshot of “Super Street Fighter” is shown in Fig. 5.

Fig. 5. "Super Street Fighter" screenshot

Three participants, who had experience of playing this game with joysticks before, won this game in medium mode. Other two participants, having no experience of playing this game with joysticks before, had difficulty to fight back under the offensively attack of computer controlled game character.

In addition, participants reported that the pleasure from the game can significantly reduce the fatigue during the boring training. Immersing into the game, they found using this EEG-based BCI was a pleasure instead of tedious workload. At the same time, the two participants, who had never played similar games, reported that they had no interest in joystick-based "Super Street Fighter" but really enjoyed this EEG-based one.

4. Conclusion and Future Work

In this paper, a video game controlled by an EEG-based BCI is demonstrated. The participants learned to control their *delta* and *mu* rhythms by expressing specific emotion and imagining arms movements. They are able to master the basic operation of BCI after several short pre-training and maintain this level of control in the following imagined writing test and game control. The results of the experiment show that the combination of the emotion and motor imagery can provide BCI users more dimensions of control. The advantage offered will enable players to control a game character to do more complex actions via their brain activity, which can bring game players greater immersion into the game. At the same time, it is reported that this entertainment implementation makes the previously boring training session to be enjoyable.

This implementation of EEG-based BCI game controller has three positive effects on the BCI training session. Firstly, the EEG signals are transmitted with wireless digital data chain. Without data cable between participants and laptop, there is a more freedom for participants to choose a more comfortable posture while playing the game. In addition, digital data is more robust to the induced noises during the signal transmission. Secondly, employing Hanning window, FFT and adaptive thresholds can significantly reduce the computation and gives low delays in the control signals, which is critical for real time game control, especially for current popular mobile embedded system. Thirdly, the combination of emotion and imagined movement provides participants greater immersion into the game and provides them more options to send out control commands for complex actions. Although emotion expressing is not a kind of motor imagination, it is still helpful to assist participants to complete the training session effectively and to master the skills of BCI operation quickly.

This study has raised several questions for further exploration. Firstly, how to solve the share of computation power in multiplayer mode is an essential issue for most of current electronic games. This implementation of "Super Street Fighter" is demonstrated in single player mode, and only one calculating core is enough for signal processing which is computationally expensive. If the number of players was extended to two or more, the demand of computational power will significantly increase. Hence, how to share the computation source is important for meeting the demand of real

time game. Secondly, is it possible to develop a more user-friendly interface of the game itself instead of using the conventional control strategy? Currently, game characters are controlled by the mapping of keyboard and mouse. Using this strategy in BCI-based game control, there are two kinds of mapping: EEG signals mapping to keys and keys mapping to the actions of game characters. Mapping EEG signals directly to actions of game characters will reduce the calculation workload and offer the game player freedom to control the game character using natural behavior. This can not only increase the immersion of game for game players, but also can serve as an effective training tool for the users of BCI-based rehabilitation robots and prosthetics.

There are also some enhancements needed to be explored in the future. For instance, we can broaden the band of EEG from *mu* and *beta* to higher bands such as *gamma* band. On the other hand, universal Application Programming Interface (API) should be developed to enable this EEG-based BCI to be transplanted among different systems and games. In addition, the algorithm of the signal processing should be further optimized to reduce the command delay.

5. Acknowledgment

This work was supported in part by the Chinese Scholarship Council (CSC). Authors would like to thank Mr. Oliver Grant of the University of Auckland for his support on script proofreading.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, “Brain-computer interfaces for communication and control,” *Clinical Neurophysiology*, 113, vol.6, 2002, pp. 767-791.

[2] B. Allison, and J. Jacko, “The I of BCIs: Next Generation Interfaces for Brain–Computer Interface Systems That Adapt to Individual Users Human-Computer Interaction,” *Novel Interaction Methods and Techniques*, vol. 5611, 2009, pp. 558-568.

[3] M. A. Lebedev, J. M. Carmena, J. E. O’Doherty, M. Zacksenhouse, C.S. Henriquez, and M. A. Nicolelis, “Cortical ensemble adaptation to represent velocity of an artificial actuator controlled by a brainmachine interfaces,” *Neuroscience*, vol. 25(19), 2005, pp. 4681-4693.

[4] E. C. Lalor, S. P. Kelly, C. Finucane, R. Burke, R. B. Reilly, and G. Mcdarby, “Brain Computer Interface based on the Steady-State VEP for immersive gaming control,” *Biomedizinische Technik*, vol. 49(1), 2004, pp. 63-64.

[5] D. M. Taylor, S. I. H. Tillery, & A. B. Schwartz, “Direct Cortical Control of 3D Neuroprosthetic Devices,” *Science*, vol. 296(5574), 2002, pp. 1829-1832.

[6] E. Margalit, J. D. Weiland, R. E. Clatterbuck, G. Y. Fujii, M. Maia, and M. Tameesh, "Visual and electrical evoked response recorded from subdural electrodes implanted above the visual cortex in normal dogs under two methods of anesthesia," *Journal of Neuroscience Methods*, vol. 123(2), 2003, pp. 129-137.

[7] F. Wallois, A. Patil, C. Héberlé, and R. Grebe, "EEG-NIRS in epilepsy in children and neonates," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 40(5-6), 2010, pp. 281-292.

[8] R. A. Andersen, S. Musallam, B. Pesaran, "Selecting the signals for a brain-machine interface," *Current Opinion in Neurobiology*, vol. 14(6), 2004, pp. 720-726.

[9] E. C. Leuthardt, G. Schalk, J. R. Wolpaw, J. G. Ojemann, D. W. Moran, "A brain-computer interface using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 1(2), 2004, pp. 63-71.

[10] M. V. Gerven, J. Farquhar, R. Schaefer, R. Vlek, J. Geuze, A. Nijholt, N. Ramsey, P. Haselager, L. Vuurpijl, S. Gielen, and P. Desain, "The brain-computer interface cycle," *Journal of Neural Engineering*, vol. 6(4), 2009, pp. 041001-041010.

[11] M. Siniatchkin, P. Kropp, and W. D. Gerber, "Neurofeedback—The Significance of Reinforcement and the Search for an Appropriate Strategy for the Success of Self-regulation," *Applied Psychophysiology and Biofeedback*, vol. 25(3), 2000, pp. 167-175.
[12] J. R. Wolpaw, and D. J. McFarland, "Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101(51), 2004, pp. 17849-17854.
[13] E. E. Sutter, "The brain response interface: communication through visually-induced electrical brain responses," *Journal of Microcomputer Applications*, vol. 15(1), 1992, pp. 31-45.
[14] P. M. Dibartolo, T.A. Brown, and D.H. Barlow, "Effects of anxiety on attentional allocation and task performance: an information processing analysis," *Behaviour Research and Therapy*, vol. 35, 1997, pp. 1101–1111.
[15] O. U. Soyuer, G. Turanli, D. Yalnizoglu, E. E. Bakar, M. Topcu, "Classification and follow-up of pediatric patients with absence epilepsy," *Epilepsia*, vol. 47, 2006, pp. 152-152.
[16] D. Bussink, "Towards the first HMI BCI game," *Master's Thesis*, University of Twente, March 2008.
[17] J. Allanson and J. Mariani, "Mind over Virtual Matter: Using Virtual Environments for Neurofeedback Training," *Virtual Reality Annual International Symposium Proceedings of the 1999 IEEE Virtual Reality*, 1999, pp. 270-273.
[18] J. D. Bayliss, "Use of the evoked potential P3 component for control in a virtual apartment," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on Rehabilitation Engineering*, vol. 11(2), 2003, pp. 113-116.
[19] P. Martinez, H. Bakardjian, and A. Cichocki, "Fully online multicommand brain computer interface with visual neurofeedback using SSVEP paradigm," *Computational Intelligence and Neuroscience*, vol. 2007(1), 2007, pp. 13.
[20] L. Minyu, G. Mcmillan, G. Calhoun, and K. S. Jones, "Development of EEG Biofeedback System based on Virtual Reality Environment," *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005, 27th Annual International Conference of the*, pp. 5362-5364.
[21] E. C. Lalor, S. P. Kelly, C. Finucane, R. Burke, R. Smith, R. B. Reilly, and G. Mcdarby, "Steady State VEP based Brain Computer Interface control in an immersive 3D gaming environment," *EURASIP Journal on Applied Signal Processing*, vol. 2005(19), 2005, pp. 3156-3164.
[22] J. A. Pineda, D. S. Silverman, A. Vankov, and J. Hestenes, "Learning to control brain rhythms: making a brain computer interface possible," *Neural Systems and Rehabilitation Engineering, IEEE Transaction on Rehabilitation Engineering*, vol. 11(2), 2003, pp.181-184.
[23] O. Bai, P. Lin and et al, "A high performance sensorimotor beta rhythm-based brain–computer interface associated with human natural motor behavior." *Journal of Neural Engineering*, vol. 5(1), 2008, pp. 24.
[24] T. Kayagil, O. Bai, P. Lin, S. Furlani, S. Vorbach, and M. Hallett, "Binary EEG control for two-dimentional cursor movement: an online approach," *Complex Medical Engineering, 2007, CME 2007, IEEE/ICME International Conference on*, pp. 1542-1545.
[25] J. Lehtonen, P. Jylanki, L. Kauhanen, and M. Sams, "Online classification of single EEG trials during finger movements," *Biomedical Engineering, IEEE Transactions on*, vol. 55(2), 2008, pp. 713-720.
[26] "Emotiv systems," *Website*, Aug, 2011, http://emotiv.com.
[27] A. Nuttall, "Some windows with very good sidelobe behavior," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29(1), 1981, pp. 84-91.
[28] "Street Fighter," *Website*, Aug, 2011, http://www.streetfighter.com.
[29] "Macromedia Studio 8," *Website*, Aug, 2011, http://www.adobe.com.
[30] R. C. Oldfield, "Assessment and Analysis of Handedness - Edinburgh Inventory," *Neuropsychologia*, vol. 9(1), 1971, pp. 97-113.

Application of PARSEC Geometry Representation to General Airfoil for Aerodynamic Optimization

R. Mukesh [1, a], Dr.K. Lingadurai [2, b], A. Muruganandham [3, c]

[1] Department of Mechanical Engineering, Anna University of Technology, Dindigul, India

[2] Associate Professor, Department of Mechanical Engineering, Anna University of Technology, Dindigul, India.

[3] Associate Professor, Department of ECE, Sona College of Technology, Salem, India

[a] pr.mukeshphd@gmail.com, [b] lingadurai@gmail.com, [c] muruga_salem@rediffmail.com

Keywords: Aerodynamic Shape Optimization, Parametric Section, PSO, MDO.

Abstract. Generally if we want to optimize an airfoil we have to describe the airfoil and for that, we need to have at least hundred points of x and y co-ordinates. It is really difficult to optimize airfoils with this large number of co-ordinates. Nowadays many different schemes of parameter sets are used to describe general airfoil such as B-spline, Hicks- Henne Bump function, PARSEC etc. The main goal of these parameterization schemes is to reduce the number of needed parameters as few as possible while controlling the important aerodynamic features effectively. Here the work has been done on the PARSEC geometry representation method. The objective of this work is to introduce the knowledge of describing general airfoil using twelve parameters by representing its shape as a polynomial function. And also we have introduced the concept of Particle Swarm Optimization (PSO) to optimize the aerodynamic characteristics of a general airfoil for specific conditions. A MATLAB program has been developed to implement PARSEC, Panel Technique and PSO. This program has been tested for a standard NACA 2411 airfoil and optimized to improve its coefficient of lift.

Introduction

In the context of aerodynamic shape optimization, a need arises to represent a general three dimensional surface by means of minimum number of design parameters, which can serve as the optimization parameters to arrive at the optimum shape. It is due to that when using computational design optimization, a too large set of design variables would lead to excessive computation time to search the design space. Partial differential equation approach (time consuming and not suitable for multidisciplinary design optimization), discrete points approach (number of design variables becomes large) and polynomial approach (number of design parameters depends on the degree of the polynomial chosen and suitable for multidisciplinary design optimization) are the three basic approaches to describe the geometry of a general airfoil. The best optimum design can be obtained depends on the parameterization scheme. So the selection of parameterization scheme is very important for design optimization. The PARSEC parameterization scheme is coupled with PSO algorithm to achieve the goal of getting the optimum aerodynamic shape of NACA 2411 airfoil.

PARSEC

In PARSEC [1, 2, 3] parameterization scheme an unknown linear combination of suitable base function is used to describe the geometry of an airfoil. Twelve design variables are selected to have direct control over the shape of the airfoil. The twelve control variables are, Upper leading edge radius (R_{leu}), Lower leading edge radius (R_{lel}) , Upper crest point (Y_{up}), Lower crest point (Y_{lo}), Position of upper crest (X_{up}), Position of lower crest (X_{lo}) Upper crest curvature (YXX_{up}), Lower crest curvature (YXX_{lo}), Trailing edge offset (T_{off}), Trailing edge thickness (T_{TE}), Trailing edge direction angle (α_{TE}), Trailing edge wedge angle (β_{TE}), as shown in Fig. 1. The mathematical formulation for PARSEC is given by the polynomial,

$$y_u = \sum_{i=1}^{6} a_i x^{i-(1/2)} \tag{1}$$

$$y_l = \sum_{i=1}^{6} b_i x^{i-(1/2)} \tag{2}$$

for upper and lower surface respectively. Where y_u is the required y coordinate for the upper surface, y_l is the required y coordinate for the lower surface and a_i, b_i are the coefficients to be solved from the twelve control variables.

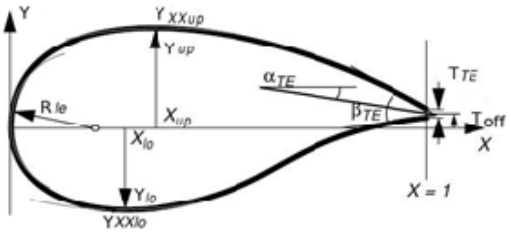


Figure 1. Control variables for PARSEC

Panel Technique

The solution procedure for panel technique consists of discretizing the surface of the airfoil into straight line segments or panels, assuming the source strength is constant over each panel but has a different value for each panel and the vortex strength is constant and equal over each panel [3, 4, 5]. The compressibility and the viscosity of air in the flow field is neglected, and the net effect of viscosity on a wing is summarized by requiring that the flow leave the sharp trailing edge of the wing smoothly. The curl of the velocity field is assumed to be zero. Here,

$$\phi = \phi_\infty + \phi_\delta + \phi_\upsilon \tag{3}$$

where ∅ is the total potential function and its three components are the potentials corresponding to the free stream, the source distribution, and the vortex distribution. These last two distributions have potentially locally varying strengths. Fig. 2 illustrates the nomenclature of an airfoil and the definition of nodes and panels for panel methods respectively.

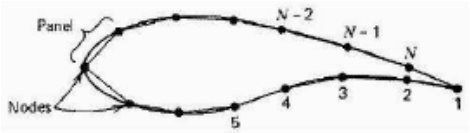


Figure 2. Nodes and Panels

The numbering system starts at the lower surface trailing edge and proceeds forward, around the leading surface and aft to the upper surface trailing edge. N+1 points define N panels. The flow tangency boundary condition is imposed on the points located at the midpoint of each of the panels. Once we found the tangential velocity (V_{ti}) at the midpoint of each panel, then we can compute the pressure coefficient at the midpoint of each panel according to the following formula,

$$C_p(x_i, y_i) = 1 - [V_{ti}^2 / V_\infty^2] \tag{4}$$

Particle Swarm Optimization

PSO is a population-based algorithm for searching global optimum. It ties to artificial life, like fish schooling or bird flocking, and has some common features of evolutionary computation such as fitness evaluation. The original idea of PSO is to simulate a simplified social behavior [6, 7]. Similar to the crossover operation of the GA, in PSO the particles are adjusted toward the best individual experience (PBEST) and the best social experience (GBEST). However, PSO is unlike a GA in that each potential solution, particle is ‘‘flying” through hyperspace with a velocity. Moreover, the particles and the swarm have memory; in the population of the GA memory does not exist.

Let $x_{j,d}(t)$ and $v_{j,d}(t)$ denote the d^{th} dimensional value of the vector of position and velocity of j^{th} particle in the swarm, respectively, at time t. The PSO model can be expressed as

$$v_{j,d}(t) = v_{j,d}(t-1) + c_1.\varphi_1.(x^*_{j,d} - x_{j,d}(t-1)) + c_2.\varphi_2.(x^{\#}_d - x_{j,d}(t-1)), \tag{5}$$

$$x_{j,d}(t) = x_{j,d}(t-1) + v_{j,d}(t), \tag{6}$$

where $x^*_{j,d}$ (PBEST) denotes the best position of j^{th} particle up to time t-1 and $x^{\#}_d$ (GBEST) denotes the best position of the whole swarm up to time t-1, φ_1 and φ_2 are random numbers, and c_1 and c_2 represent the individuality and sociality coefficients, respectively.

The population size is first determined, and the velocity and position of each particle are initialized. Each particle moves according to (5) and (6), and the fitness is then calculated. Meanwhile, the best positions of each swarm and particles are recorded. Finally, as the stopping criterion is satisfied, the best position of the swarm is the final solution. The main steps of PSO are given as follows:

a) Set the swarm size. Initialize the velocity and the position of each particle randomly.
b) For each j, evaluate the fitness value of x_j and update the individual best position $x^*_{j,d}$ if better fitness is found.
c) Find the new best position of the whole swarm. Update the swarm best position $x^{\#}$ if the fitness of the new best position is better than that of the previous swarm.
d) If the stopping criterion is satisfied, then stop.

e) For each particle, update the position and the velocity according (6) and (5). Go to step b.

Optimization of NACA 2411 Airfoil

The design conditions, optimization objectives and constraints are tabulated in Table I.

Table I. Optimization objectives and constraints

Angle of attack	5.0 deg
Flow constraint	Subsonic and incompressible
Geometric constraint	Max thickness must be less than 10% chord length
	T_{TE} and T_{off} the airfoil is zero
Aerodynamic constraint	Lift not less than original one
Objective	Maximize coefficient of lift

Results

The initial PARSEC parameters have been given approximately by specifying its lower and upper bound values. There is no need for specifying this accurately. The geometry of the airfoil expressed by the best twelve PARSEC parameters resulting from the PSO algorithm exhibits a considerable increase in the coefficient of lift. The comparison between the original NACA 2411 airfoil and the optimized airfoil is indicated in Fig. 3. The comparison of pressure distribution over the surface of the original NACA 2411 airfoil and the optimized airfoil is shown in Fig. 4. Their corresponding PARSEC parameters and coefficient of lift are tabulated in Table II and Table III respectively.

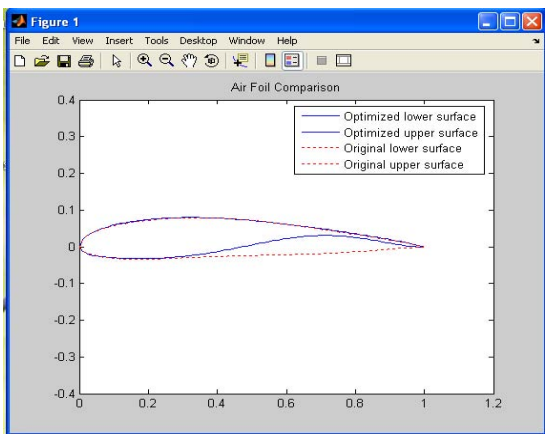


Figure 3. Original NACA 2411 airfoil Vs. Optimized airfoil

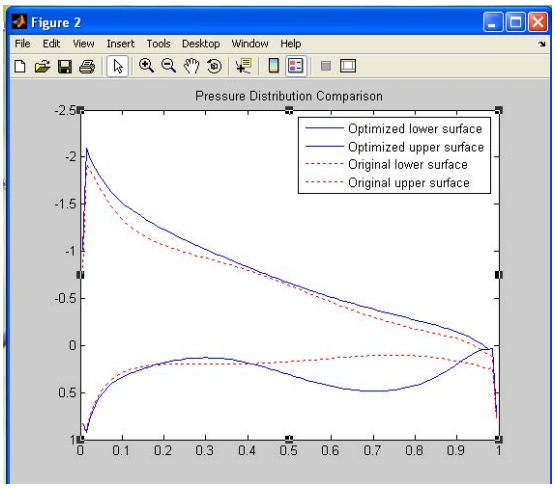


Figure 4. Comparison of pressure distribution over the surface of original NACA 2411 airfoil and Optimized airfoil

Table II. Optimized PARSEC parameters

Parameter	Value original	Value optimized
(Rleu) Upper leading edge radius	0.0216	0.020153
(Rlel) Lower leading edge radius	0.008	0.0095635
(Xup) Position of upper crest	0.3445	0.32033
(Yup) Upper crest point	0.07912	0.079734
(YXXup) Upper crest curvature	-0.6448	-0.63414
(Xlo) Position of lower crest	0.17	0.17371
(Ylo) Lower crest point	-0.033797	-0.032287
(YXXlo) Lower crest curvature	0.6748	0.67749
(TTE) Trailing edge thickness	0	0
(Toff) Trailing edge offset	0	0
(αTE) Trailing edge direction angle	-4.785	-4.7811
(βTE) Trailing edge wedge angle	15.082	15

Table III Original vs. Optimized Coefficient of Lift

Angle of attack	$Cl_{original}$	$Cl_{optimized}$
5.0	0.8420	1.0352

Conclusions

The geometry of NACA 2411 airfoil is optimized to improve its coefficient of lift for 5.0 deg angle of attack. The optimized airfoil has the improved coefficient of lift of 1.0352 as compared to the original one which had 0.8420. The PARSEC parameterization scheme is used to express the shape of the airfoil. The result shows that the PARSEC parameters show good correlation between design parameters and aerodynamic performance. The impact of individual PARSEC parameters on the aerodynamic properties of the airfoil can be predicted more easily. There is no need for baseline shape and typical geometric constrains on the airfoil shape can be expressed or approximated by simple bound or linear constrains. The panel method gives reasonable accuracy in predicting the pressure distribution over the surface of the NACA 2411 airfoil for low speed, incompressible subsonic flows. PSO algorithm is so effective in finding the best solution among many possible solutions within a search space. During this problem of optimization plenty of design data are obtained. It is possible to create a model from these design data. The potential application of this process is that it acts as a data-mining process to increase design knowledge

References

[1] R. Balu and U. Selvakumar, “Optimum hierarchical Bezier parameterization of arbitrary curves and surfaces,” 11th Annual CFD Symposium, Indian Institute of Science, Bangalore, India, pp. 46-48, August 2009.

[2] Helmut Sobieczky, “Parametric airfoils and wings”, Notes on Numerical Fluid Mechanics, Volume 68 Vieweg verlag, pp. 71-88, 1998.

[3] P.R. Mukesh, U.Selvakumar. Aerodynamic Shape Optimization using Computer Mapping of Natural Evolution Process. In: Proceedings of the 2010 International Conference on Mechanical and Aerospace Engineering at University of Electronics Science and Technology of China, China, April 16-18, 2010.Volume 5, pp 367 - 371

[4] J.L. Hess, “Panel methods in computational fluid mechanics”,Annual Review of Fluid Mechanics, Vol.22, pp. 255-274, 1990.

[5] J. Katz and A. Plotkin, “Low-speed aerodynamics from wing theory to panel methods”, McGraw-Hill, Inc., New York, 1991.

[6] J. Kennedy, R.C. Eberhart, Particle swarm optimization, in: Proceedings of IEEE International Conference on Neural Networks, Perth, Australia, vol. 4, 1995, pp. 1942–1948.

[7] R.C. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in: Proceedings of IEEE International Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995, pp. 39–43.

Application of Ant Colony Algorithm in Emotion Clustering of EEG Signal

Liu Hongxia[1, a], Wu Guowen[1, b] and Luo Xin[1, c]

[1]School of Computer Science and Technology

Donghua University

Shanghai, China

[a]hx_0223@163.com, [b]gwwu_dh@163.com, [c]xluo@dhu.edu.cn

Keywords: EEG; Ant Colony Algorithm; Clustering; Emotion State.

Abstract. With the development of image retrieval, rational organization and management of image database play a critical role in image retrieval. In order to raise efficiency of the image retrieval, ant colony algorithm is presented in emotion clustering of EEG signal after it was improved. This algorithm determines the initial ants by specific samples and simulates the process of Cemetery Organization. It is used to clustering emotion features extracted from EEGs. Computer simulation shows that the improved algorithm has better clustering results and accuracy rate.

1. Introduction

Emotion state extraction is important to human intelligence. Research of emotional image retrieval is one of the key issues in the development of information technology. Not only the way to collect global information about human's mental activities, but also emotion states extraction and qualification are huge challenges to the ability of computer retrieving images according to human understanding.

There are many sources to extract emotional states of the images, such as color, distribution of color, color saturation and brightness. With the deepening of studies of physiological signals, technology of emotion extraction directly from physiological signals is developed rapidly. Eckman and Friesen [1] described the six basic facial expressions and combined these components to explain 33 more complex emotions. Toshimitsu [2] proved that emotional state is decomposed into more elementary states, ten electrodes are used and the four elementary states, anger, sadness, joy, and relaxation, are adopted, and get better effect.

Researching brain activity from scalp potentials, proper signal processing to extract emotional states and quantify these characteristics, and emotional image clustering are prerequisites of image retrieval based on emotional semantic. Currently, many scientists have studied a good deal of approaches to clustering, such as genetic algorithm, immune algorithm [3]. However, mostly the methods are only suitable to certain problems, and there are still drawbacks needing to be solved in current cluster analysis, such as weak ability to handle different types of properties, needing to set the input parameters according to priori knowledge, bad scalability and so on.

Based on the principle of foraging ants, improved clustering analysis based on ant colony algorithm is applied to cluster analysis of EEG signal for the first time. Images are marked emotion vectors before they are emotional clustered. This allows users to retrieve emotional information more rapidly and accurately.

2. Extract Emotion Feature from EEGs

Scalp potentials, measured by an electroencephalograph(EEG), are rich in information about brain activity. Processing EEGs accurately and efficiently is significant to the research of emotion recognition. The present technique is called the emotion spectrum analysis method, or ESAM. Four elementary emotional states are extracted from EEGs, which are anger, sadness, joy, and mental relaxation. They are expressed as the following values of the emotion vector, $z = (1, 0, 0, 0)^T$,$(0, 1, 0, 0)^T$,$(0, 0, 1, 0)^T$ and $(0, 0, 0, 1)^T$, respectively. Besides, the one $(0, 0, 0, 0)^T$ is considered to be the control state in which no special emotion was activated in the subjects.

2.1 Data Preprocessing

The physiological data of Electroencephalogram signals is collected with laboratory equipment, produced by Japan, dedicated to the analysis of EEG brain waves. Ten disk electrodes are placed on the scalp at positions, FP1, FP2, F3, F4, T3, T4, P3, P4, O1, and O2 according to the International 10-20 Standard, and EEGs were recorded with a reference electrode on the right earlobe. The subjects are some students who are 22 years old, healthy, without any history of mental illness and brain damage. The experimental material is some emotional images.

Emotional data of images are obtained in the following way. Images are placed in groups of five. During the experiments, the subjects frequently close their eyes for resting every few seconds to ensure the accuracy of data obtained. Those subjects watched the picture presented before them, 5.12-s EEG segments were cut out from their ten-channel EEGs. And after that, pictures with which the previous one is paired are presented in order and EEG segments were collected in the same way. This process was repeated for the other groups of images.

2.2 Emotion Feature Extraction

The emotion vectors of EEGs features are extracted in the following ways. The segments cut out from their ten-channel EEGs are evaluated with cross-correlation coefficients on 45 channel pairs. The electric potentials are sampled at 100 Hz, and then separated in the theta (5-8 Hz), alpha (8-13 Hz) and beta (13-20 Hz) frequency bands by means of FFT. Totally, 135 such variables are obtained. The set of these 135 variables forms the 135-vector $y(y1,y2,\ldots,y135)^T$. Then the set of those 135 variables associated with emotional states can be transformed into the 4-vector $z = (z1, z2, z3, z4)$ by operating a transformation in the neural network.

The neural network is prepared in the following ways. Firstly, ten people, well trained in mental imaging, involve in preparatory work. They first image anger in their minds, and 5.12 seconds EEG segments were cut out from their ten-channel EEGs, and then they were transformed into 135-dimensional vector. This process was repeated for the other three elementary emotional states of sadness, joy, and relaxation. Secondly, an artificial neural network is established with the 135-dimensional vector as its input vectors and emotion vector as its output vector. The neural network is trained by the data collected on previous step.

3. Clustering Algorithm Based on Improved Ant Colony

Biologists found that many ant species, such as Lasius niger and Pheidole pallidula, exhibit the behavior of clustering corpses to form cemeteries. Each ant seems to move randomly in space, picking up or depositing corpses based on the current location of the local ant information. This simple behavior results in the emergence of a complex behavior of cluster formation. Based on this theory, algorithms are implemented to cluster emotion data of images to effectively improve the retrieval efficiency.

3.1 Generalized Ant Colony Clustering Model

The emotion vector of each image in the database corresponds to an ant. Assume there are two images

which emotion vectors are y_a and y_b. In general, the Euclidean distance $d(y_a, y_b)$ between data vector y_a and y_b is used most frequently to quantify dissimilarity. The clusters are considered as the cemeteries of corpses formed by the ants. In other words, the process of data objects clustering is the course of ant's cemeteries forming.

Clustering partitions a dataset $X = \{y_i | i = 1, 2, \cdots, n\}, y_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ into $c \in \{2, \cdots, n-1\}$ subsets (clusters). We consider y_i as the emotion state of images. Assume that an ant is on site i at the time t, and finds data vector y_a. The "local" density, $\lambda(y_a)$, of data vector y_a within the ant's neighborhood is then given as

$$\lambda(y_a) = \max\left\{0, \frac{1}{n_\varkappa^2}\sum_{y_b \epsilon N_{n_\varkappa \times n_\varkappa}(i)}\left(1 - \frac{d(y_a, y_b)}{\gamma}\right)\right\} \quad (1)$$

where $N_{n_\varkappa \times n_\varkappa}(i)$ is a square neighborhood current position of the ant, $d(y_a, y_b)$ is the Euclidean distance between vector y_a and y_b, and the constant γ defines the scale of dissimilarity between the two vectors.

Using the measure of similarity, $\lambda(y_a)$, the picking up and dropping probabilities are defined as

$$P_p(y_a) = \left(\frac{\gamma_1}{\gamma_1 + \lambda(y_a)}\right)^2 \quad (2)$$

$$P_d(y_a) = \begin{cases} 2\lambda(y_a) & \text{if } \lambda(y_a) < \gamma_2 \\ 1 & \text{if } \lambda(y_a) \geq \gamma_2 \end{cases} \quad (3)$$

Consider an ant moving an item, y_a. If there are some similar individuals in the square neighborhood of the $n_\varkappa \times n_\varkappa$ sites, then the values of $\lambda(y_a)$ approaches 0, and the probability of dropping the item is high. If there is no item nearby, then the values of $\lambda(y_a)$ approaches 1, and the probability of dropping the item is small. If it is an unladen ant, the probability of picking up the item is high.

3.2 Improved Ant Colony Algorithm

In basic ant colony clustering algorithm, the vectors are randomly placed on the grid, and the clusters are formed by local optimization. There are some deficiencies when clusters are formed by this method.

a) The algorithm does have the tendency to create more clusters than necessary. The cluster center is not stable and it is difficult to achieve optimal clustering. If two clusters A and B are high similarity, but A is far from B, these two parts are hard to be formed into a cluster. So measures should be taken to maximize the inter-cluster distances; that is, the different clusters should be well separated. In the beginning, place some items which trained the neural network on the grid in the form of clustering. These original clusters ensure that inter-distances are maximized and intra-distances are minimized.

b) In the Euclidean distance to quantify dissimilarity, each vector component plays the same role in the result. The whole process of getting emotion vectors is assumed to be ideal. Therefore, it's necessary to adopt appropriate methods to reduce the impact of deviation. We add a weight $\frac{x_{ik}}{\sum x_{ik}}$ for each variable of Euclidean distance to strengthen the role of the main emotion vector component and weaken the effect of noise.

Let n_t be the maximum number of iterations. A summary of the Lumer-Faieta ant colony clustering algorithm is given as follows.

a) Initialize values of $\gamma 1, \gamma 2, \gamma$ and n_t;
b) Place special data vector on a grid and form initial cluster;
c) Place each data vector y_a randomly on a grid and place n_k ants on randomly selected sites;
d) Initialize all ants clustering information and start a new round of operation.
e) Consider the ant n_i. According to neighborhood information, calculate the "local" density $\lambda(y_a)$. If ant n_i is unladen and the site is occupied by item y_a, compute $\lambda(y_a)$ and $P_p(y_a)$ using equation (1) and (2) respectively. Otherwise, compute $\lambda(y_a)$ using equation (1) and compute $P_d(y_a)$ using equation (3);

f) Compare $P_p(y_a)$, $P_d(y_a)$ and $U(0,1)$ to decide to pick up item y_a or drop it;
g) Move to a randomly selected neighboring site not occupied by another ant. Judge whether all the ants have completed the operation. If not, go to step (d).
h) Determine whether to maximum iterating times. If not, go to step (d). Otherwise, clustering process is over.

4 Result and Analysis

There are two parts of the analysis of experiment. First, it is the analysis of the Validity of emotion vectors. Second, verify the validity of the emotion vectors clustering of images by the improved ant colony clustering algorithm.

4.1 Results of Emotion Analysis

In order to prove the validity of the emotion vector extracted, a female subject listen to a sad music. After the music stopped, emotion analysis was started. EEG segments are completely cut out, including not only the time when she listened the music but also the moment before and after it. After dealing with the EEGs data, emotion vectors are extracted. Vector components are analyzed respectively and the results about "sadness" are shown in Fig. 1. As is shown in the Figure, the index of sadness (vector component z2) is high level when she listen the sad music. And the same situation appeared when measuring the others vectors.

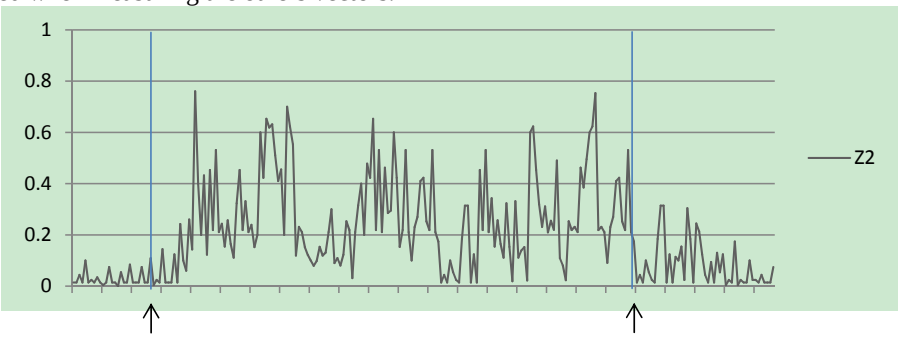


Fig.1. Music can relieve the tension

4.2 Clustering Effectiveness Analysis

In order to verify the effectiveness of image emotion vector clustering by improved ant colony clustering algorithm in this paper, we select 1000 pictures from Corel image library as the original data set, and the size of each image are 256 × 384.

First extract emotion vectors of the images with the method mentioned above and then the data clustering. Since the results of clustering may be always different, table 1 shows that the accuracy of the improved ant colony cluster algorithm is higher than the original one.

As is shown in table 1, two cluster algorithms have different impacts on accuracy of generating clusters of emotion images. For instance, accuracy rate of cluster, "joy", which is produced by the original method, is only 85.51%. However, it can be higher and more stable accuracy by the improved algorithm. The reason why this happened is that cluster centers, in some ways, are generated by manual operation in the improved algorithm. Its results are a very large optimization. Meanwhile, in the calculation of similarity, it is weakened the impact of noise of the emotion vectors by adding weight for each vector component, increasing the accuracy.

Table 1. Accuracy Comparison of the algorithm improved before and after (%)

Emotion Type	Accuracy of original ant colony cluster algorithm	Accuracy of improved ant colony cluster algorithm
joy	85.51	87.67
angry/nervous	68.74	74.30
sad	70.75	78.49
relaxation	81.64	82.37

The research shows that the algorithm improved in this paper has obvious advantages than the original ant colony clustering algorithm, basically reached the expected results.

5. Conclusions

Based on researches, the ant colony clustering algorithm is firstly applied in the image emotion clustering analysis of EEG signal. To improve the clustering accuracy, four groups of data having maximum similarity are selected to form initial clusters. When calculating the similarity between different vectors, based on the importance of each component, vector components are added a weight to optimize results. It shows that the improved ant colony clustering algorithm can effectively generate clusters of EEG signal, and has an advantage, compared with the original algorithm. In future work, ant colony clustering algorithm should be continually improved to solve the problem of solution speed slow.

References

[1] Eckman, P., Friesen, WV (1975). Unmasking the face. *Prentice-Hall,Englewood*.

[2] T. Musha, Y. Terasaki, H. A. Haque, and G. A. Ivamitsky. Feature extraction from EEGs associated with emotions, *in Artificial Life and Robotics*, 1(1):15–19, March 1997.

[3] Handl J., Knowles J., Dorigo M.. Ant-based clustering: a comparative study of its relative importance with respect to K-means, average link and ID-SOM, *in Technical Report TR/IRID-IA/2003-24. University Libre de Bruxelles*, 2003.

[4] Li Haifang, Wang Li.: Clustering algorithm of image emotional characteristics based on ant colony, *in Journal of Computer Applications*(2009).

[5] Dorigo M., Gambardella L. M.: Ant Colony System: A Cooperative Learning Approachto the Traveling Salesman Problem, *IEEE Trans. On Evolutionary Computation*, 1997,1(1).

[6] A.K. MacKworth, R.G. Goebel, and D.I. Poole, Computational Intelligence: A Logical Approach, *Oxford Univ. Press*, 1998, pp. 319–342.

[7] Thayer, R.E.: The self regulation of mood: strategies for changing a bad mood, raising energy and reducing tension, *Pers. Soc. Psychol*, 1994, vol. 67, pp. 910–925.

[8] Chen Zeng: Application of EEG Signal in Emotion Recognition. *Computer Engineering* (2010).

[9] K.S. Chuang, H.L. Tzeng, S.W. Chen, J. Wu and T.J. Chen: Fuzzy c-means clustering with spatial information for image segmentation. *Comp. Med. Imag. Graph*.,30 (2006), pp. 9–15.

[10] P.S. Shelokar, V.K., Jayaraman and B.D. Kulkarni.: An ant colony approach for clustering. Anal. *Chim. Acta.* 509 (2004), pp. 187–195.

[11] Boryczka, U.: Finding groups in data: Cluster analysis with ants. *Applied soft computation*, 2009.

The Storage and Management of Distributed Massive 3D Models based on G/S Mode

Miao Fang[1,a], Xie Yan[2,b], Yang Wenhui[3,c] and Chai Sen[4,d]

[1]Chengdu University of Technology, Geophysical College, Chengdu Sichuan

[2]Chengdu University of Technology, Information Science and Technology College, Sichuan

[3]Chengdu University of Technology, Information Science and Technology College, Sichuan

[4]Chengdu University of Technology, Information Science and Technology College, Sichuan

[a]mf@cdut.edu.cn, [b]xiey521@126.com, [c]ywhui@cdut.edu.cn, [d]claudehotline@gmail.com

Keywords: data distribution, model segmentation, 3DMML, G/S, massive data of 3D model

Abstract. With the development of digital city and smarter city, the massive 3D models of large virtual environment caused the data access bottleneck. It has been received much concern that how to effectively manage these massive data and to implement on-demand service. Under the support of G/S model, according to the characteristics of data of 3D model, this article put forward a 3D model cutting method based on OML, a data exchange standard - 3DMML. With 3DMML as the core, making use of the "data distribution" storage method and metadata organization and management method based on 3DMML to solve the problem of storage and management of massive 3D models, in turn to implement the on-demand dynamic polymerization service for virtual environment based on Storage Cloud at terminals, collaborating analysis and aided decision making.

Introduction

The application of three-dimensional model is wide with the rapid development of Digital City and Smart City. To implement modelling and simulation about virtual reality scene, expression and processing of three-dimensional models in the city environment is necessary. System need to express not only individual building, but also the groups of buildings and models of the whole city. People can do planning and design, traffic control and other decision support, etc. But, the amount of data building a virtual reality scene is huge, with the characteristics of distributed heterogeneous, diverse formats, structured and unstructured. And the data have a strong time and special characteristics. Data using and sources are diverse, the structure of data is very complex, leading to tremendous problems in storing and accessing. People can't effectively organize and manage. We have to and must solve the storage and access problems of mass three-dimensional model in order to achieve true three-dimensional virtual environments, digital and intelligent, and transcend time and space limitations. It convenient to the human understand, alternate, and use our environment.

This paper proposes a model segmentation that a large number of three-dimensional model data split into several files. In a distributed environment, people use metadata to describe three-dimensional model ID, Serial Number, Location, Size, and other information according to the characteristics and attributes of three-dimensional model data. It can determine the location, nature and other functions. This paper also proposes distributed spatial data mark-up language——3DMML (Three Dimensions Model Mark-up Language) in order to organize and manage metadata effectively. 3DMML follows the grammars and rules of HGML with XML as a standard. People

can achieve the description, visualization, interaction, sharing, distributed integrated and management about three-dimensional with the mechanism of metadata organization and management on 3DMML.In the support of G/S (G: Geography-information Browser, S:Service Cloud) , people use cloud services architecture, take advantage of redundant storage, dynamic collaboration, load balancing, adaptive, virtualization management of distributed data and data scheduling mechanism about information gathered to achieve the client information gathering and dynamic service aggregation.

The mechanism of data distributed storage and management based on G/S

G/S mode is a netted spatial information service mode based on the Internet. With HGML as the core and the conception of "request-polymerization-service" of client side polymerization, it is able to organize and manage data and resources distributed in the network, and provided a collaborative computation environment for spatial information service based on 3D model data, in turn to implement client side dynamic polymerization service. G is geographic information browser, giving users a human-computer interaction interface and necessary services for a 3d virtual environment in unified geographic coordinate system; S, according to the current situation of front-end data storage, in the manner of distributed clusters - service cloud - uses slicing, redundant storage, synchronized processing, multi-point download, load balancing to process spatial data to eliminate network access bottleneck, therefore improved the service capability of terminals. Under the support of G/S mode, in front of the massive, multi-source, distributed isomerous, of various types data which has strong time-spatial characteristics required by virtual reality scenes, this article adopts service cloud architecture, and making use of distributed storage management mechanism, effectively solved the storage and management of massive data of 3D models.

Distributed storage is different from traditional distributed storage, it core idea is the distribution of data themselves and storage node. Particularly, it is massive data themselves are divided into several small data blocks, and then stored in servers of different functionalities. Combined with distributed file system technology, NoSQL technology and redundant algorithm under distributed environment, using dynamic redundant data management mechanism, according to the access frequency to dynamically increase or decrease the number of copies of data, and on-demand aggregating the data when needed. By doing this, not only is it in favor of the data transmissions on the Internet, but also improved the storage efficiency of storage media and enhanced the reliability of data.

Data distributed storage is realized though divide storage space in to unified continuous abstract blocks of the same size, and the size can be adjusted according to applications. And each abstract data block is given a start tag as metadata used for storing in metadata database; after data has been stored in database, the corresponding relationships between files and the number of block are also stored in database as metadata. When file are stored in block manner, supplemental storage is used, that is next block is used, only the first block is full, and in order to implement accurate positioning files stored cross different blocks and files storage cross different blocks and servers, metadata has to record the offset of the first and last block, by doing this, the efficiency of storage media improved a lot. The implementation principles of the distributed storage of massive data in distributed 3D models servers, as illustrated in FIG. 1.

The storage, invoke and management of massive 3D model data are dealt with 3DMML, when new data needs processing, regardless of its structure and type, 3DMML is used to encapsulate the data and store in database, this process is called register. Once registered the data becomes to the data resource in service cloud architecture and can be accessed, updated and deleted through 3DMML.

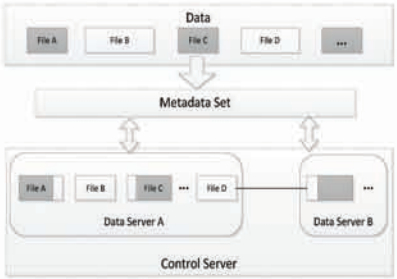


Fig 1. The distributed storage of massive data in distributed 3D models servers

Model Segmentation

The massive data of three-dimensional model affect the 3D visualization, query, and analysis and interoperability functions in creating virtual reality scene based on the above. The formats of 3D model are variety, such as .3max, .skp and so on. The data format of 3D model is unified, then are divided into several files with 3DMML as a core and with solving the massive data storage of three-dimensional as a ultimate goal based on G/S. Finally, computer can achieve distributed storage of 3D model data. Specific processes as illustrated in FIG. 3

This paper proposes OML (Object Modeling Language). OML is a 3D shape language on object-oriented, is used for creating models of real-world scenes or fictional-world scenes, is a platform-independent interpreted language. The object of OML is called node. Sub-set of nodes can pose complex scenes. Node can be reused by instantiating. Computer assigned them to names after creating simulation world of 3D. The description methods of OML are structured. Its files are more readable and can be extracted by the program. It is easy to file splitter. OML file format is .oml or .OML. OML has more ability to express three-dimensional, and support color, texture, light and other common property. OML is consisted of header (must exist), prototype and modeling and so on. This file format learns XML-encoded advantages. It is convenient to information management, control and exchange. This file format provides encryption and compression to improve the security and storage, access speed. A simple experimental result is shown under the guidance above theory, as illustrated in FIG. 2.

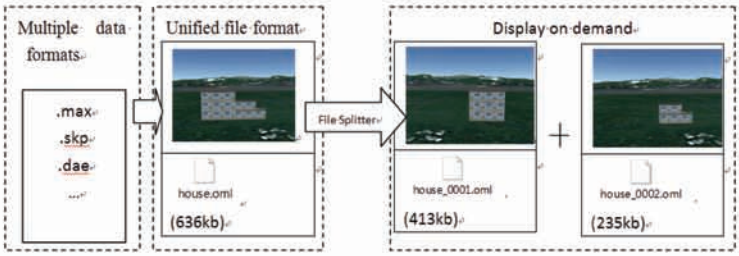


Fig 2. Model segmentation and service on-demand

It describes the process of data segmentation in FIG 2. In the first place, the format is unified; In the second place, computer extracts multiple files automatically through analyzing. That is dispersion; ultimately, system completes 3D display based on actual needs. That is services on-demand.

Three-dimensional Model Mark-up Language-----3DMML

The mass data are produced because of model segmentation, people need to descript and manage them according to the features and attributes of three-dimensional model data, then formulate organization of digital information. For this reason, people use metadata generally. This paper

proves the mechanism of metadata organization and management on 3DMML to achieve that the format of 3D model data are exchanged, data are distributed, information is gathered, and services are aggregated

Metadata Structure. Metadata is used to describe the data, it is the structured data, and provides information about some resource. The purpose of using metadata is that descript the name, location, organization about 3D models were cut up, and identify, evaluate, track, manage, discovery and search resources effectively. People definite the structure of metadata, according to segmentation of 3D model data, distributed storage, as illustrated in FIG. 3.

Object is an abstract representation of all elements. All elements are derived from Object, distinguished with ID; MetaData and its derivative element are used to describe the metadata; Feature represents abstractly all elements that have relevant structure, its derived elements are Placemark, NetworkLink, Container and Overlay, they are used to describe label names, link network resources, create nested hierarchy and reload images, elevation data and 3D model data; Geometry and its derivative element are used to describe vector data; ColorStyle provides the description of all the display style of spatial data; StyleSelector is used to establish the link between the display style and spatial data; TimePrimitive is used to describe the time information.

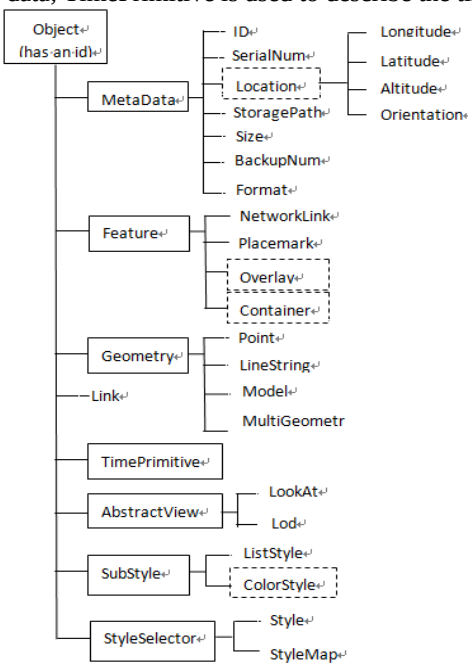


Fig 3. Metadata structure

Metadata Organization and Management on 3DMML. Miao Fang, Chengdu University of Technology, proposed HGML（Hyper Geographic Markup Language） as spatial data exchange standards (different from the Hyper Graphic Markup Language) when solves the problem of spatial information network services architecture with G/S. It played an important role to achieve the exchange and management of massive heterogeneous spatial data. Therefore, this paper puts forward 3DMML as the exchange standards of 3D model refer to HGML, which achieves data management, those massive, heterogeneous, diverse formats. 3DMML is important in the application, operation, sharing and interaction, etc.

The data structure of 3DMML is semi-structured. Definition of 3DMML is flexible label mode; it encapsulates heterogeneous data into structured data to complete unified format of structured data and unstructured data. The basic unit of 3DMML used for describing data is an element which is an

abstract concept. People divide elements into parent and child elements in accordance with the different elements describe the level and scope of data. Child elements share the structure and method of parent elements automatically. 3DMML defines single and complex labels to achieve inheritance in the process of instance. Specific definition as illustrated in FIG. 3.

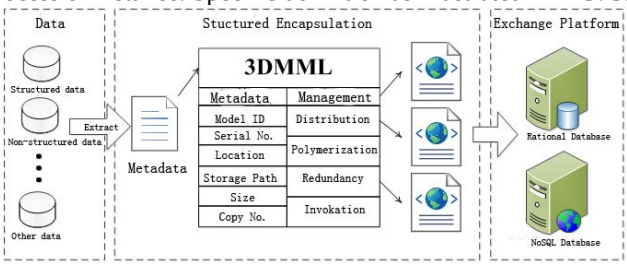


Fig 4. The flow chart of data exchange

This paper researches 3D model data ID, serials number, size, format, location, owners, update and other metadata information and so on, establish automatic and semi-automatic extraction method of metadata. People design approach to building 3DMML metadata-base, manage 3DMML through relational database or NoSQL. The process of data exchanging as illustrated in FIG. 4.

Real-time is considered often in many applications of 3D model data, it includes the real-time of data retrieval and data transmission, which meets the needs of individual users and on-demand services.

People integrate 3D model data of 3DMML format and are distributed on the network through URL to achieve distributed organization and management. Specifically, the data source is transparent to users. 3DMML data source can be local data and also be a remote data. NetworkLink is a 3DMML element. People do integration of distributed with NetworkLink and its derived elements to complete the integration of multi-source data.

Complex label <NetworkLink> is used to describe the integration of distributed data. Specifically, it defines parameters of integration of distributed data through defining <linked>, <href>, <httpQuery>, etc.

Simple label <linkID> is used to define subsets of complete integration of distributed data once to achieve integration of many 3DMML files.

Simple label <httpQuery> is used to define parameters when defined http queries. clientVersion and clientName is used to describe attribute to achieve retrieval and positioning of 3DMML files accurately on the client.

Three-dimensional Model Data on-demand. In the G/S mode support, system can meet different end-users the requirements of diversity, individuality and synergy with real-time transmission and quick search and through the approach of information gathering and services aggregation on client-demand. It is convenient to use 3D model data scientifically and effectively and achieve decision support functions. Information gathering on client-demand is which does classification and retrieval of data which has been registered, stored and managed on the basis of 3DMML. System sets personalized 3DMML data index list automatically according to the specific needs of each end-user, resolves its position in the storage cloud and related properties in accordance with the data model of organization and management, then extracts useful information from data and gathers to the client to achieve visualize expression; Services aggregation on demand is that system sets client-specific functional requirements into personalized 3DMML function index list through cloud service platform and according to loose coupling principles of SOA. It is combined organic that the client itself and all kinds of functions and services from different services subject, which completes specific services and meets ubiquitous and personalized requirements of client. The flow of information gathering and service aggregation based on G/S, as illustrated in FIG. 5.

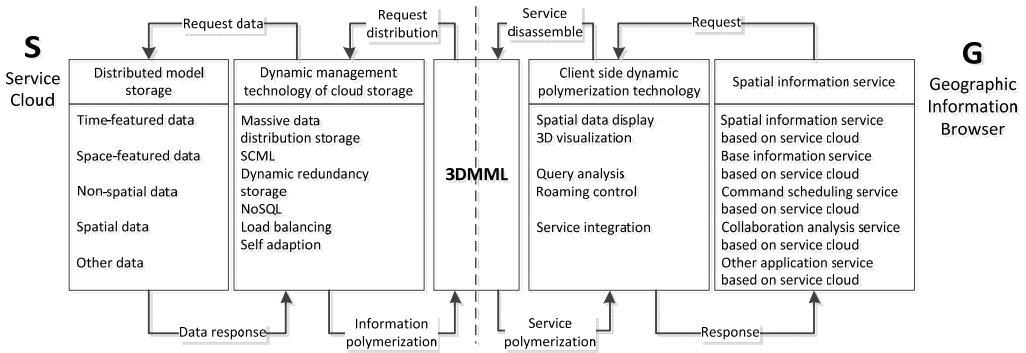


Fig 5. The flow chart of information gathering and services aggregation on client-demand based on G/S

Summary

The computer encounters bottlenecks of network access in practical applications, which restricts the development of Digital Cities and Smart Cities. In the support of G/S mode, this paper uses the architecture of cloud services and "distributed storage" method, and combines with redundant storage, dynamic collaboration, load balancing, adaptive and other advanced technologies. What's more, this paper also proposes segmentation method of three-dimensional model based on OML, data exchange standards—3DMML， the mechanism of metadata organization and management with 3DMML as the core, which solves storage and management of massive data of 3D models and achieves information gathering and services aggregation on client-demand.

Acknowledgment

The authors wish to acknowledge with appreciation the financial support (Grant No. NSFC 61071121) provided by the National Natural Science Foundation of China.

References

[1] Yang WenHui, Xie Yan, Miao Fang, Preliminary Study of HGML-Based Virtual Scene Construction and Interaction Display, DEIT 2011.

[2] Guo XiRong, Mang Fang, Wang HuaJun, LiuRui. The research of digital travel services platform based on G/S[J]. Remote sensing technology and application, 2009(4):490-495.

[3] Guo XiRong, Mang Fang, Wang HuaJun, Xu YiXing. The preliminary exploration of spatial information network access mode G/S. 2009(10):72-74.

[4] Guo Xi-Rong; Miao Fang; Wang Hua-Jun; Du Gen-yuan, Initial Discussion on the Architecture of a New Spatial Information Network Service Mode Based on the Digital Earth,ESIAT2009, Published: IEEE Computer Society CPS, 2009,V3, p406-410.

[5] Tan Li. Tutor: Miao Fang. The research of key technology of creating digital moon platform, 2011.6.

[6] Miao Fang, Ye ChengMing, Liu Rui. The exploration of new generation digital earth platform and the technical architecture of Digital China[J], Surveying and Science. 2007, (6):157-158.

The Current Situation of Green Tourism in China

Diao Zhibo[1, a]

[1]Tourism and Cuisine School, Harbin University of Commerce, Harbin, China

[a]dzb1977@sina.cn

Keywords: Green Tourism; Environmental Protection; Sustainable Development; China.

Abstract. The paper aims at analyzing the current situation of green tourism in China. By means of literature review and focus group discussion, the facts and findings are clear. Conclusions are as followings: NGOs, governmental agencies, businesses and mass media have made great contributions to green tourism. The study is helpful to government agencies and commercial development.

Introduction

There is no specific definition of green tourism. Strictly speaking, green tourism is different from ecotourism or rural tourism though they are usually recognized as equal and interchangeable. Green usually means nature, life, safety, non-pollution, environmental protection, etc. Green tourism is just a figure of speech. Green tourism is closely interrelated with sustainable development, environmental protection, biological diversity, human health, etc. Green tourism includes a series of ideas, methods and measures.

From the perspective of tourists, green tourism refers to those activities based on sustainable development, environmental protection, biological diversity and other principles, even green tourism can be one method to achieve goals such as sustainable development. From the perspective of tourism businesses, green tourism refers to providing services and products by means of advanced technologies and methods such as low carbon, energy saving and circular economy. In brief, green tourism is environment-friendly to leave a beautiful world to future generations. Green tourism is an inevitable result of human beings' return to nature.

By the underlying resources, green tourism can be classified into green tourism on the sea, lake, island, hill, grassland, etc. By the organizers, green tourism can be classified into green tourism by tourism businesses, NGOs, governmental agencies and green tourists themselves. Green tourism is not limited to rural and unexploited areas. Green tourism can be organized in the city, such as visiting rubbish or wastewater treatment plants.

Methodology

The paper was based on available literature and focus group discussion. It made full use of qualitative analysis. The literature was selected from those influential periodicals, and it covered major provinces and cities of China. Therefore a complete picture of the current situation and problems of green tourism could be reflected through the literature review. Besides, five experts were invited to take part in a focus group discussion about the result of literature review. Each expert was required to give his comment and additional ideas. The facts and findings were summarized based on the literature review and suggestions of the experts.

Literature Review

The relevant literature involves several aspects of green tourism, such as the meaning and definitions of green tourism, the development of green tourism, regional green tourism, the impact of green tourism, the branches of green tourism, etc. Some scholars discussed basic questions about green tourism (Zou Tongqian, 2005[1]; Chen ling, 2004[2]; Zhang Linbo and Zhang Lili, 2009[3]). Most papers were related to how to develop regional green tourism (Xiao Shenghe and Lian Yunkai, 2001[4]), but they lack of case study and empirical research. Few papers were about the impact of green tourism on tourists and environmental protection (Cai yonghai and Zhang zhao[5], 2009). Some papers provided a contrast between China and developed countries (Shi Linyun, 2008[6]; Xu Keshuai and Zhu Haisen, 2008[7]). Most papers adopted normative study and basic qualitative research methods to generalize; only few adopted basic quantitative research method. Therefore more intensive study is very necessary.

Current Situation of Green Tourism in China

NGOs' Contribution to Green Tourism. Green tourism in China has developed with the spread of sustainable development, environmental protection, and corporate social responsibility. Green tourism is one application of those ideas. During such a process, local and international NGOs play a critical role. They devote to the publicity of those ideas among the masses. They have been working hard to promote environmental awareness about China's most pressing environmental problems. They have established many professional web sites, edited magazines and newspapers, held meetings and seminars, enrolled large numbers of members, and organized public welfare activities. By means of these efforts, the public have more chances to know and concern about sustainable development and environmental protection. Some green tourists emerge from the cultivated masses. Besides, these NGOs regularly organize some activities focusing on environmental protection, energy saving, circular economy and biological diversity. Those participants get firsthand experience from activities such as visiting and interviews.

For example, famous domestic NGOs include Friends of Nature (FON, the oldest environmental NGO in China), Global Village of Beijing Environmental Education Center (GVB), and Green Web. They have regularly organized some activities like bird watching, mountain climbing, and wild animal investigation. Large amounts of overseas environmentalists swarm into China every year to fulfill their green travel.

Some NGOs propose behavior of green tourists: choosing environment-friendly hotels, restaurants and transportation; counteracting carbon footprint by tree planting; never eating wild animals; never buying wild animals and products made of them; buying local food and products; no littering, esp. battery; never using throwaway plastic bags.

Governmental Agencies' Contribution to Green Tourism. In 1999, China National Tourism Administration (CNTA) declared the "Year of Ecological Environment". Then in 2009, CNTA declared the "Year of Eco-tourism". Its slogan was "Be a green traveler and experience eco-civilization". CNTA hopes to promote the concept of environment-friendly travel and encourage resource-saving and energy efficient tourism operations. CNTA also devotes to transform the China's tourism sector into a green industry with sustainable development.

In 2008, China Tourism Association (CTA) presented proposals of green travel to all tourists and tourism businesses. CTA called on them to contribute to resource-saving and environment-friendly society. In 2001, Tourism Administration of Zhejiang Province issued Criteria of Green Hotels. It was the first local criteria of green hotels. As of the end of 2006, the number of certificated green hotels amounted to 229. In 2006, CNTA issued National Criteria of Green Hotels. Since that time, those star-rated hotels throughout China began to adjust to those criteria.

Besides, local governments of all level have promoted development of green tourism locally. Many cities aim at transition to green tourism. For example, the local government of Guilin City

greatly emphasizes environmental protection as well as economic growth. The air quality of the city ranks No. 1 among the inland cities. The water quality of Lijiang River is also top ranked. During the process of upgrading tourism industry, low carbon and green are their choice. Green and low carbon travel are very popular, such as bicycle riding, hiking and rural tourism.

Businesses' Contribution to Green Tourism. Green Tourism businesses, such as travel agencies, hotels, transportation, nature reserves and tourist attractions, have made some improvement of energy saving, water saving , environmental protection, soil protection, garbage disposal and pollutant emission. They have provided many kinds of green products. However, those businesses still have a strong motivation to make a profit. To some extent, their conception is contradictory to sustainable development.

Some businesses have got one or more certification such as ISO14000 and Green Global 21. It showed that some businesses have recognized the importance and urgency of environmental problems. Green Global 21 is the only globally accepted certification program for tourism. Green Globe 21 aims at sustainable travel and tourism for consumers, businesses, and communities. It is based on Agenda 21 and principles for Sustainable Development. There are some businesses certified by Green Globe 21 Company Standard, including Jiuzhaigou National Nature Reserve, Huanglong National Scenic Area, Sanxingdui Heritage Site Museum, South Sichuan Bamboo Sea National Scenic Area, Jiuzhai Paradise International Resort and Convention Centre, Zhejiang World Trade Centre Grand Hotel (five star), Shenzhen Pavilion Hotel (five star).

Mass Media's Contribution to Green Tourism. Mass media, including TV stations, newspapers, radios and websites, plays an important role in developing green tourism, esp. those mass media sponsored by government agencies. For example, those public television broadcasting stations like China Central Television (a national TV station providing over 20 channels), often attach great importance to public advertisements in the prime hours of the evening. With the help of mass media, many volunteers have become good examples followed by young people. Meanwhile those areas suitable for green tourism are paid more attention to get rid of poverty. Usually those news reports are concerned by the audience. These media make a nationwide call for green tourism and green tourists. Chinese people have begun to care more about the environment. More and more tourists realized the problems of mass tourism. They incline to travel by themselves and try to avoid group tour.

Summary

Green tours have become increasingly popular among domestic and international tourists. NGOs, governmental agencies, businesses and mass media have made a great contribution to the spread of green tourism in China. They separately or jointly promote the development of green tourism which means a new direction of tourism. Green tourism does not only mean an idea or concept, but also a series of activities. There is still a long way before green tourism becomes dominant over traditional tourism.

Acknowledgement

This research was sponsored by Heilongjiang Planning Office of Philosophy and Social Science (No. 10C001), "Development Patterns and Effect Assessment of Rural Tourism for New Villages in Heilongjiang Province".

References

[1] Zou Tongqian: On the Guidelines and Action Plans of Green Tourism in China. China Population, Resources and Environment. Vol. 4 (2005), p. 43-47

[2] Chen ling: Green Tourism and Certification System in 21st Century. Ecological Economy. Vol. 11 (2004), p. 73-75

[3] Zhang Linbo, Zhang Lili: Green Tourism and Its Prospect in China. Ecological Economy. Vol. 1 (2009), p. 265-267

[4] Xiao Shenghe, Lian Yunkai: On Developing Green Tourism in Guangxi. Journal of Guilin Institute of Tourism. Vol. 4 (2001), p. 14-16

[5] Cai yonghai, Zhang Zhao: Thought on Green Tourism. Theoretical Exploration. Vol. 4 (2009), p. 105-106

[6] Shi Linyun: Japan's Green Tourism and Its Enlightenments to China. Ecological Economy. Vol. 2 (2008), p. 198-201

[7] Xu Keshuai, Zhu Haisen: Green Tourism Development in Japan and Its Insight to China's Rural Tourism Development. World Regional Studies. Vol. 2 (2008), p. 102-109

Improvement of the Mutual Authentication Protocol for RFID

Juseok Shin[1, a], Sejin Oh[2, b], Cheolho Jeong[3, c], Kyungho Chung[4, d], Yonghwan Kim[2, e], Sanghoon Kim[2, f] and Kwangseon Ahn[2, g]

[1]Electronics and Telecommunications Research Institute, Daegu, Korea

[2]Granduate School of Electrical Engineering and Computer Science, Kyungpook National University, Korea

[3]School of Electronic Engineering, Kyungnam University, Korea

[4]School of Computer Engineering, Kyungwoon University, Korea

[a]jsshin@etri.re.kr, [b]170m3@knu.ac.kr, [c]jch21@kyungnam.ac.kr, [d]khjung@ikw.ac.kr, [e]hypnus@knu.ac.kr, [f]ksh3000@knu.ac.kr, [g]gsahn@knu.ac.kr

Keywords: RFID; Mutual Authentication; Protocol; Hash Function; Security

Abstract. In 2010, Wei et al. proposed a new RFID authentication Protocol which is AMAP (A Mutual Authentication Protocol for RFID) to resolve a variety of problem related to security using mutual authentication scheme and updating secret value. In addition, Wei et al. proved AMAP was safe for a variety of attacks including Replay Attacks through safety analysis. However, this paper demonstrates that Wei et al.'s protocol has a serious problem such as asynchronous secret value between Back-end Server and Tag. In this paper, we also propose simply improved Wei et al.'s protocol to resolve problem of AMAP.

Introduction

RFID (Radio Frequency Identification) system, a kind of contactless automatic identification system, consists of Reader, Tag and Back-end Server. Recently, RFID systems to replace the bar code technology, has been applied and has been used in distribution, logistics, transportation and security, etc. [1], [2]. However, the data transmitted in the air between the tag and the reader could easily be intercepted and eavesdropped. Thus, EPC-Global Class 1 Gen 2-UHF Specification provides Kill password, Access password and simple protocol [3]. Case of Kill password, it provides greater safety, but it has a disadvantage that tag can't be reuse. A similar approach, there are the Faraday Cage technology, Blocker Tag and Active Jamming technology, etc. Access password and simple protocol are used for accessing and communicating the tag, but safety is vulnerable. Because EPC (Electronic Product Code) is transmitted without encryption and it does not provide mutual authentication between the reader and the tag in simple protocol. Therefore, to reuse the tag and to resolve the problem of security in RFID system, protocols should be designed using the cryptographic methods (hash functions, public key encryption, symmetric encryption, etc.) and mutual authentication scheme. Thus, recently, many Researchers have proposed numerous RFID protocols using these cryptographic methods and have demonstrated the safety of the proposed protocol. In 2010, Wei et al. proposed the hash-function-based mutual authentication protocol [4]. The proposed protocol provides mutual authentication between the reader and the tag through the Back-end Server and updates secret value for each session between the tag and the Back-end Server to protect several attacks such as Tag Tracking, Replay Attacks, etc. and provide Forward Security. Nevertheless, this paper demonstrates that Wei et al.'s protocol have a serious

problem such as asynchronous secret value between the tag and Back-end Server due to Replay Attacks unlike their claims. In addition, we propose simply improved Wei et al.'s protocol and safety and efficiency is compared with their protocol, it can prevent the security problem and more efficiently.

The remaining sections of the paper are organized as follows: Section 2 briefly reviews Wei et al.'s protocol. Section3 gives problem on Wei et al.'s protocol. The improved protocol is presented in Section4. And Section5 discusses the safety and efficiency of the propose protocol with Wei et al.'s protocol. Finally, Conclusion is presented in Section6.

Review of AMAP (A Mutual Authentication Protocol)

This section reviews AMAP, Table 1 defines the notations used in AMAP. In AMAP, assume that communication channels are insecure not only between reader and tag but also between reader and back-end server (database).

Table 1 Notations used in AMAP.

Notation	Description
[ID]	Unique identifier of tag
[RID]	Unique identifier of reader
[S_{new}]	Newly updated secret value
[S_{old}]	Old secret value
[$R_{r(new)}$]	New random number, generated Reader
[$R_{r(old)}$]	Old random number, generated Reader
[R_{db}]	Random number, generated Back-end Server
[R_t]	Random number, generated Tag
[H()]	One-way hash function
[A =? B]	Whether A equals B or not
[⊕]	Exclusive-or operation

The information kept within perspective devices in initialization phase is follows.

• Back-end Server: ID, RID, Snew, Sold, $R_{r(new)}$, $R_{r(old)}$

• Reader: RID

• Tag: ID, S

Fig.1 illustrates the AMAP. The detailed steps of authentication are presented as follows.

Step1. Reader → Tag

The reader generates random number R_r and sends it to the tag.

Step2. Tag → Reader

After receiving R_r, the tag generates random number R_t and computes $M1 = H(S \oplus R_r \oplus R_t)$ then sends R_t and M1 to the reader.

Step3. Reader → Back-end Server

The reader computes $M2 = H(RID \oplus R_r)$, then sends it together with R_r, generated Step1, and (M1, R_t) received from the tag, to the Back-end Server.

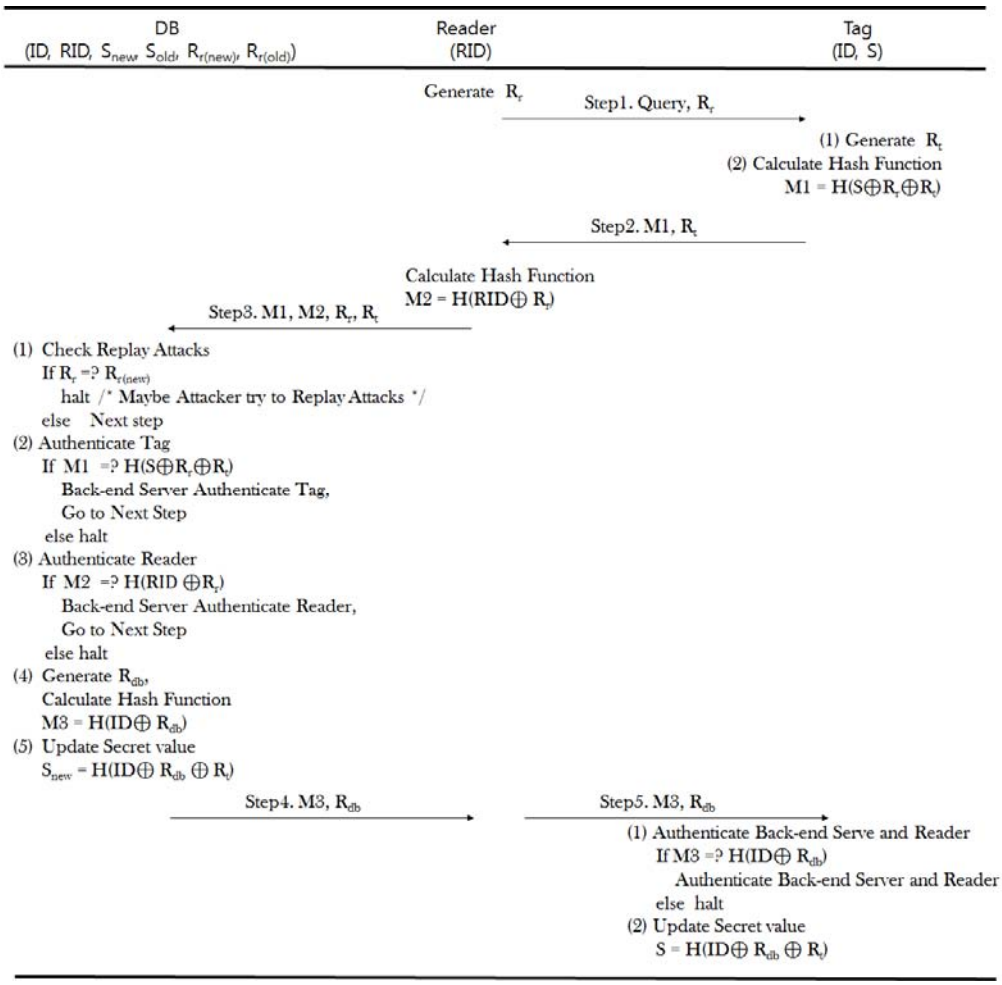


Fig. 1 Wei et al.'s proposed mutual authentication protocol

Step4. Back-end Server → Reader
After receiving M1, M2, R_r and R_t from the reader, the Back-end server performs the following:

(1) Compares R_r with $R_{r(old)}$ to know whether attempt Replay Attacks.
(2) If the two values do not match, it retrieves each stored secret value S to compute $H(S \oplus R_r \oplus R_t)$ with R_r, R_t, and compare with M1 and obtain tag's ID.
(3) Computes $H(RID \oplus R_r)$ with R_r and compare M2 with $H(RID \oplus R_r)$. If M2 is equal to $H(RID \oplus R_r)$, Back-end Server consider the reader is legal.
(4) Generates a random number R_{db} and computes $M3 = H(ID \oplus R_{db})$.
(5) Updates secret value $S_{new} = H(ID \oplus R_{db} \oplus R_t)$ with R_t, generated Step2 and sends M3 and R_{db} to the reader.

Step5. Reader → Tag
After receiving M3 and R_{db}, it sends these data to the tag.

When the tag receives the message from the reader, the tag computes $H(ID \oplus R_{db})$ with R_{db}, generated Step4 and compares M3, received from reader with $H(ID \oplus R_{db})$ to authenticate the

Back-end Server and Reader. Mutual authentication occurs if the comparison of M3 == $H(ID \oplus R_{db})$ is successful. Finally, the tag also updates the secret value $S_{new} = H(ID \oplus R_{db} \oplus R_t)$ to synchronize Back-end Server's secret value S_{new} and prevent attack such as Tag Tracking and Forward Security.

Weaknesses of AMAP

This section shows that AMAP has the serious problem which is asynchronous the secret value between Back-end Server and Tag by Replay attacks. Therefore, even legal the tags, it can't normally be communicated anymore. Fig. 2 shows the scenario of the Replay attacks on AMAP. Replay attack is that adversaries intercept the transmitted data and resend it illegitimately in an attempt to a legitimate device and pass the authentication.

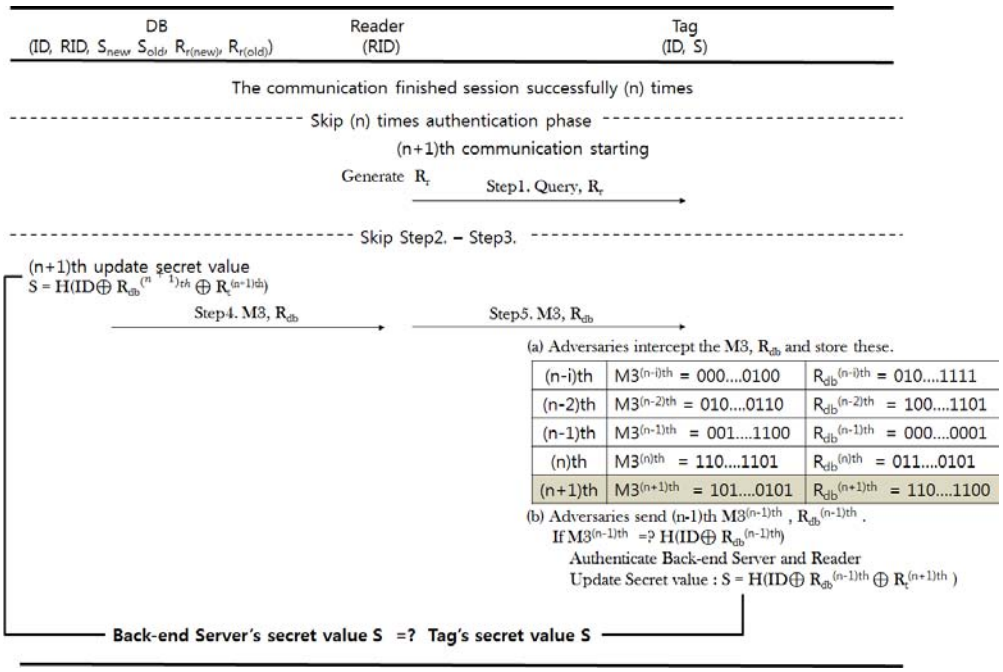


Fig. 2 Wei et al.'s proposed mutual authentication protocol

It is possible to make Replay attack on Step5 in AMAP. The detailed steps of Replay attacks on AMAP are presented as follows:

(1) Adversaries can intercept M3 and R_{db} on Step5 and store these data on memory several times.

(2) If adversaries modify the M3 or R_{db} and send it to the tag, it is easy to distinguish whether data is legal or not because of comparing M3 =? $H(ID \oplus R_{db})$.

(3) For example, already the communication finished session successfully (n) times and adversaries also save all of the data (M3, R_{db}). After (n+1)th communication starting, in Step5, adversaries send saved (n-i)th data(M3, R_{db}) to tag.

(4) In this time, Back-end Server already updates (n+1)th secret value. However, the tag updates (n-i)th secret value again, so it have previous session secret value. Thus, Back-end

Server and Tag have different secret value.

(5) Next time, if legal the tag reply to reader's request (Query), Back-end Server can't authenticate it due to asynchronous secret value each other. Thus, some tags are could not be reused anymore.

Propose IAMAP (Improved AMAP)

This section proposes IAMAP that can resolve the problem of asynchronous secret value between B ack-end Server and the tag. The information kept within respective devices is almost same as AMA P but difference one is that Back-end Server has only one information which is $RID_{(old)}$ more than A MAP. Table 2 defines notation used in IAMAP.

Table 2 Notations used in IAMAP and other notations are same as AMAP.

Notation	Description
$[RID_{(new)}]$	Newly updated RID(Unique identifier of reader)
$[RID_{(old)}]$	Old RID(Unique identifier of reader)
[Info]	Tag's Information
[||]	Concatenate operation

Fig. 3 depicts improved mutual authentication phase. The detailed steps of the improved mutual authentication phase are presented as follows.

Step1. ~ Step3.
This steps are almost the same as AMAP but one difference is that compute M1, M2 using not exclusive-or operation but concatenate operation ($M1 = H(S||R_r||R_t)$, $M2 = H(RID||R_r)$).

Step4. Back-end Server → Reader
After receiving M1, M2, R_r and R_t from the reader, the Back-end server performs the following operations:

(1) Compares R_r with $R_{r(old)}$. If R_r is equal to $R_{r(old)}$, consider that Replay Attacks arises. Then the protocol will be aborted.

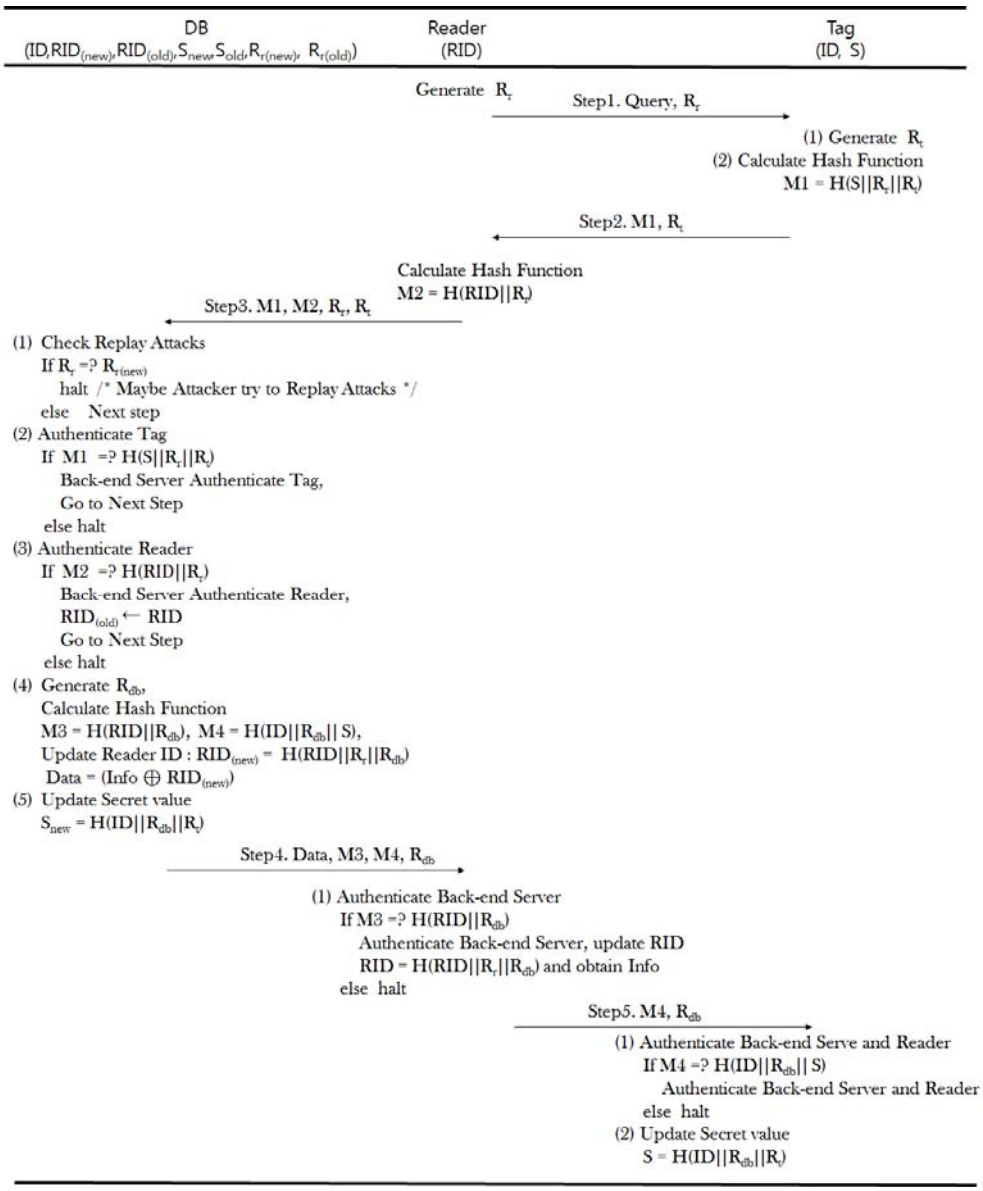


Fig. 3 The proposed IAMAP phase

(2) If R_r is not the same as the $R_{r(old)}$, retrieves each stored secret value S to compute $H(S||R_r||R_t)$ with R_r and R_t and compare with M1 to verify whether the tag is legal or not and to obtain tag's ID. If there is no matching secret value S in the Back-end Server, the protocol also abort.

(3) Computes $H(RID||R_r)$ with R_r, generated Step1, and compare M2, received from the reader, with $H(RID||R_r)$. If M2 is equal to computed $H(RID||R_r)$, authenticate the reader. After considering legal reader, RID is stored to $RID_{(old)}$.

(4) Generates a random number R_{db} and computes $M3 = H(RID||R_{db})$ and $M4 = H(ID||R_{db}||S)$.

Updates reader's ID: $RID_{(new)} = H(RID||R_r||R_{db})$ and computes $Data = (Info \oplus R_{db})$.

(5) Secret value also update $S_{new} = H(ID||R_{db}||R_t)$ and sends Data, M3, M4 and R_{db} to the reader.

Step5. Reader → Tag
After receiving Data, M3, M4 and R_{db}, the reader computes $H(RID||R_{db})$ with R_{db}, received from Back-end Sever and compare M3 with $H(RID||R_{db})$ to attempt to authenticate it. If hashed value $H(RID||R_{db})$ matches the M3, the reader consider that Back-end Server is legal. After authenticating the Back-end Server, updates $RID = H(RID||R_r||R_{db})$ and can obtain the Info using exclusive-or operation. The next phase, send M4 and R_{db} to the tag and it computes $H(ID||R_{db}||S)$ with R_{db}, generated Step4 and own secret value S. Then, compares M4, received from reader with $H(ID||R_{db}||S)$ to authenticate the Back-end Server and Reader. Finally, the tag also updates the secret value $S_{new} = H(ID \oplus R_{db} \oplus R_t)$ if the M4 is equal to $H(ID||R_{db}||S)$.

Safety and Efficiency of IAMAP

This section discusses the comparison of AMAP and IAMAP for safety and efficiency and we prove that IAMAP is more safe and efficient than AMAP. Table 3 shows security properties of the AMAP and IAMAP and comparison of efficiency is presented in the Table 4.

Table 3 Security properties of the AMAP and IAMAP

	eavesdropping	Replay Attacks	Tag Tracking	Forward Security
AMAP	*[Insecure]*	*[Insecure]*	[Secure]	[Provide]
IAMAP	*[Secure]*	*[Secure]*	[Secure]	[Provide]

The AMAP's problem of the Replay attacks was examined in section 3 and Eavesdropping is used in Replay Attacks, so AMAP is also vulnerable to it. However, IAMAP is safe for Replay Attacks. We define security analysis of the IAMAP as follows.
Proof. Assume that Adversary intercept the M4 and R_{db} between the reader and tag on Step5 and save it to memory at every session. And then, Adversary send saved (n-i)th data(M4, R_{db}) to tag after successfully finished communication (n)times. In this case, asynchronous secret value between Back-end Server and the tag on AMAP. In IAMAP, We can easily see that $M4^{(n-1)th}$ is not equal to $M4^{(n+1)th}$ because M4 is made using secret value S and it is updated for every session.

Table 4 Security properties of the AMAP and IAMAP

Computation Overhead	AMAP	IAMAP
Tag	[3*H() + 4*XOR(operation)]	[3*H()]
Reader	[1*H() + 1*XOR(operation)]	[3*H()]
Back-end Server	[4*H() + 6*XOR(operation)]	[6*H() + 1*XOR(operation)]

In the IAMAP, both the Back-end Server and the reader require twice of the hash operation to update RID, respectively. Rest computations are more efficient due to not using exclusive-or operation but using concatenate operation. The reason of the updating RID and using concatenate operation is that protect exhaustive search even very difficult or impossible. In addition, IAMAP provide that the reader can obtain the tag's information.

Conclusion

This paper demonstrated that AMAP has problem such as asynchronous secret value between Back-end Server and the tag because of the Replay attacks and the reader can't obtain the tag's information. Thus, we presented a simply improved mutual authentication protocol to resolve above the problems. As a result, proposed protocol IAMAP is not only enhanced secure against

well-known several attacks but also provides good efficiency since it provides mutual authentication based on hash function. Finally, we expect that IAMAP is used in a variety of RFID systems for ensuring safety.

References

[1] S. A. Weis.: Security an Privacy in Radio-Frequency Identification Devices. MS Thesis. MIT. (2003).

[2] S. A. Weis, S. E. Sarma, R. L. Rivest, D. W.Engels.: Security and Privacy Aspects of Low-Cost Radio Frequency Identification Systems. Security in Pervasive Computing 2003, LNCS 2802, pp. 201-212, Springer-Verlag Heidelberg (2004).

[3] F. Klaus.: RFID handbook. Second Edition, Jone Willey & Sons (2003).

[4] Chia-Hui Wei, Min-Shiang Hwang, Chin, A.Y.: A Mutual Authentication Protocol for RFID.Computing & Processing, vol.13, pp. 20-24. IEEE Computer Society (2011).

[5] Gene Tsudik, YA-TRAP.: Yet Another Trivial RFID Authentication Protocol. Proceedings of the 4th annual IEEE international conference on Pervasive Computing and Communications Workshops, pp. 640-643 (2006).

[6] J. Aragones, A. Martinez-Balleste, A. Solanas.: A brief survey on rfid privacy and security, In World Congress on Engineering (2007).

[7] M. Ohkubo, K. Suzuki, S. Kinoshita.: Hash-chain based forward secure privacy protection sheme for low-cost RFID. Proceedings of the 2004 Symposium on Cryptography and formation Security.Sendai, pp. 719-724 (2004).

[8] Jung-Sik Cho, Sang-Soo Yeo, Sung-Kwon Kim.: Securing against brute-force attack: A hash-based RFID mutual authentication protocol using a secret value. Computer Communications, Vol. 34, No. 3, pp.391-397 (2011).

[9] A. Juels.: RFID security and privacy: a research survey. IEEE Journal on Selected Areas in Communications, Vol. 24, No. 2, pp.381--394 (2006).

A Mutual Authentication Protocol in RFID Using CRC and Variable Certification Key

Sejin Oh[1, a], Juseok Shin[2, b], Cheolho Jeong[3, c], Jaekang Lee[1, d], Sungsoo Kim[1, e], Seungwoo Lee[1, f] and Kwangseon Ahn[1, g]

[1]Graduate School of Electrical Engineering and Computer Science, Kyungpook National University, Korea

[2]Electronics and Telecommunications Research Institute, Korea

[3]School of Electronic Engineering, Kyungnam University, Korea

[a]170m3@knu.ac.kr, [b]jsshin@etri.re.kr, [c]jch21@kyungnam.ac.kr, [d]10004oke@knu.ac.kr, [e]ninny@knu.ac.kr, [f]zpa007@knu.ac.kr, [g]gsahn@knu.ac.kr

Keywords: RFID; Protocol; Mutual Authentication; Hash Function; CRC

Abstract. An RFID system is data transmission technology to use radio frequency. Then, it has many problems like eavesdropping, location tracking, spoofing attack and replay attack. Recent research has being progress about cryptography and mutual authentication to resolve these problems. Despite many studies, previously proposed protocols were vulnerable to security. In this paper, we proposed RFID protocol using CRC code and variable certification key. Proposed protocol safely encrypts data on wireless and prevents various attacks by mutual authentication using CRC code. In addition, our protocol is secure in various attacks than past protocol and is efficient in computation.

Introduction

The RFID(Radio Frequency Identification) technology can alternate with barcode recognition technology to pass the limit of it. RFID systems can be divided into sub-parts-tag, reader and server. The tags contain the information of the object and reader reads the tag's information. Finally, The server manages the tag's unique information[1]. RFID systems has problem like eavesdropping, location tracking, spoofing attack and replay attack because of using radio frequency signal. In order to resolve this problem, cryptographic and physical methods must be used. Kill password, faraday cage, blocker tag and active jamming are physical methods. However, physical methods make the tag can't reuse. So, RFID protocol should be used cryptographic methods and mutual authentication for security and privacy. Cryptographic methods are hash function, AES(Advanced Encryption Standard) and public key algorithm. Recently published paper, there are problems such as location tracking, exposure of tag's ID in RFID protocol[2]. In this paper, we propose a hash function-based protocol. Our protocol uses a CRC(Cyclic Redundancy Checking) code and variable certification key for mutual authentication. In addition, it's excellent from security and efficiency.

Related Works

The hash-based protocol has been activated by the Hash-Lock Protocol[2]. The advantage of a hash-based protocol provides forward-secrecy. Therefore, the hash function is used in many RFID protocols. This section describes the SRAP(Security RFID Authentication Protocol) and HMAP(Hash function-based Mutual Authentication Protocol).

SRAP(Security RFID Authentication Protocol). SRAP transmits encrypted tag's unique identification by hash function and perform mutual authentication using reader's and tag's unique ID[3]. However, this protocol has a problem in location tracking. The reason is because of the fixed hash values(h(IDT)).
Besides, the attacker passes the authentication procedure if it retransmits IDT⊕R to the tag. The specific work process of the protocol is shown in Fig. 1.

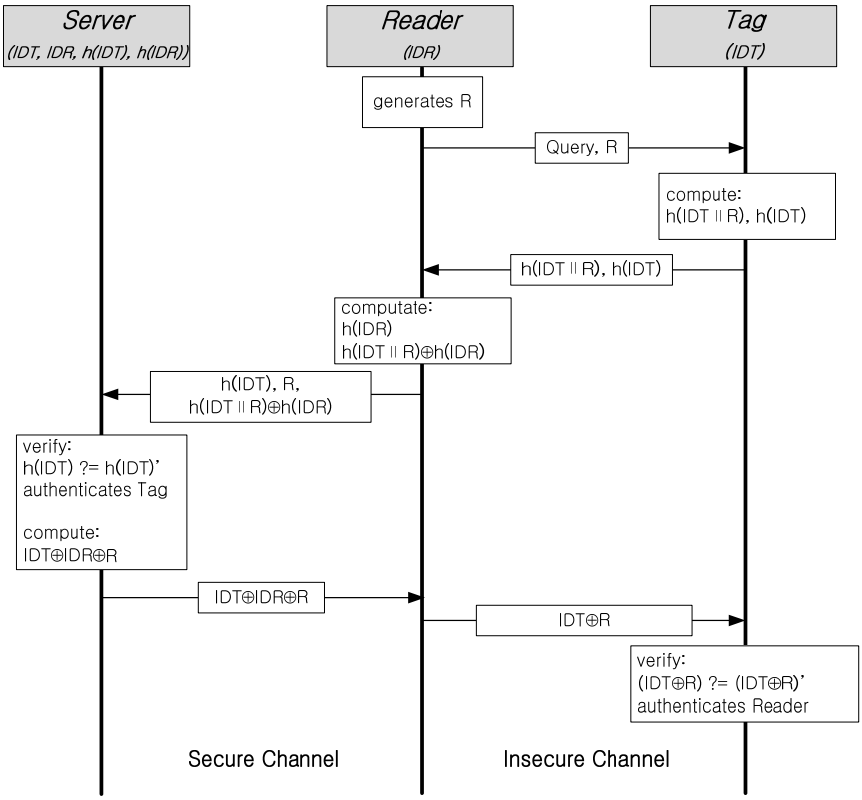


Fig. 1. The Work Process of Security RFID Authentication Protocol

HMAP(Hash function-based Mutual Authentication Protocol). HMAP is protocol to use reader's random number and tag's unique information[4]. They claimed that the proposed HMAP withstand against location tracking, replay attack and spoofing attack. The reason is the variable response in the tag. However, $r \| h(ID_k)$ is operated by simple concatenation operation. Therefore, HMAP is vulnerable to location tracking because next of the r-bits is fixed tag's $h(ID_k)$, result of hash function. Also, the spoofing attack and the replay attack are possible in HMAP because attacker obtain the data from eavesdropping attack. Attacker can compute $r' \| h(ID_k)$ using previously obtained $h(ID_k)$ and r' of next session, transmit it to the reader. Then, spoofing attack is possible, and attacker can pass the authentication process. The specific work process of the protocol is shown in Fig. 2.

Proposed Protocol

We proposed protocol against various attacks in RFID systems. Our protocol safely encrypts using hash function, and perform mutual authentication to use CRC code, variable certification key. This chapter describes the assumptions, notation and proposed protocol.

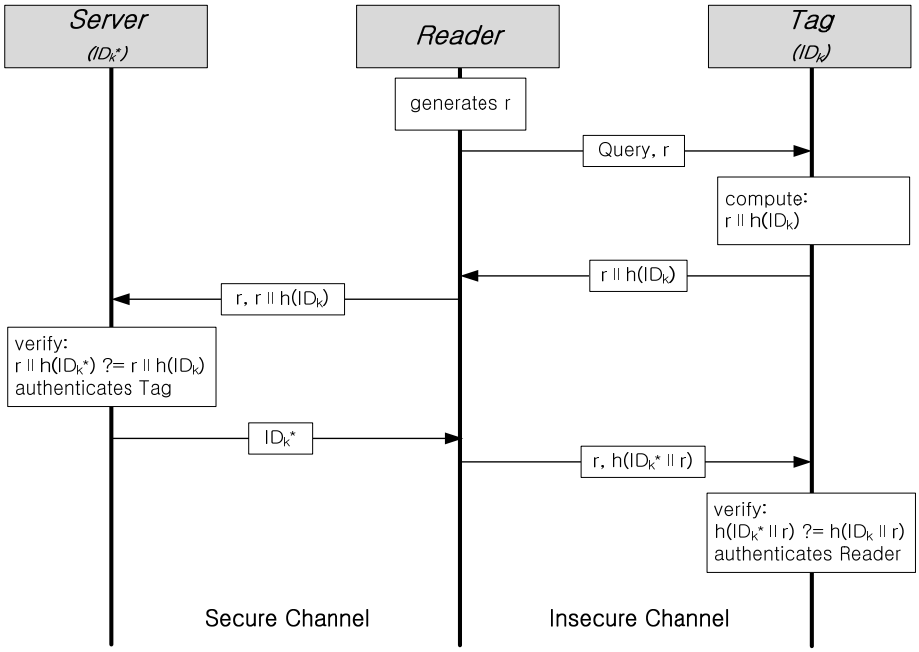


Fig. 2. The Work Process of Hash function-based Mutual Authentication Protocol

Prior Conditions and Assumptions. In our protocol, there are five assumptions. Table 1 is notation.

1) Secure communication channel between the server and reader is used.
2) Insecure communication channel between the reader and the tag is used.
3) The server and the tag can calculate hash function.
4) The tag, reader and server can generate a random number.
5) The tag, server can operate a CRC code and key-shift operation.

Table 1. Notation Used in Proposed Protocol

Notation	Description
IDt	Tag's unique identification
IDs	Tag's unique identification in Server
K	Tag's and Server's symmetric key
Rr	Reader's random number
Rt	Tag's random number
Rs	Server's random number (Rs(10)={0,1,2, …, 126, 127})
H( )	Hash function
crc()	CRC function that is provided by EPC-Global Class 1 Gen 2
P	Pivot bit
C	Variable certification key
Tag Info	Information of Tag

A Variable Certification Key Generates by Key-Shift Operation. C value is used mutual authentication in proposed protocol. First, the server generates a random number Rs in decimal. The Rs range is from 0 to 127. The Rs value is selected pivot bit, and the server perform key-shift operation. The Left_bits store from MSB(Most Significant Bit) to P-bit of symmetric key K. Then,

The Right_bits store from P-1 bit to LSB(Least Significant Bit) of IDs. Finally, The Left_bits and Right_bits are concatenated. C is this. The key-shift process and C to generate are shown in Fig. 3.

The Proposed Protocol. The proposed protocol is shown in Fig. 4.
Step 1. Reader → Tag : Query, Rr
The reader generates random number Rr. The reader sends a Query message along with Rr to the tag.

Step 2. Tag → Reader : H(IDt ‖ Rr ‖ Rt), Rt
With IDt, Rr, and Rt, the tag computes H(IDt ‖ Rr ‖ Rt), Rt, and sends it to the reader.

Step 3. Reader → Server : H(IDt ‖ Rr ‖ Rt), Rt, Rr
The reader sends H(IDt ‖ Rr ‖ Rt), Rt and Rr to the server.

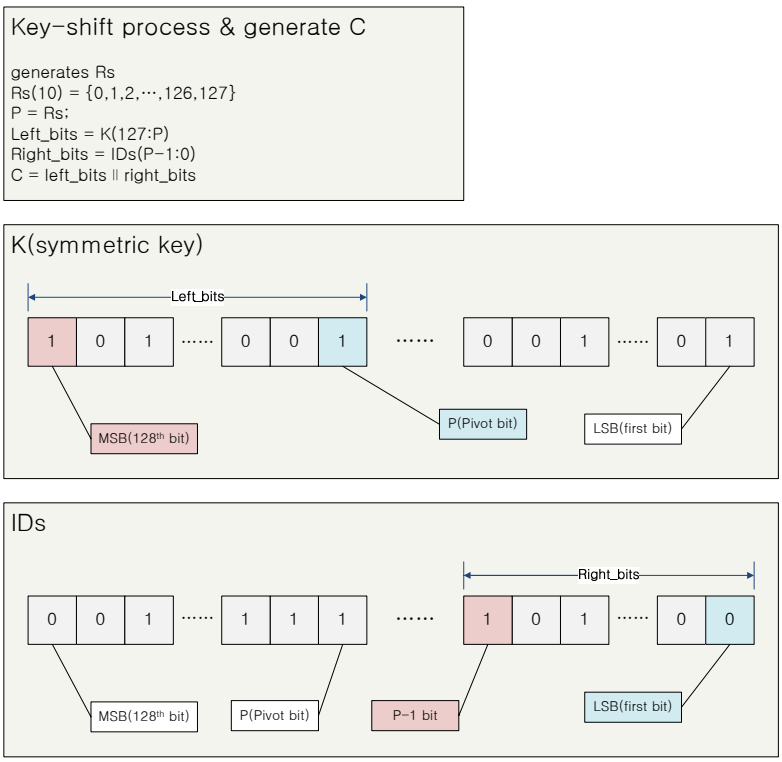


Fig. 3. The Key-shift Process and C to Generate in Proposed Protocol

Step 4. Server → Reader : H(CRC ‖ C ‖ Rr ‖ Rt), Rs, Tag Info
The server use IDs, Rr and Rt saved to calculate by to find whether there is a H(IDs ‖ Rr ‖ Rt) which satisfies the equation H(IDs ‖ Rr ‖ Rt) = H(IDt ‖ Rr ‖ Rt) or not If H(IDs ‖ Rr ‖ Rt) = H(IDt ‖ Rr ‖ Rt), it means that the tag is a legal one. The server sends H(CRC ‖ C ‖ Rr ‖ Rt), Rs, and Tag Info to the reader.

Step 5. Reader → Tag : H(CRC ‖ C ‖ Rr ‖ Rt), Rs
The Reader obtained Tag Info, and sends H(CRC ‖ C ‖ Rr ‖ Rt), Rs to the tag. After receiving

H(CRC ‖ C ‖ Rr ‖ Rt), Rs, the tag computes C, CRC, and the server is authenticated if H(CRC ‖ C ‖ Rr ‖ Rt)' is verified. If H(CRC ‖ C ‖ Rr ‖ Rt) = H(CRC ‖ C ‖ Rr ‖ Rt)', it means that the reader is a legal one. The tag-to-reader authentication process is successful. Until now, the mutual authentication process is completed.

Safety and Computation Analysis of The Proposed Protocol

1) Eavesdropping : In the process of the proposed protocol the information has been encoded by hash function which makes the adversary to get the original value impossible because of the one-way characteristic; In the process of (2) and (5), the data is dealt with hash operation, so the attacker also can't know the real information.
2) Location Tracking : In the proposed protocol, the response message of the tag changes in every session and the output of the hash function is not predictable. Therefore, location tracking is impossible.
3) Spoofing and Replay Attack : The proposed protocol uses Rr, Rt, C and CRC. Even if the adversary has acquired the messages of the previous session, as Rr, Rt, C and CRC change every session, it is impossible to obtain authentication from the legitimate reader and tag.

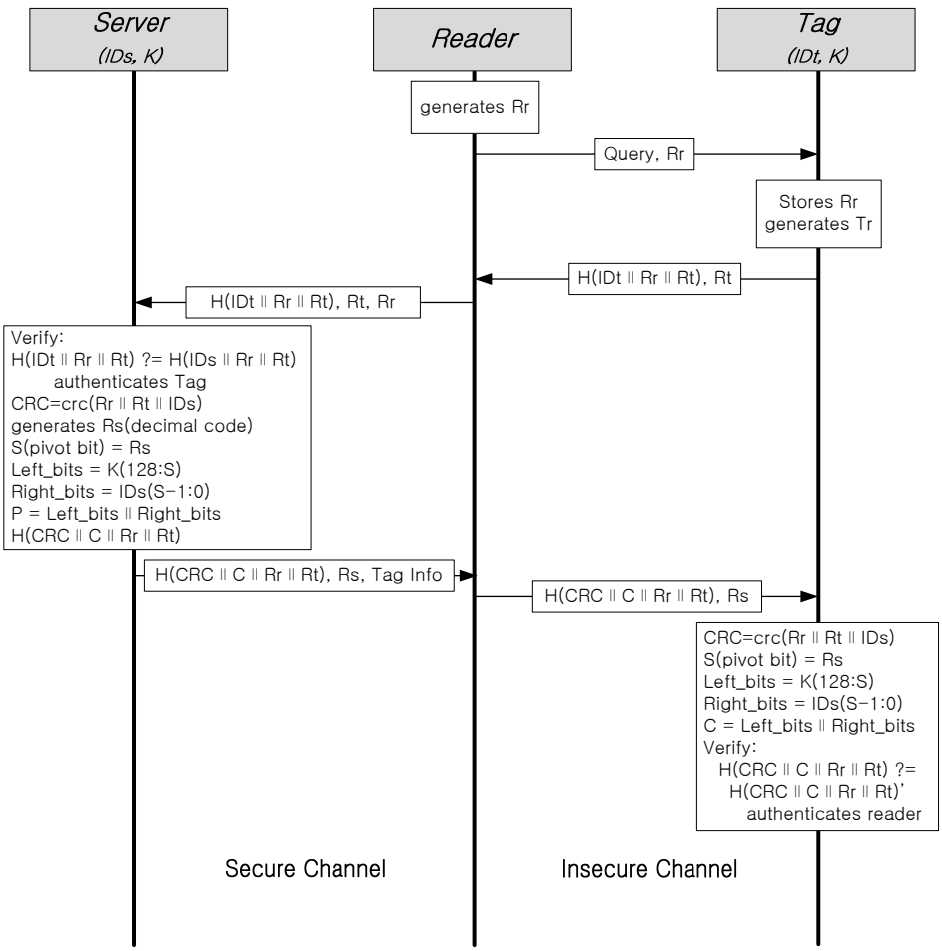


Fig. 4. The Proposed Protocol

4) Efficiency : The existing protocols basically require more than two hash computations and some of them additionally perform XOR computation. However, the proposed protocol needs not XOR computation, and reader not perform any computation.

Safety and computation analysis of the proposed protocol is shown in Table 2 and Table 3.

Table 2. Comparison of Efficiency

Protocol	Eavesdropping	Location Tracking	Spoofing Attack	Replay Attack	Tag's ID exposure	Mutual Authentication
SRAP	Safe	Unsafe	Unsafe	Unsafe	Unsafe	Partially
HMAP	Safe	Unsafe	Unsafe	Unsafe	Unsafe	Unsafe
The proposed protocol	Safe	Safe	Safe	Safe	Safe	Safe

Table 3. Comparison of Efficiency

Protocol	SRAP	HMAP	The proposed protocol
Random number	1R	1R	1T, 1R, 1S
XOR Computation	1T, 1R, 3S	-	-
Hash Computation	2T, 1R, |n/2|+2S	2T, 1R, |n/2|S	2T, |n/2|+1S

Conclusion

This paper designed a RFID mutual authentication protocol based on hash function. The proposed protocol has good security and privacy protection properties. The computation results have shown that the proposed protocol is more secure than previously protocols and has practical advantages over them. By using hash function, random number, CRC, key-shift computation, the data secrecy of a tag is protected from eavesdropping, location tracking, spoofing attack and replay attack.

References

[1] J. Aragones, A. Martinez-Balleste, and A. Solanas. A brief survey on rfid privacy and security. In World congress on Engineering, 2007.

[2] S. Weis, S. Sarma, R. Rivest, D. Engels, Security and privacy aspects of low-cost radio frequency identification systems, in: International Conference on Security in Pervasive Computing, pp. 201-212, Mar. 2003.

[3] T. Yu, Q. Feng, "A Security RFID Authentication Protocol Based on Hash Function", 2009 International Symposium on Information Engineering and Electronic Commerce, pp. 804-807, May. 2009.

[4] L. Liu, X. Lai, D. Yan, Z. Chen, L. Yang, "Mutual Authentication Protocol Based on Hash Function of RFID Systems", Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, pp. 501-506, July. 2011.

A Low Power 32bit Microcontroller and Its Application on Handheld Financial Transaction Terminal

Yinchao Lu[2, a], Weiwei Shan[2, 1, b], and Haolin Gu[2, c]

[2] National ASIC system and research engineering center, Southeast University

210096 Nanjing, China

[a] lyc@seu.edu.cn, [b] wwshan@seu.edu.cn, [c] ghlnnu@163.com

Keywords: System-on-Chip; hardware; low power; financial system.

Abstract. A 32-bit system-on-chip microcontroller using different kinds of low power and high security techniques is implemented and tested for the application on handheld financial transaction terminal. Four power domains and six power modes are designed to hit the low-power targets and meet different functional requirements. Both chip-level and system-level security and transactional methods are utilized to protect all the terminals in the environment against malicious players and criminals. This chip occupies an area of 20 mm^2 in a 0.18μm CMOS process. Test results show that the microcontroller works reliably and safely at the frequency of 70MHz, performs well in all the power modes and only consumes 1.67μA leakage current in the OFF mode. Its application on handheld financial transaction terminal is also tested, which fulfills all the necessary functions.

Introduction

Low power System-on-Chips (SoC) [1-5] are becoming increasingly important to the financial transaction terminals such as electronic cash registers, PC-based point-of-sale (POS) terminals, Mobile Security Cards and so on. Some low power SoCs such as S3C2410, NXP LPC21XX、LPC22XX are widely used in industry control and handheld applications. However, these chips have no functional module or interface for financial applications and the lack of which has to be compensated by extending peripherals on PCB level, thus complicates the design and results in additional cost.

In this paper, a low power 32-bit microcontroller satisfying the requirements of handheld financial applications is designed. Test results show that this SoC chip fulfills financial application functions as well as a low power ability of consuming only 1.67μA leakage current in the OFF mode, which is suitable for handheld devices.

SoC Design of the Financial Terminal

As to the handheld and battery-supplied applications, low power and security are the most important features. Low power technologies such as clock gating, multiple supply voltages (MSV), dynamic voltage and frequency scaling (DVFS), power gating (PG) are explored in the hardware design. The chip architecture is shown in Fig.1, which is mainly divided into the digital unit part and the analog unit part.

1 Corresponding author. Project supported by the National Natural Science Foundation of China (Grant No. 61006029) and the Natural Science Foundation of Jiangsu province, China (Grant No. BK2010165).

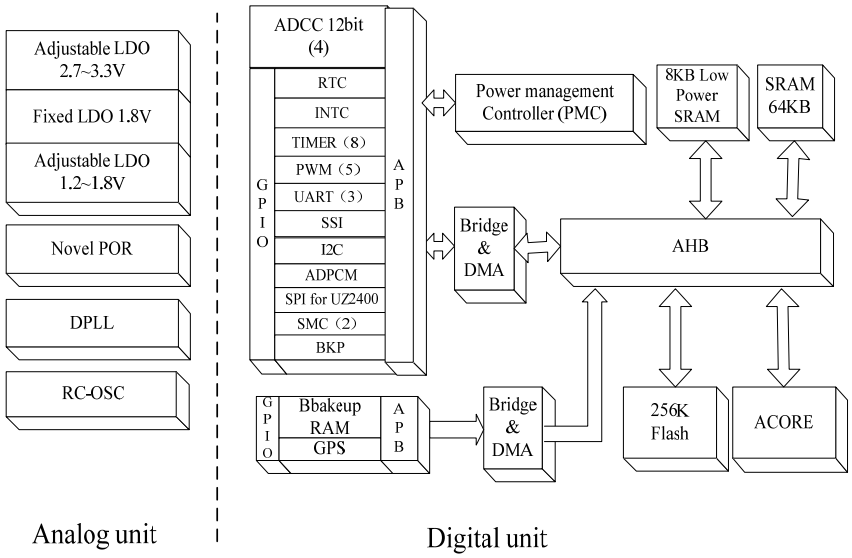


Fig.1 Chip block diagram

The digital unit, which occupies more than 90 percent of the chip's area, consists of a high performance 32-bit microcontroller with a three-stage pipeline structure. Internal memories such as 8KB Synchronous Static Random Access Memory (SSRAM), 64KB ESRAM and 256KB on-chip programmable FLASH are attached to the AMBA High-Performance Bus (AHB). Peripherals like Universal Asynchronous Receiver/Transmitter (UART) interface and sixteen 16-bit backup registers (BKP) are included by way of the 32-bit on chip AMBA Peripherals Bus (APB). In order to meet the financial transaction requirements, an ISO7816-3 and EMV protocols compatible smart card controller (SMC) with flexible programmable interface is designed.

In the analog unit, a low jitter phase-locked-loop (PLL) is integrated for on-chip clock generation. Three low power LDOs are also integrated to generate two adjustable and one fixed output voltages to supply different parts of the chip.

Low Power Solutions. Different power domains are designed with power gating technique [2-3] in our chip to power off certain modules when they are not in use. The power distribution network is mainly divided into four parts as the always-on domain, the memory domain, the digital core domain and the backup domain.

Six typical power modes are designed according to different operations of the four power domains, i.e.: high-speed mode, slow mode, idle mode, low-power mode 1 (PM1), low-power mode 2 (PM2) and OFF mode. A power management controller (PMC), which is responsible for the management of the chip's clock, reset and power supply, controls the processor to shift among different power modes.

Different logic modules are placed in different power domains and supplied separately so that the chip's dynamic power is decreased by way of MSV. Due to the existence of the three LDOs and the DPLL, working frequency and the supply voltage could be adjusted flexibly from 1.2V to 1.8V to achieve DVFS. In addition, by clock gating, the clock signal may be applied to each digital unit when needed so the dynamic power of the ones who have no clock is saved.

The division of power domain also provides a mechanism to eliminate static power. By using Power Gating technique, the non-used blocks are powered off through power gates. But additional isolation cells are demanded to insert at the input of always-on block and the output of three 2K-bit SSRAM to prevent floating, non-powered signals from affecting the powered-down ones.

Chip Level Security Design. A specially designed smartcard controller compatible with the ISO7816-3 and EMV standards supports the whole workflow of a smartcard controller. The simplified and normalized workflow of the controller is shown in Fig.2.

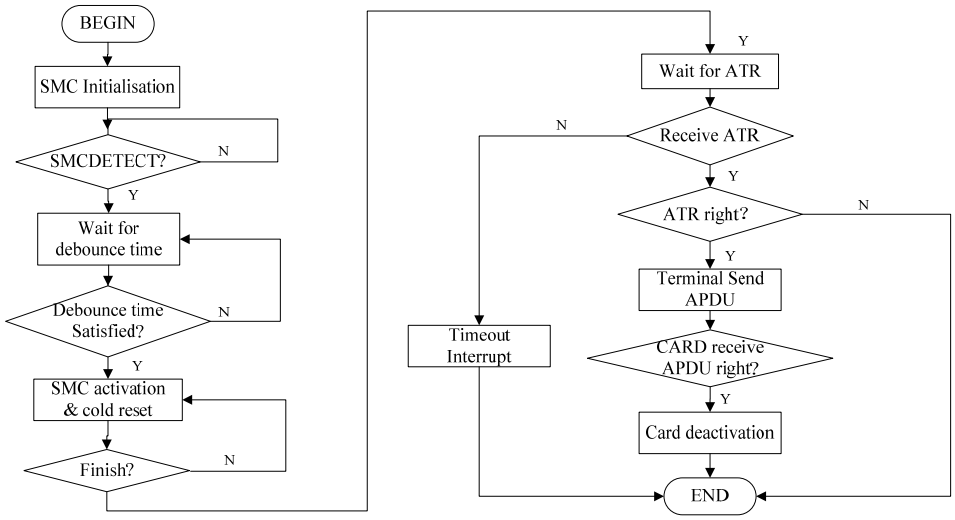


Fig.2 Simplified and normal workflow of SMC

On each working stage of the smartcard controller, from cold reset to card deactivation, if an abnormal case occurs, corresponding interrupt is generated. For example, if the smart card has finished the cold reset process while the SMC has not received ATR sequence fed back from the card after a certain time configured through ATR_WAIT register, the controller will then generate a time-out interrupt to remind the microcontroller to take active steps. The SMC's fast speed, feasibility, plug and play supporting, and automatic configuration characteristics brings more convenience and safety to the financial transaction terminal.

Sixteen 16-bit backup registers (BKP) are designed to store the secure information for encrypt and decrypt, for example, the private key. These data are preserved in the back up domain which is always supplied by the battery even when the main supply voltage is cut off. The data in these registers will never be changed or reset unless been written through the regular bus-operation ways. But if a tamper detection event is detected by the chip's temper pin, the content of the registers will all be cleared right away to assure the information privacy.

Generally, during the back-end design of integrated circuits, SoCs are tested by way of scanning, but many unsafe factors are induced by this way. The internal key information may be observed through the scan chains. Therefore, in this chip a new method is adopted by isolating the interfaces reserved for die test and the JTAG interfaces reserved for debug to prevent illegal test. Furthermore, by covering a layer of metal above certain cells, the SoC is protected from Focused Ion Beam (FIB) destruction to save sensitive data effectively. Besides, a large amount of dummy logic cells are inserted in the chip layout to avoid the layout extraction from the attacker.

ESD protection [4-5] is important especially for chips with smart-card interfaces. In this chip, the ESD clamp circuits with typical RC-based detection, MOS feedback, and cascaded PMOS feedback are embedded in all power/ground/IO cells to construct the I/O ring ESD protection scheme. Besides, in order to enhance ESD protection level, the I/O cells are implemented in a ring structure to ensure that the global ESD bus is continuous throughout the entire I/O domain. Power cells are inserted to form shortest ESD path to provide test modes. With these ESD protection methods, this chip has the highest level of Human Body Model ESD, by passing ±8000V pulses under the ESD test equipment of Thermo Keytek Zapmaster 7/4 tester.

System Design of the Financial Terminal

System Overview. The architecture of the financial terminal system based on our chip is shown in Fig.3(a). Special integrated smart card and magnetic-stripe card (MSC) interfaces which bring higher safety and greater convenience to the software development are designed to collect essential information for the microcontroller to process. Using keyboard, consumers can interact with the financial terminal. By using FSK module or wireless module, the terminal communicates with the financial server. The sensors are integrated to detect the potential dangerous in the environment to protect secret information against crashes and criminals.

In addition, the microcontroller is designed with different types of interfaces which can support external devices. For example, the Electrically Erasable and Programmable ROM (EEPROM) can be extended by I^2C interface to store Chinese Character Library called and displayed by LCD. Furthermore, a lot of general purpose I/Os are designed to support modules such as printer, keyboard and sensors for the financial transaction terminal.

System Level Security Design. As a complement of chip level security strategies, various methods are applied in system level to enhance the security of the total terminal. A special boot procedure is designed to ensure safety burning and updating of the operating system as well as utility software. Dissymmetric-key and PBOC2.0 based encryption algorithms are also certificate each login identity to ensure the integrity and secrecy of the data transferred. Furthermore, all the terminal running states are monitored and all the safety related operations are recorded to prevent the illegal access. Each terminal device binds a physical ID to trace every business deal.

Memories are responsible for a significant percentage of a system's security features. According to the current problems of terminal data protection, memory auto-placement is put forward. Target binary file generating by the compiler is executed in the off-chip memory to get the corresponding CPU access information. Then the bin-format file is divided into a series of data and instruction nodes according to the link information and CPU access record. Converted by compound matrices, a new bin-format file is derived. By replacing the storage location of the executable file, the power distribution is changed when the CPU runs a program, so the nearly homogeneous power characteristics have a very strong immunity from different kinds of power attack.

Experimental Results

Fig.3 (b) shows the photograph of the SoC die. This chip was fabricated in a 2-poly-6-metal 0.18-micron CMOS process. Fig.3 (c) presents the platform built to test the whole chip's performance. All the features below are measured at room temperature. The typical characteristics are presented in Table 1. The typical measured power consumption in different power modes is also shown in Table 1 with only 1.67μA leakage current in the OFF mode.

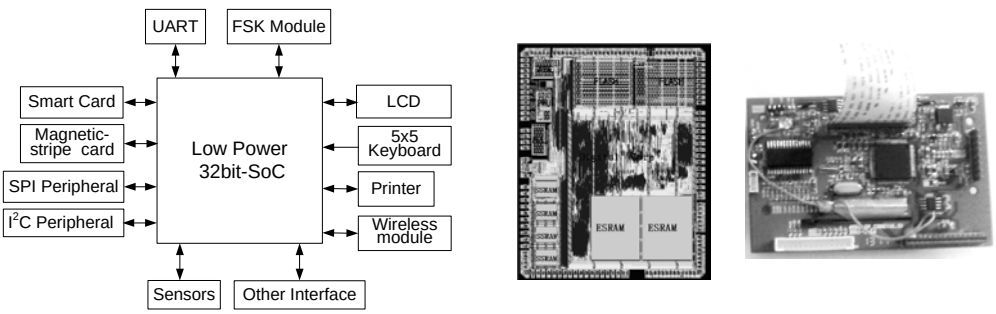


Fig.3 (a) Architecture of the financial terminal (b) Chip die photo (c) Measurement Platform

Table 1. System-on-a-chip characteristics

	Architecture	Die Size	Transistors	Package	Power Consumption
Charac-teristic	32-bit RISC CPU	$20mm^2$	1.3million	QFP100pin	High-Speed 68mA@70MHz PM2:15μA; OFF:1.67 μA

Table 2. Measurement methods of the proposed financial terminal

Modules	Measurements
SMC	Communicates well with the smart card with display information on the LCD
FSK	Fulfills the functions of a telephone
MSC	Communicates with the bank card and displays message on the LCD
UART	Communicates with personal computer and prints information on the screen
LCD	Displays menus and related results
Printer	Connects to thermal printer and print transaction record
Keyboard	Selects menus and manipulates peripherals

Important interfaces are connected to corresponding peripherals to perform their functions. For example, a smart card can communicate with the CPU by SMC interface integrated in the SoC. Measurement methods based on the financial terminal demonstrated in Fig.3 are listed in Table 2. All the modules work cooperatively to achieve the performance and functional requirements of the financial terminal.

Conclusions

This paper presents the implementation of a low power SoC for the handheld financial transaction terminal. Multiple low power techniques such as clock gating, DVS, MSV and PG are applied to satisfy battery-supplied requirements. Both chip level and system level security and transactional means are used to enhance the security features of the terminal. In addition, the SoC shows good performance at different power modes and only consumes 1.67μA leakage current in the OFF mode. Therefore, it is suitable for application on handheld financial transaction terminal.

References

[1] L. Chandrasena, P. Chandrasena, and M. Liebelt. An energy efficient rate selection algorithm for voltage quantized dynamic voltage scaling [C]. Proc. ISSS, 2001, pp. 124–129.

[2] A. Mukheijee et al. Clock and Power Gating with Timing Closure [J]. Design & Test of Computers, IEEE, May-June, 2003, vol. 20, no. 3.

[3] P. Royannez et al. 90 nm low leakage SoC design techniques for wireless applications[C]. IEEE ISSCC'2005 Dig., San Francisco, CA, Feb. 2005, pp. 138–139.

[4] Ming-Dou Ker. ESD (Electrostatic Discharge) Protection Design for Nanoelectronics in CMOS Technology [C]. Advanced Signal Processing, Circuits, and System Design Techniques for Communications, 2006, pp. 217.

[5] Ming-Dou Ker, Cheng-Cheng Yen. Investigation and Design of On-Chip Power-Rail ESD Clamp Circuits without Suffering Latchup-Like Failure During System-Level ESD Test [J]. IEEE J. Solid State Circuits, Nov. 2008, Vol. 43, pp. 2533.

Digital Video Watermarking Algorithm Based on Blocked Wavelet Transform

Sun Cheng[1,a], Gao Fei[1,b], Gong Zhaoqian[1,c]

[1]School of Information and Electronics, Beijing Institute of Technology,
5 South Zhongguancun Street, Haidian District, Beijing 100081, P. R. China

[a]chpsun@bit.edu.cn, [b]gaofei@bit.edu.cn, [c]qsir007@sina.com

Key words: Digital Video Watermarking, Blocked Wavelet Transform, DWT, Robustness, Human Visual Characteristic

Abstract: The rapid development of computer technology and network communication, all kinds of digital products bring to people's life conveniently, but the digital product copyright protection issues have become increasingly prominent. In order to protect the copyright of digital video more effectively, this paper presents a method of digital video watermarking algorithm based on blocked wavelet transform. The algorithm take a video image frame into blocks, obtain the watermark embedding feature block according to low-frequency wavelet coefficients, finally embed the watermark information by modifying the feature wavelet coefficients. After simulation, this algorithm has good invisibility, and can withstand the MPEG-2 compression, so it has strong robustness.

Introduction

In recent years, the appearance and wide application of digital products makes criminals can obtain easily the copy of genuine digital products and spread digital products illegally, therefore, the rights of digital products supplier are infringed. As an important means to protect digital products, digital watermarking technology attracts wide attention of scholars, and a lot of research has been done on the static image watermarking. Digital video can be considered as a consecutive sequence of static images in time domain, at present most of the static image watermarking technique can be applied directly into the video watermark. But the video sequence has a large amount of information, real-time processing and the need for compressed coding, which make the video watermarking technique has its particularity [1].

Time-frequency localization, the specific feature of wavelet transform, make DWT achieve better results than the traditional DCT transform in the image processing field, especially in the digital image and video coding field [2]. In this paper, we design blocked digital video watermarking algorithm on the basis of two-dimensional Discrete Wavelet Transform (DWT). In Section 2，we describe the concept of Blocked DWT. Section 3 explains implementation of digital video watermarking algorithm. Section 4 explains the experimental result of this video watermarking algorithm. Finally, we conclude the paper in Section 5.

Blocked Discrete Wavelet Transform

Two Dimensional Discrete Wavelet Transform. The DWT of a one-dimensional sample signal x is calculated by passing it through related filters: low pass filters and high pass filters. The result from the high-pass filters are detail coefficients and the result from the low-pass filters are approximation coefficients. Expanded to two-dimensional signal, the 2-D DWT in every level is decomposed the

image into four sub bands: LL, HL, LH and HH. They represent low frequency part and edge property of the original image.

The pyramid structure diagram of three-level wavelet transform is shown in Fig. 1. As can be seen from here, from the beginning of the second level, each level of wavelet transform take the low frequency coefficient matrix of last level as the original image. Thus, we can get a series of wavelet coefficients of different resolution, which facilitates the subsequent analysis research.

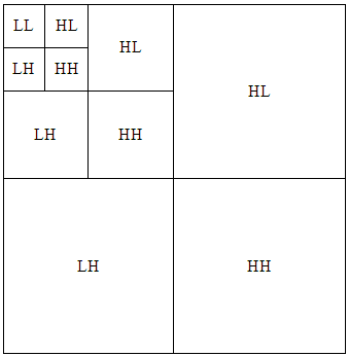


Fig. 1 Wavelet Decomposition

Blocked Wavelet Transform and Characteristic Value Calculation. From the front of this paper, an image is decomposed into four different spatial directional images after once two-dimensional wavelet transform. In this, LL is low frequency part of the image, it is the main energy distribution area and reflecting the average luminance of the original image; LH is high frequency that has edge property of horizontal direction, HL is high frequency that has edge property of vertical direction and HH is high frequency that has edge property of diagonal direction [3]. These bands are not completely irrelevant, they are the description of the same edge, outline and texture image information in different directions and different resolution.

Cox once put forward the important principle of strong watermarking algorithm [4]: In order for a watermark to be robust, the watermark should be placed explicitly in the perceptually most significant components of the original data. In the frequency domain space, the significant part is the low frequency component, but this method leads to lack of invisibility. If the watermark information is added to the most significant components (such as high-frequency coefficient), which can ensure the image quality loss minimum, but it would induce low robustness. Therefore, take into account the watermark robustness and invisibility, this paper put forward the watermark sequence is embedded into the small pieces of complex texture, and takes the low frequency wavelet coefficient to distinguish complex texture feature of blocked images. The blocked method of Video frame and characteristic value calculation process are as follows:

(1) The video image frames are divided into blocks, each size is m × n, denote by W_t. In order to the small pieces contain enough texture characteristic, the value of m and n must be appropriate.

(2) Process two-dimensional discrete wavelet transform for each small block respectively, the selected wavelet is Haar wavelet. Each block gets the wavelet coefficients after wavelet transform.

(3) Calculate standard deviation of the elements in low frequency sub-band LL_3 as the block texture feature value:

$$S_t = STD(LL_3(t)) \tag{1}$$

Where, t $(1 \leq t \leq L)$ represents the serial number of block. L represents the total block numbers of each video frame; $LL_3(t)$ represents the third level's low frequency coefficient matrix of the block image of No. t. STD (•) is used to calculate the standard deviation of all elements of matrix.

Through the above process, we divide the video image frame into many small pieces, and calculate

the values of texture feature of each piece, and then we can embed watermark sequences into original video frame.

Video Watermark Algorithm

Watermark generation. The watermark is inserted in intermediate frequency DWT domain by using Chaotic Sequence. Chaos is a deterministic, random-like process found in non-linear, dynamical system, which is non-period, non-converging and bounded. Moreover, it has a very sensitive dependence upon its initial condition and parameter [5]. When the initial value has slight change, it will have almost completely different chaotic sequences [6].

A chaotic map is a discrete-time dynamical system running in chaotic state.

$$x_{k+1} = f(x_k), \quad 0 < x_k < 1, \ k = 0, \ 1, \ 2, \cdots \tag{2}$$

The chaotic sequence { x_k: k=0,1,2, ⋯} can be used as watermark sequence.

Due to the chaotic map has good sensitivity to initial conditions, we can take the initial value as the key, and generate the chaotic sequence as the watermark information, and then insert the sequence into the characteristic positions of original video image. It can improve the security of digital watermark; to a certain extent, can be very good to prevent unauthorized users to extract the watermark successful. We generate a 64-bit sequence *watermark0* (consist of 1 and -1) through the Eq.2, and this chaotic sequence will be the watermark.

Watermark insertion. As is known to everyone, component Y of video sequence is the brightness information of video, component U and V are two chrominance components of video, and component Y contains most of the information of the original video. Therefore, the new algorithm is inserting the watermark into component Y of the original video sequences. Fig. 2 shows the whole watermark embedding process.

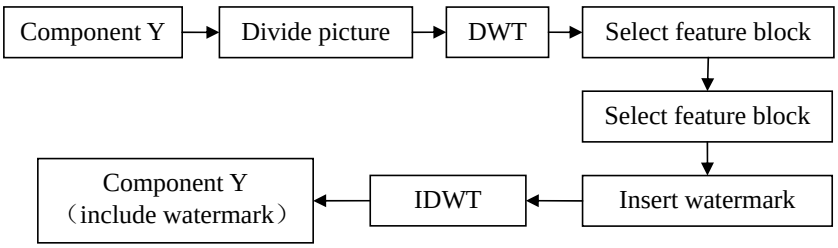


Fig. 2 Watermark Embedding Process

Description of embedding algorithm:

(1) Divide image: Read the Y component of the original video frames, and divide the Y component into 16 × 16 blocks, denote by *W(t)* (t is the serial number and *t=1, 2, 3… L*).

(2) DWT: Process two-dimensional discrete wavelet transform for each block respectively, the selected wavelet is Haar wavelet. We can get the wavelet coefficients of each block finally.

(3) Calculate feature value: Calculate standard deviation of all elements of low frequency sub-band LL3 as the block feature value, denote by S (t) (t=*1, 2, 3… L*).

(4) Select feature block: Select 64 feature blocks according to the value of S (t) and save the corresponding serial number. The selected block will be the watermarking embedding region.

(5) Embedding watermark: according to the serial number saved in step (4), we find the corresponding coefficient matrix and embed the watermark sequences into the intermediate frequency sub-band HL_2. If the watermark symbol is 1, we make $|HL_2(2, 3)| > |HL_2(3, 2)|$; If the watermark symbol is -1, we make $|HL_2(2, 3)| < |HL_2(3, 2)|$.

(6) IDWT: After the watermark sequence are embedded, process two-dimensional discrete wavelet inverse transform respectively for the 64 blocks, then get the new component *Y'* that contain watermark information, finally re-synthesis for the YUV video frame.

(7) Repeat step (1) to step (6), until all the video sequences are embedded completely.

Watermark Extraction. Due to the large amount data of video sequences, we cannot get the original video data generally when watermark detect operation is taken, and it is impossible to store the original video data. We recommend use the blind detection method [7]. The watermark extracting process is the inverse procedure of the watermark embedding process. The step (5) should be replaced by the following: Find the corresponding block and the intermediate frequency coefficient matrix HL_2, compare the value of $|HL_2(2, 3)|$ and $|HL_2(3, 2)|$. If $|HL_2(2, 3)| > |HL_2(3, 2)|$, the extracted watermark symbol is 1, otherwise the extracted watermark symbol is -1.

For a video sequence, we can get an extracted watermark sequence *WD (j)* for every frame of the video sequence (*j* is the frames number of video sequence). Then, we need to synthesis the watermark sequences *WD (j)* in order to get the watermark information *watermark1* finally.

Experimental Results

Computer simulations were carried out to demonstrate the performance of the proposed algorithm. We used the Walk video sequence as the test video sequence. Each frame is of 352 × 240. The watermark is a 64-bit chaotic sequence.

Invisibility. The peak signal to noise ratio (PSNR) was used as an objective measure of the invisibility. Usually, if the value of PSNR is more than 35dB, we couldn't tell the difference in vision between two video frames.

original video sequence

watermaked video sequence

Fig. 3 the original frame and the watermarked frame

Fig. 3 shows the original frame and the watermarked frame of Walk video sequence respectively, It shows that the proposed algorithm result in an almost invisible difference between the original frame and the watermarked frame subjectively.

Fig. 4 represents the PSNR result of 100 frames watermarked by the proposed algorithm. It shows that the proposed algorithm result in an almost invisible difference between the original frame and the watermarked frame objectively.

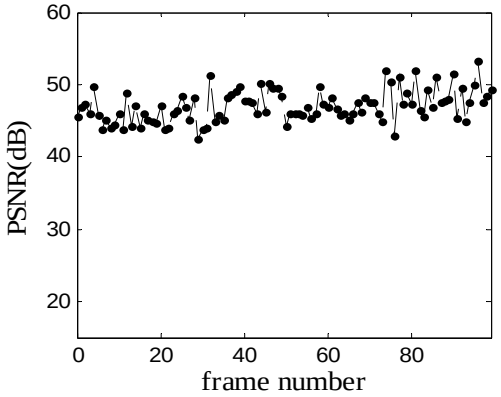


Fig. 4 PSNR of 100 frames watermarked by proposed method

Robustness. In order to demonstrate the robustness performance of the proposed algorithm, we divide the video sequence into several video shots that consist of 25 or 50 or 75 or 100 frames. And we use the normalized correlation (NC) as the objective measurement of robustness. The NC is

$$NC = \frac{\sum_i W_0(i) W_1(i)}{\sum_i (W_0(i))^2} \tag{3}$$

Where, $W_0(i)$ is the original watermark sequence and $W_1(i)$ is the extracted watermark sequence.

Table 1 the average NC for several frames after the various attacks

frames	No process	MPEG-2 coding	Frame Averaging	Frame Dropping
25	0.984	0.890	0.734	0.906
50	0.984	0.937	0.812	0.921
75	1	0.968	0.896	0.937
100	1	0.984	0.906	0.953

Table 1 shows the average NC for several frames after the various attacks. From the experimental data, it is known that the algorithm has high extraction accuracy, and the proposed algorithm is robust against all these attacks especially against MPEG-2 coding and frame dropping. Along with the increase of the video sequence's length, the proposed algorithm could revise the watermark sequence result in improve gradually the watermark extraction accuracy.

Conclusion

This paper adopts the block wavelet transform coefficients to divide the different texture regions, embed and extract the watermark information in the intermediate frequency coefficient matrix by using the stability of intermediate frequency. This method can effectively resist the MPEG-2 video compression and frame dropping. At the same time, the calculation complexity is not high and save memory storage space because of blind watermark extraction algorithm. With the approval experimental results show the algorithm has good invisibility and robustness.

References

[1] Zhao Jie, Gao Fei, Su Guang-chuan, Zhou Xing-fu: Video Watermarking Algorithm Based on Feature points of Moving Regions. J. Microcomputer Information. 27, p.210--212 (2011)

[2] Jin Xiao-hua: Video Watermark Technology Based on DWT Domain. J. Journal of Suzhou Vocational University. 20, p.19--22 (2009)

[3] Xie Yong-hua, Chen Fu-bin, Zhang Sheng-liang, Yang Jing-yu: Feature Extraction and Recognition of Human Face based on Intersected Wavelet Transform and SVD Threshold Compression. J. Computer Applications and Software. 25, p.30--32 (2008)

[4] Ingemar J. Cox, Joe Kilian, F. Thomson Leighton, Talal Shamoon: Secure Spread Spectrum Watermarking for multimedia. ICIP' 97, vol. 6, p.1673--1687(1997)

[5] H. G. Schuster: Deterministic Chaos, an introduction. Second Revised Edition, Physica- Verlag GmbH, Weinheim, Germany (1988)

[6] Yang Hua-Qian, Zhang Wei, Wei Peng-Cheng, Huang Song: An Image Encryption Scheme Based on Random Chaotic Sequence. J. Computer Science. 33, p.205--209 (2006)

[7] Lan Hong-xing, Chen Song-qiao, Hu Ai-na, Li Tao-shen: Research on the Second Generation Digital Watermarking Algorithm Based on DWT Domain. J. 35, p.1799--1803 (2007)

Data Mining based Crime-Dependent Triage in Digital Forensics Analysis

Rosamaria Bertè[1,a], Fabio Marturana[1,b], Gianluigi Me[1,c], Simone Tacconi[2,d]

[1]Department of Computer Science, Systems and Production
University of Tor Vergata, Rome, Italy

[2] Servizio Polizia Postale e delle Comunicazioni, Rome, Italy

[a]rosamariaberte@libero.it, [b]marturana@libero.it, [c]me@disp.uniroma2.it,
[d]simone.tacconi@interno.it.

Keywords: Computer Forensics, Data Mining, "Post-mortem" Triage.

Abstract. Over the last few years, law enforcement registered a growing number of crimes related to the worldwide diffusion of high storage capacity low-cost digital devices. As a consequence Computer Forensics, the investigative discipline that aims to find evidence among seized devices is becoming increasingly complex. In this paper, we propose a new approach to digital investigations, based on the application of Data Mining and Knowledge Management theory, which aims to give a theoretical foundation to "*post-mortem*" Triage. This new practice has the potential utility to speed-up investigations by assigning a priority to each seized device, with a positive impact on the upcoming forensic analysis. The paper shows how the proposed methodology could create intelligence from the extracted data and predict the model's dependent variable (i.e. the *class*) and its relation with the independent variables (i.e. *system configuration files*, *installed software*, *file statistics*, *browser history* and *system event log*). We identify the *class* variable with the likelihood that a computer has been used to commit specific crimes such as *child pornography*, *copyright violation*, *hacking*, *murder* and *terrorism*. The paper is based on a case study carried out in collaboration with the *Servizio Polizia Postale e delle Comunicazioni* (the Italian Cybercrime Police Unit).

1. Introduction

Computer Forensics is the application of specific investigative techniques aiming to analyze seized computers and gather evidence for presentation in a court of law. The goal of Computer Forensics is therefore to perform a structured investigation to track computer system and user activity, maintaining a documented chain of evidence.

In 1955, in an article in *The Economist*, Cyril Northcote Parkinson first wrote that "*work expands so as to fill the time available for its completion*" (*Parkinson's Law*). The article was referring to public administration but today's corollary to this law might be that "*computer forensic examinations expand in proportion to the increase in size of forensic units thus maintaining a significant backlog*" [1].

Considering how the large availability of low-cost, sophisticated and heterogeneous digital devices with large storage capacity has contributed to the spread of computer crimes [2], we can

certainly argue that Parkinson's Law corollary is particularly true dealing with Digital Forensics. A serious problem complained by law enforcement specialists, indeed, is that, considering the effort to analyze hundreds of Terabytes of data, usually only few records of relevant intelligence about the crime under investigation could be extracted. The reasons are basically twofold: on one side, the enormous amount of available data to be processed in the lab and, on the other, the forensic investigator's habit to look for potential evidence by means of traditional, manually intensive and time-consuming procedures. As a consequence, finding theories and techniques with the aim to narrow the area of interest and deepen the search only where needed is considered a crucial aspect to reverse this negative trend.

In this paper, we propose an application of Data Mining theory to "*post mortem*" Computer Forensics Triage, inheriting some concepts from a previous research in the field of Mobile Forensics whose effectiveness has already been tested and verified in the field [3,4].

The proposed approach aims to build a priority list among the seized computers, before being processed in the lab, highlighting their relative relevance according to a crime-dependent three-dimensional model of categorization concerning *timeline*, *crime's features* and *suspect's private sphere* (habits, skills and interests). This new point of view could represent a complementary activity in digital investigations aiming to identify immediately the most relevant computers.

"Post mortem" Computer Forensics Triage consists of the following four phases: *forensic acquisition*, *feature extraction and normalization*, *context and priority definition*, *data classification and triaging*. The first one is the classical forensic hard disk image creation. The second is devoted to acquire the whole set of available features (system configuration files, installed software, file statistics, browser history and system event log) from the disk image and normalize them creating a two-dimensional matrix, called *complete matrix*. The third is in charge of introducing in the model the timeline of interest and the crime-specific features in order to focus the attention only on a part of the aforementioned matrix, called *reduced matrix*. The fourth is assigned the task of analyzing the reduced matrix's features and calculating the *class* variable by means of Data Mining algorithms.

2. Related Work

Over the past few years, Computer Forensics has been supported by several theories and methods developed in order to find evidence quickly on seized computers as far as specific crimes such as murder, child abductions, missing persons, death threats etc. are concerned. In such cases the need for the timely identification, analysis and interpretation of digital evidence is crucial since it could be the difference between life and death for the victim.

As far as Mobile Forensic is concerned, our research group has recently proposed two possible applications of Data Mining based classifications to "*post mortem*" Triage.

The first one [3] concerned a methodological approach to Mobile Forensics in order to identify the most interesting mobiles from an investigative point of view by predicting the *device owner's usage profile*. The proposed methodology can help investigators split up relevant and less important aspects of the case under investigation by assigning a priority to every involved device, person and crime. By means of a quick memory search on the whole set of seized devices it is possible to create a list of phones, ordered by probative value, which require additional processing at Forensic Lab.

The second study [4], extensively discussed with Italian law enforcement cybercrime specialists, was based on Mobile Forensics Triaging and self-knowledge algorithms for mobiles classification and concerned a viable methodology to determine the likelihood that a mobile phone has been used to commit child pornography.

In 2006 a research group proposed a new methodology called *Cyber Forensic Field Triage Process Model* (*CFFTPM*) [5], which deals with "live" on-site or field activities for providing the identification, analysis and interpretation of digital evidence in a short time frame, without the

requirement of taking the system back to the lab for an in-depth examination or acquiring a complete forensic image. The proposed methodology, although entailing a real risk of exhibit pollution, is justified by the need to provide investigative leads quickly in time critical situations.

Recently a new trend combining computer forensic principles and Data Mining statistical approach & knowledge discovery process is taking hold in the research community. According to recently published papers, the proposed model allows to extract, store and analyze digital devices data with forensically sound methods that could be used in a court of law [6].

Finally an important research was conducted about the application of Data Mining theory to digital text analysis, based on clustering text mining techniques for investigational purposes. The work addressed text clustering for forensics analysis based on a dynamic, adaptive clustering model to arrange unstructured documents into content-based homogeneous groups [7,8].

3. Proposed process model

This paragraph describes the proposed four-phases model allegedly applied on a set of data extracted from forensic images of seized hard disks and regarding crimes such as child pornography, copyright violation, hacking, murder and terrorism.

The whole process is carried out in a forensic lab in parallel with the traditional forensic procedures (acquisition, retrieval and analysis) [9,10,11] and is summarized in the following figure:

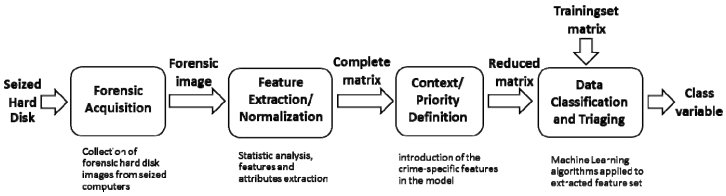


Fig.1 – "Post-mortem" triaging model

The *first stage* of the process is, therefore, the *forensic acquisition*. According to NIST [12], indeed, the first step in a Computer Forensics investigation is to make a disk image in order to preserve digital evidence integrity and guarantee the analysis repeatability.

The *second stage* of our workflow, called *feature extraction and normalization*, is in charge of extracting relevant data from disk images.

This stage of the process is inspired by official best practices concerning computer incident response which strongly suggests to:

- Look at the last date of change on critical files;
- Examine configuration and start-up files;
- Look for hacking tools (password crackers, copies of passwords, etc.);
- Examine the password file for unauthorized accounts;
- Search for keywords appropriate to the incident;
- Search for hidden areas, slack space, and cache;
- Look for changes to files, critical file deletions, and unknown new files;
- Collect a list of all e-mail addresses, FTP sites and URLs, visited from the computer.

We concentrate on the following computer usage parameters: *system configuration files*, *installed software*, *file statistics*, *browser history* and *system event log* which represent the independent variable of the model. Occurrences of each parameter, for instance the number of deleted images or installed hacking tools, the presence of a system log file or the number of visited URL etc. is counted and summarized within a two-dimensional matrix called *complete matrix*.

During drive image analysis, we suppose to collect the whole set of *features* concerning user's

habits, technical skills and interests (*suspect's private sphere*).

We propose to associate user's habits with the following set of parameters:

- percentage of modified files, ordered by time slot (morning, afternoon, evening, night);
- internet connections, ordered by time slot;
- monthly login frequency;
- system utilization ordered by time slot.

With regards to user's technical skills, we consider:

- system configuration files;
- system log settings;

With regards to user's interests we focus our attention on:

- stored and deleted files statistics (audio, video, images, documents, executable);
- installed applications statistics;
- visited URLs statistics.

The *third stage* of our workflow is called *context and priority definition* since it introduces in the model the *timeline* of interest (i.e. the time frame during which the crime happened and thus we suppose to find more evidence) and the *crime-specific features* (i.e. the crime related fingerprint such as the presence of child images in a child pornography case or illegally downloaded software or movies in a copyright violation). After the aforementioned three stages we will be able to create the following data structures (*complete* and *reduced matrix)* which summarizes the set of parameters that could be processed by the model:

	Computer #1	Computer #2	Computer #3
Feature #1	true	true	false
Feature #2	true	true	false
Feature #3	false	false	false
Feature #4	10	20	100
Feature #5	1000	1500	0
Feature #6	true	true	false
.....	-	-	-
.....	-	-	-
Feature #N	1	1	10
Class	-	-	-

	Computer #1	Computer #2	Computer #3
Feature #1	true	true	false
Feature #3	false	false	false
Feature #4	10	20	100
.....	-	-	-
.....	-	-	-
Feature #N	1	1	10
Class	-	-	-

Fig.2 - complete and reduced matrix

The feature nickname is indicated in the leftmost column while the other columns show a set of instance samples of retrieved hard disk images.

The *fourth stage* of our workflow is called *data classification and triaging* and it is in charge of elaborating the reduced matrix in order to provide the final classification of the input data. According to our experience, this phase could be based on a collection of machine learning algorithms for data mining tasks such as Waikato Environment for Knowledge Analysis [13].

4. Methodology algorithmic foundations

We assume that our model's dependent variable, called *class*, is the likelihood that a computer has been used to commit the crimes of *child pornography*, *copyright violation*, *hacking*, *murder* and *terrorism*. The goal is therefore to assign a *class* to each inspected exhibit, a sort of relative score used to create an ordered list of items. To meet the model requirements, during the *context and priority definition* phase, it is necessary to gather the whole set of features extracted during the previous phase, creating a context-dependent features subset, correlated with the crime under observation/evaluation and the specific timeframe i.e. the time the crime occurred. In other words, the goal is to determine the presence of crime's digital fingerprints in the exhibit. Talking about child pornography, for example, there is a high likelihood to find private pictures and videos on the sized computer, while in case of copyright violation the presence of a large number of hard disk

stored music files and films is a relevance indicator from an investigative point of view. In other cases such as hacking crime we could probably find installed firewalls and system or hacking applications while, in case of terrorism, we will search the computer for documents concerning the criminal conduct. Finally, considering murder specific digital fingerprints, we will probably find specific log events or chat/instant messaging stored sessions, social networks visited URLs or received/sent emails etc. The outcome of *context and priority definition* phase is therefore the contextualization of the *complete matrix* which will be deprived of unnecessary items and provided as input to the next classification phase (i.e. *reduced matrix*).

The Data Mining *supervised* classifier requires first a *trainingset*, i.e. a collection of crime-dependent representative patterns with a known class, in order to train the underlying algorithm. Once trained the classifier will elaborate the collection of real patterns.

The first step of data classification and triaging is, therefore, the collection of a consistent set of representative exhibits with a known *class* concerning the crimes that we want to classify (i.e. *trainingset* creation). This is a crucial activity since, if performed by inexperienced analysts, it could negatively influence the whole learning process. With a well-formed *trainingset* at our disposal, it now is possible to train a classifier by adopting, for instance, the iterative and predictive method called *10 folds cross-validation* [13] which splits the *trainingset* into ten approximately equal partitions, each in turn used for testing and the remainder for training. The procedure is repeated ten times so that, in the end, every instance has been used exactly once for testing.

It is also possible to use 10 folds cross-validation output to evaluate classifiers' effectiveness (i.e. the ratio between performance and acquired knowledge) calculating performance indicators such as *Precision*, *Recall* and *F-Measure* [13].

5. Conclusions

This paper deals with a new methodological approach to Computer Forensics, the branch of Digital Forensics concerning evidence extracted from digital devices according to well defined and standardized methods and with probative value in court.

The research questions that we addressed were the following: "*which are the potential benefits of applying Data Mining and Machine Learning algorithms to the actual computer forensic techniques? Can this methods eliminate bottlenecks and increase investigation efficiency?"*.

To answer the questions we interviewed experts and law enforcement, in order to assess their investigation procedures and we realized that the actual forensics lab data analysis process, based on a rigid 4-steps workflow (HD Image creation, Extraction, Analysis and Reporting) implies that, in each case, terabytes of seized data must be searched to isolate a single evidence.

In this paper we propose a new model which redefines the aforementioned forensics workflow, introducing the concept of "*post mortem*" *Triage*. Based on Machine Learning algorithms and data analysis to identify crime-dependent activity patterns not discernible through a manual review process, Triage aims indeed at giving a priority to each inspected computer based upon the likelihood that a computer has been used to commit one of the following crimes: *child pornography, copyright violation, hacking, murder and terrorism*. We identify this likelihood with the model's dependent variable under observation and we call it *class*.

The class is calculated upon a set of independent variables based on *system configuration files, installed software, file statistics, browser history* and *system event log*.

In particular, the proposed crime-dependent categorization model, concerning *timeline, crime's features* and *suspect's private sphere* (habits, skills and interests) consists of the following four phases: *forensic acquisition, feature extraction and normalization, context and priority definition, data classification and triaging*. The latter two phases are the most important since, during the third one we are able to introduce in the model the *timeline* of interest and the *crime-specific features* creating the, so called, *reduced matrix* while in the fourth we analyze the matrix and calculate the *class* variable by means of classifiers, assigning a relative score to each analyzed exhibit.

6. Future Work

Our research relates to the theoretical foundation of Computer Forensics "*post mortem*" Triage and draw the guidelines for future implementations.

The proposed methodology could be implemented with specific digital investigations support tools allowing the automated disk categorization and the creation of a user activities' timeline, based on the WEKA java library downloadable at [14]. It is important to highlight that classification task's success depends on the number of "real" observations and sized data that will be used to build the *trainingset* which is the core of the knowledge discovery process. The higher is the number of analyzed hard disk images, indeed, the better the knowledge discovery process works.

References

[1] H. Parsonage: Computer Forensics Case Assessment and Triage - some ideas for discussion.

[2] IC3 (Internet Crime Compliance Center) – Internet Crime Report (2009).

[3] R. Bertè, F. Marturana, G. Me, S. Tacconi: Mobile Forensics "triaging": new directions for methodology. ITAIS-2011Conference Proceedings, Rome, Italy, ISBN: 978-88-6105-063-1.

[4] R. Bertè, F. Marturana, G. Me, S. Tacconi: A quantitative approach to Triaging in Mobile Forensics. IEEE Computer Society TrustCom-2011 Conference Proceedings, Changsha, China.

[5] M. K. Rogers, J. Goldman, R. Mislan, T. Wedge: Computer Forensics Field Triage Process Model. Conference on Digital Forensics, Security and Law (2006).

[6] Veena H Bhat, IAENG, Prasanth G Rao, Abhilash R V, P Deepa Shenoy, Venugopal K R and L M Patnaik: A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application. IACSIT International Journal of Engineering and Technology, Vol.2, No.3, ISSN: 1793-8236 (2010).

[7] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo R. Zunino: Text Clustering for Digital Forensic Analysis, Journal of Information Assurance and Security (JIAS), Vol. 5, No. 4, pages 384-391, Dynamic Publishers (2010).

[8] S. Decherchi, S. Tacconi, J. Redi, A. Leoncini, F. Sangiacomo R. Zunino: Text Clustering for Digital Forensic Analysis. CISIS-2009 Burgos (Spain), vol. n. 63 Advances in Intelligent and Soft Computing (AISC), Springer.

[9] Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau: Crime data mining: a general framework and some examples. M. IEEE Trans. Computer. 37, 50—56 (2004).

[10] Seifert, J. W.: Data Mining and Homeland Security: An Overview. CRS Report RL31798 (2007).

[11] Mena, J. Butterworth-Heinemann: Investigative Data Mining for Security and Criminal Detection. (2003).

[12] K. Kent, S. Chevalier, T. Grance H. Dang: Guide to Integrating Forensic Techniques into Incident Response. Special Publication 800-86. Recommendations of the NIST.

[13] Ian H. Witten, Eibe Frank, Mark A. Hall.: Data Mining Practical Machine Learning Tools and Techniques. 3rd Edition- Elsevier.

[14] Weka suite. http://www.cs.waikato.ac.nz/ml/weka/

Deployment of QoS Bandwidth Guarantee Based on Burst Traffic Detection Technology

Wang Sunan[1, a], Zhang Jianhui[1, b], Zhao xin[1, c] and Zhang Xiaohui[1, d]

[1]National Digital Switching System Engineering & Technological R&D Center, 450002, Zhengzhou, China

[a]wangsunan@163.com, [b]zhangjh365@gmail.com, [c]zhaoxin@msn.cn, [d]zxhback@163.com

Keywords: QoS; Burst traffic; PPBP model; scale of traffic calculate; Bandwidth planning

Abstract: In the case of abnormal traffic burst, the network link bandwidth is undoubtedly facing a challenge. However, to effectively solve the planning problems of bandwidth, to facilitate the beneficial implementation of network resources deployment and improve the overall network *Quality of Service* (QoS). In this paper, behavior characteristics of abnormal traffic burst were analyzed first, traffic characteristics was detected using PPBP process model, the bandwidth resource planning was solved using limited time-scale set, bandwidth planning was realized in the way of counting local cache information, and bandwidth resources in QoS were guaranteed applying instances of network deployment to efficiently prove the optimality of the network resource deployment.

1. Introduction

In the early 1980s, with the deployment of broadband integrated services digital network (B-ISDN) [1], high-speed data transmission services extended to simple voice service network, in traditional voice service network, the model for each user occupying a fixed bandwidth was no longer applicable. Presently, network connectivity has integrated into all aspects of large-scale commerce in the life of thousands of families with the pace of development more than Moore's Law. It has a non-technology-driven feature so that the network construction, equipment development and business provision related theoretical studies lag behind actual applications, being difficult to achieve their goals under the guidance of systematic theoretical system. It is not only difficult to guarantee the quality of customer service provided but also waste resources often.

Quality of Service, This network security mechanism, it includes transmission bandwidth, transmission delay and data packet loss rate. That can be used to solve question such as network of latency and congestion. Exception for the case of traffic burst, Link bandwidth resource planning is the core issue of how the characteristics of a best-effort service. When the packet transmission network to solve the unexpected burst traffic, QoS of requirement is guaranteed that it need of constraint packet of loss rate and delay rate. We will stand to facilitate engineering implementation point of view, then we completed a systematic work. Start off with burst traffic characteristics researching, Finite Timescale Set analyzing and Link bandwidth resources planning, etc.

2. Traffic Characteristic Detection Model

Timothy Neame first proposed the method of self-similar traffic being described by Poisson Pareto Burst Process (PPBP) model based on traffic burst detection technology in literature [2]. PPBP model shows custom input stream obeys the Poisson process in which the parameter is λ. The system has an infinite number of service equipments, and the service time-T of each equipment is

independent and identically distributed and Pareto heavy-tailed distributed. PPBP constitute the sequences of the total number of customers in queue system on the timeline [3]. The construction of PPBP model is shown in **Fig.1**.

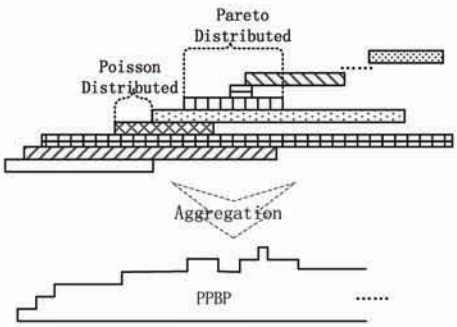


Fig. 1. PPBP model aggregated process.

Random process $\{N_t : N_t \in Z^+, t \geq 0\}$ indicates the amount of traffic bursts of time - t, of which, Z^+ indicates a positive integer. $\{A_T^i : A_T^i \in R, i = 0,1,2,\ldots\}$ indicates traffic arrival time, and $\{L_T^j : L_T^j \in R, j = 0,1,2,\ldots\}$ indicates traffic leave time. If D_T^i expresses the duration of burst traffic, then $L_T^i = A_T^i + D_T^i$ [4]. Though A_T^i is an incremental sequence, the duration of burst traffic A_D^i is a random process, so L_T^i is also random. Therefore, random process X_n indicates the continuous-time random process of the total cumulative arrival amount within the time $[0,t]$, the amount of traffic bursts- N_t of t time can be expressed as follows:

$$N_t = \sum_{i=0}^{\infty} X_i \text{, in which } i \in [A_T^i, L_T^i],$$

$$X_i = \begin{cases} 1, & i \in R \\ 0, & else \end{cases} \tag{2.1}$$

In PPBP model, all durations of burst traffic D_T^i are independent and are random variables obeying Pareto distribution. Then D_T obey:

$$P(D_T > x) \begin{cases} (\frac{x}{\delta})^{-\gamma}, & x \geq \delta \\ 1, & else \end{cases} \tag{2.2}$$

In which, $\delta > 0$, $1 < \gamma < 2$. And the mean of D_T obey:

$$E(D_T) = \frac{\delta\gamma}{\gamma - 1} \tag{2.3}$$

If the burst process is steady, then the initial duration of burst traffic will be D_T^0, and D_T^0 is the Poisson random process which mean is $\lambda E(D_T)$. Durations of these burst traffic are independent. So D_T and L_T have the same distribution, for $i \in \{1,\ldots,D_T^0\}$, L_T meets:

$$\mathrm{P}(L_T > x)\begin{cases} \frac{1}{\gamma}\left(\frac{x}{\delta}\right)^{1-\gamma}, & x \geq \delta \\ \frac{\gamma-1}{\gamma}\left(1-\frac{x}{\delta}\right)+\frac{1}{\gamma}, & else \end{cases} \tag{2.4}$$

In addiction, X_n mean the continuous-time random process of the total cumulative arrival amount within time $[0,t]$. If the arrival rates of all traffic are the mean r, then the amount of traffic bursts depicted by PPBP model can be expressed as:

$$X_n = \lambda r \int_0^t D_T dt \tag{2.5}$$

3. Traffic Scale Computation

The obtaining of traffic scale is to realize the mathematical description of the statistical multiplexing under large-scale streaming media convergence based on QoS guarantees. It is also a prerequisite for the network deployment. Therefore, it can be known from this section on the basis of traffic characteristics analyzed in the previous text that the estimation of traffic scale is essentially decided by the amount of access bandwidth required under the constraints of the QoS metrics. In order to obtain the estimation of traffic scale, first, based on MV-LD operator [5],

$$\log \mathrm{P}\{Q_\tau > b\} = \inf \frac{(b+(C_P-\mu)t)^2}{(-2\sigma_X^2(t))} \tag{3.1}$$

$\mathrm{var}(t)$ is estimated by the transformation of the equation (3.1):

$$\mathrm{var}(t) \approx \inf \frac{(b+(C_P-\mu)t)^2}{(-2\log \mathrm{P}\{Q_\tau > b\})} \tag{3.2}$$

Then, in accordance with QoS constraints:

$$\log \mathrm{P}\{Q_\tau > b\} \approx \log \mathrm{P}\{\sum_{i=1}^{\lambda} X_i \geq nc\} \leq \varepsilon \tag{3.3}$$

Traffic scale is estimated. It is considered that Gaussian aggregated traffic is composed of λ independent and identically distributed sub-streams, and the mean and variance of sub-streams were μ, σ^2, respectively. The queue service rate and the number of service queue are c, n respectively. If λ which is the best condition to meet the above case is the minimum guarantee of aggregated traffic scale, then, according to the nature of variance

$$\sqrt{\mathrm{var}(t)+\mathrm{var}'(t)} \leq \sqrt{\mathrm{var}(t)} + \sqrt{\mathrm{var}'(t)} \tag{3.4}$$

and the formula (3.1),

$$\mathrm{P}\{\sum_{i=1}^{\lambda} X_i \geq nc\} \approx e^{\left(-\frac{(nc-\lambda\mu)^2}{2\lambda\sigma}\right)} \tag{3.5}$$

Obviously, it can be obtained from the formula (3.5):

$$\lambda(n) = \left(\frac{c}{\mu}\right)n - \left(\sqrt{\varepsilon c \frac{\mathrm{var}}{\mu^3}}\right)\sqrt{n} + o(\sqrt{n}) \tag{3.6}$$

$\lambda(n)$ obtained by computing is the approximation of traffic scale when the bandwidth guarantee is $C = nc$ under QoS constraints. It can reference [2].

4. Analysis of Experimental Data

This section, we are through FTS (Finite Timescale Set) demonstrate the effectiveness of the bandwidth planning, so use validity of statistical conclusions to solve traffic problems [7].

First, we collected data on traffic conditions described. We conducted a number of different real traffic collection and analyzed separately. Show Table 1.

Table 1. Collect real traffic.

Name of traffic data	Collecting start time	Collecting duration	Collecting environment	Data File Format
Abilene-I	2008-5-14, 20:00	15 (minute)	OC-48 Backbone Network	ERF
Abilene-III	2010-6-1, 19:31	4 (hour)	OC-192 Backbone Network	ERF
Auckland-VIII	2009-12-1, 18:00	2 (week)	100MAccess Network	ERF
Campus	2007-7-12, 21:16	45 (minute)	100MAccess Network	TCP Dump

The following example for test results to Table1. Abilene-I OC48-080514-2000 trace analysis shows.

Table 2. Trace OC48-080514-2000 content.

Environment	Time	Bytes(G)	Packets(M)	MeanRate(Mb)	TCP Bytes(%)
OC-48 Backbone Network	20:00~20:15	92.8	149.2	622	26.5

Fig.2 Trace OC48-080514-2000 queuing analysis result, then FTS of estimation value. For ease of calculation selected FTS is RaNDS time range $\{2^1,2^2,2^3,2^4,2^5,2^6,2^7\}$. It slows $\mathrm{P}\{Q^{\mathrm{FTS}} > b\}$ can be a good at Gaussianity input traffic (Real traffic) that is queuing performance. Thus, by way of FTS similar analysis, taking into account the time scale of the selected implementation complexity and accuracy of both requirements, it solves validity problems of traffic statistical conclusions. So it ensures the accuracy of the premise, the use of less time-scale.

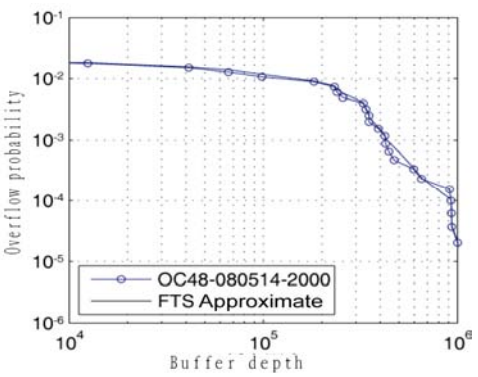


Fig. 2. FTS Approximate.

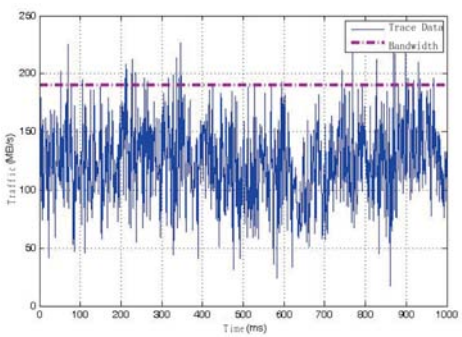


Fig. 3. Timescale of 128ms.

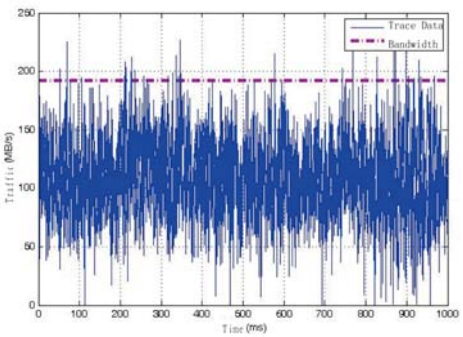


Fig. 4. Timescale of 32ms.

For example trace OC48-080514-2000, FTS given link planning in the timescale: 128ms and 32ms, the results were as follows **Fig.3** and **Fig.4**.

Through two different timescales for the duration of 1000ms, the planning link bandwidths were 195MB and 199 MB. It can provide the protection of expected value $\zeta = 0.01$. Different timescales in which the link bandwidth planning results to differ, in addition the actual statistics $\zeta = 0.008 \, and \, 0.006$ are better than expected.

5. Instances of Network Deployment

The network deployment problem based on the obtaining of traffic scale, from several possible arrangements or schemes is to look for how to effectively plan, manage and control network systems to maximize the efficiency of the optimal arrangements or schemes, known as network optimization problem mathematically, that is, to minimize the system overhead by finding the optimal network deployment scheme under traffic scale meeting the QoS bandwidth constraints. System overhead involved in this section mainly includes the following parts: equipment deployment costs, business deployment costs and bandwidth costs of long-distance transmission. To facilitate the optimization of network deployment, the dynamic programming (DP) method was selected in this section [7, 8], as a branch of operations research, it is the famous optimization method American mathematicians REBellman *et al* proposed in the study of the optimization problem of multistep decision process in the early 1950s. Its core idea is to transform the problems into a multistep process and decompose it into a set of single-step processes, and then, the relationship between steps is used to solve them one by one.

Although DP method is easy to be realized by multistep recursion, a formal DP is not present to solve various optimization problems. It needs to smartly define a model description with the optimal solution for different natures of optimization problems. Because the design of the dynamic planning process is a way or method rather than a special algorithm to solve optimization problems, it does not have a clear and standard mathematical expression. Therefore, detailed analysis was conducted on the optimal network deployment problems in this section to define optimization elements and optimal network deployment strategies were given [9].

As shown in equation (4.1). If the optimal network deployment is simplified as integer program (IP) first for analysis, the optimization goal can be expressed as:

$$\min \sum_{i=1}^{n} c_i x_i \qquad s.t. \;\; Ax \geq b, \;\; x \geq 0, \;\; x \in Z^n \tag{4.1}$$

In which, all parameters of linear IP are integers,

$c, x \in Z^n, b \in Z^m, A \in Z^{m \times n} \; c, x \in R^n, b \in R^m, A \in R^{m \times n}$

A is full row-rank, and $m \leq n, b \geq 0$. Therefore, **Fig.5** shows the algorithm elements of DP idea. Though the sub-strategies of the superior strategies of the possible actual optimization outcome are not necessarily the best, that is, they do not meet the criteria of optimization principle but can be realized by recursive ways based on DP idea and calculation.

Algorithm elements include required goal, variables, control factors and state in deployment algorithm.

- Goal:

$$Min[\sum_{i=1}^{n} S_i x_i + \sum_{i=1}^{n} \sum_{p=1}^{P} P_{ip} y_{ip} + \sum_{i=1}^{n} \sum_{p=1}^{P} C_{ip} z_{ip}] \tag{4.2}$$

- Variables:

  S_i: Device hardware costs deployed at node i;

  P_{ip}: Business p costs deployed at node i;

C_{ip}: No hardware or business deployed at node i, but the bandwidth costs of transmission business p.

- Control factors:

x_i: $x_i = \{0,1\}$, $x_i = 1$ show deployment hardware at node i;

y_{ip}: $y_{ip} = \{0,1\}$, $y_i = 1$ show deployment business p at node i;

z_{ip}: $z_{ip} = \{0,1\}$, $z_{ip} = 1$ show transmission business p at node i;

$y_{ip} \le z_j$, $j = 0,1,\cdots N$, $p = 0,1,\cdots P$;

N_j: $N_j = (n_{j1}, n_{j2}, \cdots n_{jp})$ show the type set of business required at node i.

- State:

n_{jp}: $n_{jp} = \{\mathrm{N},\mathrm{M},\mathrm{T}\}$, in which, $n_{jp} = \mathrm{N}$ indicates that node j needs no deployment business p and does not have to transfer, $n_{jp} = \mathrm{M}$ indicates that node j needs deployment business p and hardware devices, $n_{jp} = \mathrm{T}$ indicates that node j needs to transmit from other nodes to use business p. And the initial state is set as at least a device, a business and a transmission in the network.

Therefore, if node i needs business set N_j, the corresponding deployment cost, that is the goal, can be expressed as:

$$R_j(N_j) = S_j + \sum_{j=1}^{P} P_j + \sum_{j=1}^{n} C_j \tag{4.3}$$

Transformation law of dynamic programming algorithm, namely recursive formula is:

$$R_j(N_j) = R_j(N_j) + \sum_{n} \min R_n(N_n) \tag{4.4}$$

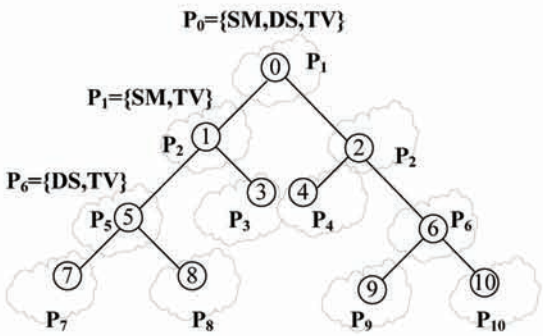


Fig. 5. Network Profession Deployment.

In which, $n \in CHILD(j)$, $N_j \in EXIST_j$, $j = n, n-1, \ldots, 0$. Because of the deployment situation of actual business, not all N_j are present. The set presenting N_j is expressed as $EXIST_m(N_j)$, N_m is a better result, $N_m \in EXIST_m(N_j)$, $CHILD(j)$ shows the set of all sub-nodes under node j, as shown in **Fig.5** $CHILD(0) = \{1,2\}$.

- Decision:

It is a choice from the state to a state of the next step after the determination of the state of the step, that is, N_m of the step.

Taken **Fig.5** as an example, based on dynamic programming ideas, the optimal deployment of three businesses $P_j = \{\mathrm{SM,DS,TV}\}$ shown in the form of network topology in the figure is divided into four steps. In which, SM, namely Streaming Media, stands for streaming media transmission business, DS, namely Data Share, stands for data transmission business, and TV stands for traditional TV business. Initial state is P0 = {SM, DS, TV}, business demand is P7 = P8 = {SM, TV}, P9 = P4 = {SM}, P10 = P4 = {DS}, P5 = P6 = {DS, TV}. To facilitate the description, $n_{jp} = \{\mathrm{N,M,T}\}$ is noted as $n_{jp} = \{0,1,2\}$, $P_j = \{\mathrm{SM,DS,TV}\}$ as $P_j = \{0,1,2\}$, then N_j corresponding vector value can be used to express the state variable n_{jp} and corresponding location to represent desired business type. Meanwhile, all hardware deployment costs are defined as $S_i = S = 6$, $P_{ip} = P = 2$, the transmission bandwidth costs are $C_{0\,1} = 6$, $C_{0\,2} = 4$, $C_{1\,3} = 3$, $C_{1\,5} = 1$, $C_{2\,4} = 1$, $C_{2\,6} = 4$ and $C_{5\,7} = C_{5\,8} = C_{6\,9} = C_{6\,10} = 3$, respectively.

As shown below **Table3**, Costs of all valid states of node set {7, 8, 9, 10} at the end are calculated. The final deployment decisions are identified by $*$ in the table and, as shown in **Table3**, that is $N_0 = \{\mathrm{M,M,M}\}$, $N_1 = \{\mathrm{M,N,M}\}$, $N_2 = \{\mathrm{N,T,N}\}$, $N_3 = \{\mathrm{T,N,N}\}$, $N_4 = \{\mathrm{N,T,N}\}$, $N_5 = \{\mathrm{T,N,T}\}$, $N_6 = \{\mathrm{N,M,M}\}$, $N_7 = N_8 = \{\mathrm{T,N,T}\}$, $N_{10} = \{\mathrm{N,T,N}\}$. That is, equipments and three businesses $P_0 = \{\mathrm{SM,DS,TV}\}$ are deployed at the root node 0, equipments and two types of business $P_1 = \{\mathrm{SM,TV}\}$ are deployed at node 1, equipments and two types of business $P_6 = \{\mathrm{DS,TV}\}$ are deployed at node 6, and the transmission of the remaining nodes minimize the network deployment overhead using the decisions of business required.

Table 3. Node calculation and network deployment decision.

State N_j	State sets $EXIST_j$	Results $R_j(N_j)$	Computational method	Decision
(0,1,0)	$EXIST_{10}$	8	$S+P$	
(0,2,0)		3	C_{10}	$*$
(1,0,0)	$EXIST_9$	8	$S+P$	
(2,0,0)		3	C_9	$*$
(1,0,1)	$EXIST_8$	10	$S+P+P$	
(1,0,2)		11	$S+P+C_8$	
(2,0,1)		11	$S+P+C_8$	
(2,0,2)		6	C_8+C_8	$*$
(1,0,1)	$EXIST_7$	10	$S+P+P$	
(1,0,2)		11	$S+P+C_7$	
(2,0,1)		11	$S+P+C_7$	
(2,0,2)		6	C_7+C_7	$*$

6. Conclusion

In this paper, the traffic scale estimation method in QoS guarantee bandwidth planning, followed by the optimal network deployment strategies provided based on dynamic programming ideas verify the effectiveness of link bandwidth planning. And from business and network level, based on meeting the point of view of QoS and the access to traffic scale, by trading space complexity for computation complexity, an optimal network deployment method which facilitates the project implementation is provided.

References

[1] Agner K. Erlang: *The Theory of Probabilities and Telephone Conversations*, Nyt Tidsskrift for Matematik, Vol. 20, No. B. (1909), pp. 33-39.

[2] Addie R G, Neame T D, Zukerman M: *Performance analysis of a Poisson–Pareto queue over the full range of system parameters*. Computer Networks, 2009, 53(7):1099-1134.

[3] R G Addie.T D Neame.M Zukerman: Performance Evaluation of a Queue Fed by a Poisson Pareto Burst Process, Computer Networks 2002, 40(3): 377-397.

[4] ZhaoXin: *An Analysis of Aggregation of Streaming Media Traffic* in MINES2009, [EI, ISTP included]

[5] B. K. Ryu, A. Elwalid, The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities, ACM SIGCOMM Computer Communication Review 26 (4) (1996) 3–14.

[6] Zhao Xin. *Bandwidth planning study of limited-scale integrated traffic*, Journal of Electronics and Information.

[7] Richard Ernest Bellman, Dynamic Programming, N Y: Dover Publications, 2003.

[8] Xie JX, Xing WX. Network optimization, Beijing: Tsinghua University Press, 2010.

[9] Koonlachat Meesublak. *Network Design under Demand Uncertainty*. Proc. of the Asia-Pacific Advanced Network Meeting, 2008, 11-18.

An Inverted Index Method for Mass Spectra K-Nearest Neighbor Queries

Houjun Tang, Xi Liu, HonglongXu, Kezhong Lu, Gang Liu,
YuhongFeng, Hong Zhou, Rui Mao[1]
National High Performance Computing Center at Shenzhen
College of Computer Science and Software Engineering
Shenzhen University
3688 Nanhai Road, Shenzhen, 518060, China
houj.tang@gmail.com, xii.liu@hotmail.com, longer597@163.com,
{kzlu, gliu,yfeng, hzhou, mao}@szu.edu.cn

Keywords:K-nearest neighbor search, metric-space indexing, mass spectra, sparse matrix, inverted index.

Abstract. Finding k-nearest neighbors (k-nn) in metric-space is frequently used in modern biological applications due to its general applicability. Processing such queries with general purpose methods usually requires more time and space than domain-specific methods. This paper presents an inverted index method which exploits the sparsity of mass spectra binary format data and compares it with an existing metric-space method. This metric-space method acts as a coarse filter and can be followed by any fine ranking scheme. In experiments, we find that the new method outperforms the metric-space method in both query speed and index size.

Introduction

Tandem mass spectrometry, also known as MS/MS or MS^2, is used to produce structural information about a compound. It has been used as a common technique in proteins and peptide sequences identification in complicated samples. A mass spectrum obtained by an experiment contains a list of peaks corresponding to the peptide fragment ions which are pairs of real numbers m/z ratios and their intensity of occurrence, where m denotes mass and z charge [10]. By labeling each spectrum with its correct amino-acid sequence, we can identify a peptide's presence in the protein sample.

The spectrum identification step can be easily modeled as a similarity search problem such as the k-nearest neighbor (k-nn) search. In k-nn search, k objects which are most similar to the given object are retrieved from a large database which is measured by a distance function. In our case, experimentally generated spectra are compared to a database of theoretical spectra.

In the last twenty years, protein databases have been growing exponentially [7], which leads to a great increase of computation when identifying a biological sequence from theoretical databases. Performing k-nn search on experimental spectra against theoretical ones costsmore time than ever and linear scan is no longer acceptable. As a result, various protein identification methods supporting rapid similarity search have been developed, such as TurboSEQUEST [19], MASCOT [11], ProFound [21], and clustering method [4]. The similarity measures include cosine distances based on shared peak count [12,15] and the Hausdorff distance [10].

Metric-space indexing, also known as distance-based indexing [2,6,16,20] usually focus on its general applicability and take little use of data domain information. All information the indexing need is the metric function to compute the distance between objects. The data set is clustered and a data structure complied during an off-line process while an on-line search makes use of the triangle

[1]The correspondence author

inequality to eliminate clusters of data and return possible results.

In MoBIoS project [8], Ramakrishnan et al. proposed a metric-space technique called MSFound and use it for similarity search of mass spectra [15]. In their method, a cosine distance-based semi-metric distance is introduced and the MVP Tree [1] is used as data structure to perform range and k-nn queries. The method produce an initial candidate set which can be ranked by any fine ranking scheme afterward [15].

Another metric-space indexing method proposed by Dutta and Chen [3] named Locality Sensitive Hashing (LSH). It groups data points in the metric-space into 'buckets' based on the specific distance corresponding to the data. Points that are close to each other under the chosen metric are mapped to the same bucket with high probability. The LSH algorithm [5] is proved to be efficient for nearest neighbor queries and returns elegant probabilistic bounds on the error of the results [18].

As is often the case that a domain specific method performs better than a general one, it is of interest to design a domain-specific method which takes use of the feature of data and compare it to the general method. We have already done similar work in range query processing [14]. In this paper, we present an inverted index method that exploits the sparsity of high dimensional binary vector to investigate its performance in k-nn queries. Since the sparsity of the vector can be as high as 99.9%, our inverted index method can make full use of it and reduce the number of distance computations. We demonstrate the high search efficiency and less requirement of index storage of our method on two mass spectra datasets.

Mass Spectra and Tandem Cosine Distance

Ramakrishnan et al. [15] have already clarified the representation on mass spectra and the tandem cosine distance. For the completeness of this paper, we make a brief introduction here.

The mass spectra data generated from experiment consist of two parts: a precursor mass (M) and a list of m/z peaks, P = {pi}. They both are real numbers. For tandem cosine distance, the list is stored in a high-dimensional sparse boolean vector S. For example, given a peak range [0, 3000] Da and a resolution of 0.1 Da, the mass spectra is represented by a 30,000-dimensional boolean vector having 1s where the m/z has a nonzero value in the mass spectra. Each mass spectra data record can be represented as a pair {M, S}.

Given two mass spectra data records A and B, where A = {MA, SA} and B = {MB, SB}, and a peak mass tolerance τms ≥≥ Mres, a shared peak count within a peak tolerance window SPCτ(A,B) is defined as [15]:

$$\mathrm{SPC}_\tau(\mathrm{A},\mathrm{B}) = \sum_i match(a_i, b_j);\ j \in \left[i - \frac{\tau_{ms}}{M_{res}}, i + \frac{\tau_{ms}}{M_{res}}\right]. \quad (1)$$

$$\mathrm{match}(\mathrm{a}_i, \mathrm{b}_j) = \begin{cases} 1 & \text{if } \mathrm{a}_i = \mathrm{b}_j = 1\ match(\mathrm{a}_m, \mathrm{b}_j) = 0, m \in [1, i) \\ 0 & \text{else} \end{cases}$$

And then the fuzzy cosine distance D_{ms}(A,B) can be defined as [15] :

$$\mathrm{D}_{ms}(\mathrm{A},\mathrm{B}) = arccos\left(\frac{\mathrm{SPC}_\tau(\mathrm{A},\mathrm{B})}{\|S_A\|\|S_B\|}\right). \quad (2)$$

The precursor mass distance D_{pm}(A,B) is also introduced to compute the absolute distance in precursor mass within a tolerance window τ_{pm}[15] :

$$\mathrm{D}_{\mathrm{pm}}(\mathrm{A},\mathrm{B}) = \begin{cases} 0, \text{if } |M_A - M_B| \le \tau_{\mathrm{pm}} \\ |M_A - M_B|, \quad \text{else} \end{cases}. \quad (3)$$

Peptides are unlikely to be similar if there precursor masses differ from each other greatly. Therefore, a linear combination of precursor mass distance and fuzzy cosine distance is proposed [15] :

$$\mathrm{D}_{\mathrm{tcd}}(\mathrm{A},\mathrm{B}) = \mathrm{C}_1\mathrm{D}_{ms}(\mathrm{A},\mathrm{B}) + \mathrm{C}_2\mathrm{D}_{pm}(\mathrm{A},\mathrm{B}). \quad (4)$$

D_{pm} is a semi-pseudometric due to the precursor mass tolerance τ_{pm}, and so is the tandem cosine distance D_{tcd}[15]. The value of the constants C_1 and C_2 are both 1 and we set $\tau_{pm} = 2$ and t =1 in our case.

An Inverted Index Method

To take use of the sparseness of the mass spectra binary peak vector, we use the inverted index method to store a compressed vector representing the list of nonzero peaks. To make it simple for later processing, the vector is in ascending order.

To compute the shared peak count with tolerance, a recursive algorithm can be written. The algorithm counts the match between two vectors from the beginning, if their difference is within the tolerance, remove both entries of the vectors and return their subsequence with count value plus one. If not, remove the entry of vector with smaller value and return the two vectors. Then algorithm recursively repeat the above process until one of the vectors has no element left.

Bulkloading the index. Comparing with the computation of SPC_t, computing the precursor mass distance takes much less time. And since most of the precursor mass distance is much greater than the fuzzy cosine distance which is bounded by [0, π/2], it is more efficient to compute the precursor mass distance and prune data before the computationally expensive fuzzy cosine distance computation. To speed up this process, the data is first sorted by precursor mass.

The inverted index D is defined as follow:

$$D = \{D_j \mid D_j = \{i \mid S_i[j] = 1, i = 1 \dots m\}, j = 1 \dots n\}$$

Compressed vectors		Inverted index
S_1 = [2,4,5,7]		L_1 = [5,7]
S_2 = [3,6]		L_2 = [1,6]
S_3 = [3]	⇨	L_3 = [2,3,5]
S_4 = [4]		L_4 = [1,4,5,6,7]
S_5 = [1,3,4,6]		L_5 = [1]
S_6 = [2,4]		L_6 = [2,5,7]
S_7 = [1,4,6,7]		L_7 = [1,7]

Fig.1 An example of inverted index

K-nearest neighbor query processing. We perform k-nn query within a radius r, *kNN(q,k,r)*, to increase the efficiency of the k-nn process. The value of radius is dynamically decreased during the k-nn process. Two pruning rules are used to reduce the computations of SPC_t and they are given as theorems as follow.

Theorem 1: A is not a result of k-nn query if $M_A > M_q + \max(\tau_{pm}, r/C_2)$ or $M_A < M_q - \max(\tau_{pm}, r/C_2)$.

Definition: Given a mass spectra query q, the set of lists in the inverted index that are within tolerance t to one of q's non-zero entry L(q). [14]

$$L(q) = \{Li \mid Li \in L, \exists j, \mid i - j \mid \leq t, S_q[j] = 1\}$$

We define GSPCt(q, A), the gross shared peak count with tolerance t of q and a mass spectra A in the database, which is the number of appearances of A in lists of q. Using the inverted index, GSPCt can be easily computed and it is used in Theorem 2 for pruning. [14]

Theorem 2: A is not a result of k-nn query kNN(q,k,r) if

$$\arccos\left\{\frac{GSPC_t(q,A)}{\|S_A\|\|S_q\|}\right\} + D_{pm}(q,A) > r$$

Theorem 1 is proved by the fact that $D_{ms} \geq 0$ and Theorem 2 is proved using the fact that $SPC_t(q, A) \leq GSPC_t(q, A)$.

In this paper we present and test two kinds of k-nn methods, one is ordinary k-nn method and the

other is a radius bounded method. The second method uses a radius that is prior obtained.

The complete k-nn algorithm is given as the following steps:

1. Preprocess the data. Given a database A, bulkload the index and sort it by precursor mass. The data is stored in an array list. Set initial radius r according to different circumstances.
2. Find the data within tolerance window τ_{pm} to the query's precursor mass. As the data is sorted by precursor mass, a binary search can be used to quickly find the right data.
3. Compute tandem cosine distance D_{tcd} of k elements near element found in step 2 and put these elements into a priority queue. Then set the radius r to the maximum of the k distances.
4. Prune data using Theorem 1. Put data satisfying Theorem 1(1) into a candidate set together with their precursor mass distance to q.
5. Compute $GSPC_t$ for any data A in the candidate set satisfies $GSPC_t(q,A) > 0$ using inverted index.
6. Prune data in the candidate set using Theorem 2.
7. Compute the tandem cosine distance of all the data left in candidate set and insert them into the priority queue.

In step 1, we set $r = \infty$ for ordinary k-nn search kNN(q,k), and for radius bounded k-nn search, r is set according to the specific data type. In step 4, binary search tree or B-tree can be used in different cases. In step 5, the head of each list can be reached in constant time due to the linear storage of the inverted index. We use a hash map to count the gross shared peak count. We customized the priority queue to make it only stores k elements unless there are multiple elements with the same D_{tcd} at the end of the queue. So the results can be easily reached after the above steps.

Cost analysis. Assume p is the probability that an entry of a binary peak vector is "1". And the 1s in the vector is randomly distributed. Suppose N is the dimension of the binary peak vectors and M is the database size. The expected number of non-zero entries of each binary peak vector, or the expected length of each compressed peak vector, is Np. The expected size of each list of the inverted index is Mp, and expected total number of ids stored in the inverted index is MNp. We have already proved that the overall bulkload time is O(MlgM) and the space needed for precursor mass is O(M) and O(MNp) for the inverted index.[14]

In step 2, finding the right data takes O(NlgN) using binary search. In step 3, computing tandem cosine distance of k elements takes O(1) time. In step 4, determining the range of candidate and result in the sorted data array takes O(lgM) time. Putting candidates into candidate set and computing their precursor mass distances takes O(#candidate) time. In step 5 for t = 0, the expected number (can be with duplicates) of ids in L(q) is $O(MNp^2)$. Step 6 takes O(#candidate) time. In step 7, the shared peak count algorithm running time is O(Np), and the total time of step 7 is O(Np#computed).

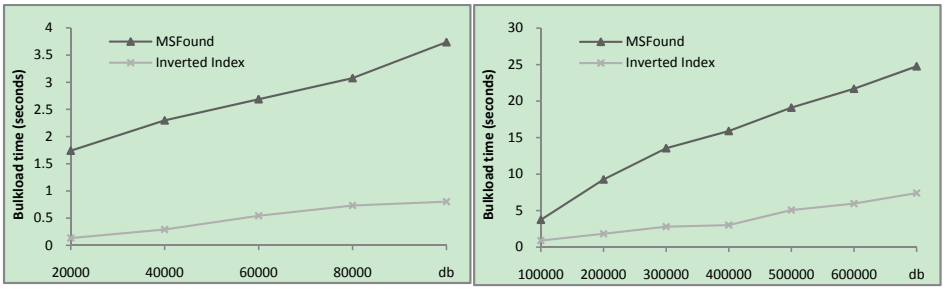
The overall cost of above is $O(MNp^2)$ for t = 0 under the circumstance that #candidate < M, #computed is empirically verified small and ignoring the I/O cost. For sparse vectors like the case we have, the effect of N and M can be greatly reduced.

Empirical Results

In our study, we use a query set which contains 49 query objects and run the search in two datasets. Dataset I has 92768 MS/MS spectra from protein sequences of a seven protein mixture from the Sashimi proteomics repository. The dataset is concatenated with a control database of spectra from the Escherichia coli K12 (E. coli) genome. Dataset II combines Database I with a larger control database of theoretical mass spectra from the human genome and its size is 654276. The query and test set databases in their original format are available from the Open Proteomics Database [13] and the Sashimi mass spectra repository [17].

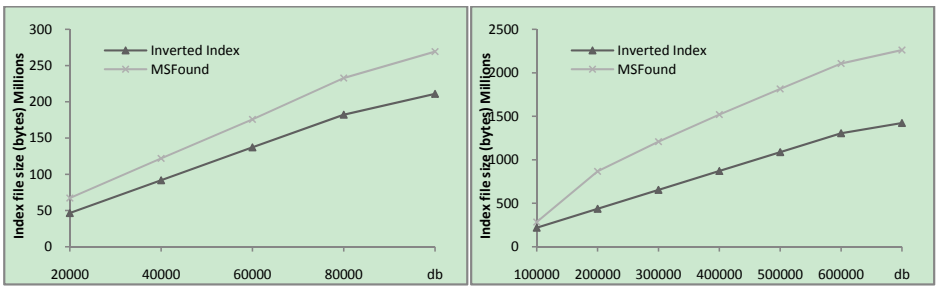
The mass range in both sets is [100, 5000] Da with the resolution of 0.2 Da. So the dimension of the peak vector, N, is about 25000. Analyzing the sparsity of the binary vectors in the database which is above 99% we confirmed the fact that the database is of high sparsity. All the algorithms were implemented in JAVA on top of Sun JVM 1.6. All the experiments were carried out on a Lenovo computer running Window 7 with 2G memory and a dual core Intel E5800 CPU at 3.2 GHz. The code and data in binary format are available at the MoBIoS repository [9].

We compare the bulkload time between the inverted index method and the MSFound method as is illustrated in Fig.2. In Dataset I, the inverted index method has about 10X speedups than the MSFound and in Dataset II it is about 5X speedups.



(a) Bulkload time of Dataset I (b) Bulkload time of Dataset II

Fig.2 Bulkload time of two datasets



(a) Index files sizes of Dataset I (b) Index files sizes of Dataset II

Fig.3 Index files sizes of two datasets

We estimate the space consumption by the size of the index file. Both datasets with various database sizes are illustrated in Fig.3. Our inverted index method needs 40% less space than MSFound. The linear trend between the index file size and database size is consistent with the analysis inCost analysissection.

K-nn query performance. Both ordinary and radius bounded k-nn (r-kNN) queries are tested and compared with the MSFound method. We set original radius r = ∞ for ordinary k-nn search and r = 1.48 for radius bounded k-nn search. This value came from Ramakrishnan et al.'s paper that in the given databases, r = 1.48 is the smallest radius which can return all positive results [15].

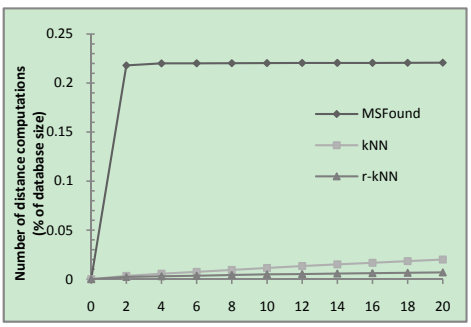


Fig.4 Number of distance computation percentages with different k value for Dataset II

We run a test to see the how much the inverted index method can reduce the number of distance computations with different k values over the MSFound method. In Table1 we can see that the two pruning rules greatly reduce the number of distance computations, it has a speedup of 10 over

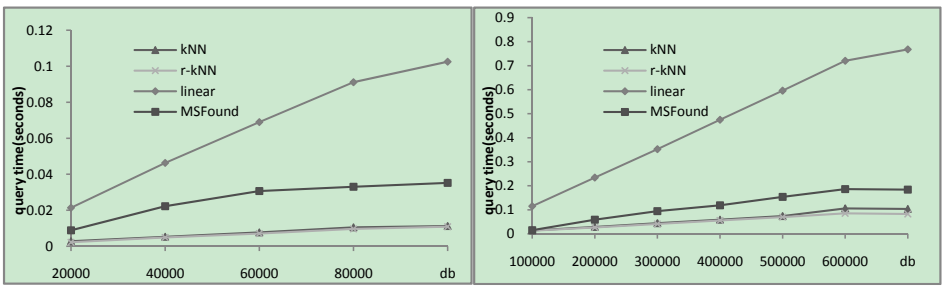
MSFound. In Fig.4 we can have a clearer view on how much the inverted index method reduces the number of distance computations. Also the radius bounded method need less computations than ordinary method.

Notice that we only count the number of tandem cosine distance computations, however, in our method, there are the computations like gross shared peak count which are not counted in the above list for they need relatively small amount of time than tandem cosine distance. The running time of different methods is given as follow for the entire performance evaluation.

Table 1. Number of distance computations with different k value for Dataset II

k	MSFound	kNN	r-kNN
2	1426	22	15
4	1440	36	19
6	1441	49	23
8	1441	62	28
10	1442	75	32
12	1443	88	35
14	1443	98	37
16	1443	109	40
18	1444	120	42
20	1445	131	45

In Fig.5 we can see that both of the k-nn methods outperform the MSFound method and the three is faster than linear scan. The k-nn search with radius bound is about 25% faster than the method without a bound when the size of the database is larger than 600000. For smaller databases their performance is almost the same. The speedups of inverted index over MSFound are about 4 for Dataset I and 2 for Dataset II. For both datasets, the inverted index method's speedups over MSFound decrease with the increase of database size.



(a) k-nn query time of Dataset I (b) k-nn query time of Dataset II

Fig.5 k-nn query running time with k = 16

Conclusions and Future Work

We have shown that the inverted index method is efficient in reducing the number of distance computations. Taking use of the high sparsity of binary format mass spectra, the domain specific indexing method, inverted index, k-nn queries can be processed much faster. By comparing it to MSFound and linear scan, we see that inverted index method possesses the fastest query time. And by knowing the radius first we can see an even better performance to set a radius bound for the k-nn search.

Our plans for further investigations include issues related to testing larger databases. We also plan to do more works in other domains and compare our schemes to others.

Acknowledgments

This research was supported by the following grants: NSF-China: 61033009, 61003272, 61170076,61103055; China NSF-GD grant: 10351806001000000; a grant from the Computer Architecture Key Lab of Chinese Academy of Sciences: “Transportation and optimization of Hadoop and GeDBIT on Loongson based platforms”; Shenzhen Foundational Research Project: JC201005280408A, JC200903120046A; a grant from the Shenzhen-Hong Kong Innovation Circle Project: ZYB200907060012A; a SZU Research Course Project: 0000132373.

References

[1] TolgaBozkaya , M.Ozsoyoglu, Indexing large metric spaces for similarity search queries, ACM Transactions on Database Systems (TODS), v.24 n.3, p.361-404, Sept. 1999

[2] E.Chávez , G.Navarro , R.Baeza-Yates , José Luis Marroquín, Searching in metric spaces, ACM Computing Surveys (CSUR), v.33 n.3, p.273-321, September 2001

[3] D.Dutta, T.Chen, Speeding up tandem mass spectrometry database search: metric embeddings and fast near neighbor search, Bioinformatics, v.23 n.5, p.612-618, Feb.2007

[4] Ari M. Frank, NunoBandeira, ZhouxinShen, Stephen Tanner, Steven P. Briggs, Richard D. Smith, and Pavel A. Pevzner. Clustering Millions of Tandem Mass Spectra. J. Proteome Res. 2008 January; 7(1): 113--122.

[5] Aristides Gionis , PiotrIndyk , Rajeev Motwani, Similarity Search in High Dimensions via Hashing, Proceedings of the 25th International Conference on Very Large Data Bases, p.518-529, September 07-10, 1999

[6] Gisli R. Hjaltason , HananSamet, Index-driven similarity search in metric spaces (Survey Article), ACM Transactions on Database Systems (TODS), v.28 n.4, p.517-580, Dec.2003

[7] D. Hoksza and T. Skopal. Index-based approach to similarity search in protein and nucleotide databases. CEUR Proc. Dateso 2007, vol. 235, pp. 67--80. 2007.

[8] Daniel Miranker , WeijiaXu , Rui Mao, MoBIoS: a metric-space DBMS to support biological discovery, Proceedings of the 15th International Conference on Scientific and Statistical Database Management, p.241-244, July 09-11, 2003, Cambridge, MA

[9] The MoBIoS repository: http://aug.csres.utexas.edu/sisap2010_ms

[10] J. Novák, D. Hoksza. Parametrised Hausdorff Distance as a Non-Metric Similarity Model for Tandem Mass Spectrometry. In the Proceedings of the Dateso 2010 Annual International Workshop on DAtabases, TExts, Specifications and Objects. Stedronin-Plazy, Czech Republic, April 21, 2010.

[11] Perkins, D. et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20, 3551--3567, 1999.

[12] Pevzner, P. et al. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. Genome Res., 11, 290--299, 2001.

[13] J. Prince, M. Carlson, R. Wang, P. Lu, and E. Marcotte. The need for a public proteomics repository. Nature Biotechnology, 22(4):471--472, 2004.

[14] Rui Mao, Smriti R. Ramakrishnan, Glen Nuckolls and Daniel P. Miranker, "Case Study: An Inverted Index for Mass Spectra Similarity Query and Comparison with a Metric-space Method", SISAP2010,pages 93-99, 2010.

[15] Smriti R. Ramakrishnan , Rui Mao , Aleksey A. Nakorchevskiy , John T. Prince , Willard S. Willard , WeijiaXu , Edward M. Marcotte , Daniel P. Miranker, A fast coarse filtering method for peptide identification by mass spectrometry, Bioinformatics, v.22 n.12, p.1524-1531, June 2006

[16] HananSamet, Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling), Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005

[17] The Sashimi mass spectra repository: http://sashimi.sourceforge.net.

[18] Gregory Shakhnarovich , Trevor Darrell , PiotrIndyk, Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing), The MIT Press, 2006

[19] Yates III, J. et al. Method to correlate tandem mass spectral data of modified peptides to amino acid sequences in the protein database. Anal. Chem., 67, 1426--1436, 1995.

[20] PavelZezula , Giuseppe Amato , VlastislavDohnal , Michal Batko, Similarity Search: The Metric Space Approach (Advances in Database Systems), Springer-Verlag New York, Inc., Secaucus, NJ, 2005

[21] Zhang, W. and Chait, B. ProFound---an expert system for protein identification using mass spectrometric peptide mapping information. Anal. Chem., 72, 2482--2489, 2000.

Fuzzy Correlation with the Issues of Study in Domestic Campus or Intent to Study Abroad

Dian-Fu Chang[1, a] and Wen-Ching Chou[2, b]

[1] Graduate Institute of Educational Policy and Leadership, Tamkang University, Taiwan
[2] Dept. of Education Policy and Administration, National Chi Nan University, Taiwan
[a]140626@mail.tku.edu.tw, [b]julinchou@yahoo.com.tw

Keywords: fuzzy measurement, fuzzy correlation, fuzzy statistics, study abroad, higher education

Abstract. This article applied the fuzzy measurement to collect the college students' perceptions of their campus life and their intention to stay at domestic campus or study abroad. This study used the self-designed fuzzy questionnaire to collect the data from 289 college students in Taiwan. The results revealed that fuzzy statistics can be applied to tackle ambiguity questions in different fields. The fuzzy correlation provides more detail information to interpret what students have done and what they intent to do. By way of fuzzy computing, the study can decipher the student engagement in different campus.

1. Introduction

Many articles displayed that the fuzzy statistics has become a useful tool for measuring ambiguous concepts in science and social science [4, 8]. Their argument is why the traditional numerical model cannot explain complex and ambiguous human and social phenomena properly? They encouraged to use the soft computing to solve the complex and ambiguity phenomena [5]. However, the ambiguous data are consistent with human logic, the study needs a powerful way to deal with such kind of data during computing process. The concept of fuzzy set proposed by Zadeh and applied to fuzzy measurement to cope with the dynamic environment [9]. Recently, the concept of fuzzy set has given a more reasonable description for different purposes of data transform [8, 12].

Studying at home or studying abroad needs to be better integrated into existing institutional research efforts. Especially, how the research can help better define what we most wanted students to learn from campus experiences? This study designed a fuzzy questionnaire to collect the college students' opinions of study at home or study abroad in terms of their perceptions of time spent in campus and intention of study abroad. Given the purposes, this study answered the following questions:

(1) Which activity students spent more time in campus?
(2) Did they have any difference among time spent in gender and different colleges?
(3) What kind of correlation have shown among the activities they engaged in?
(4) Can fuzzy statistics be used to interpret the data properly?

2. Literature Review

When undergraduates first attend university, their expectation about college life is based on popular culture or the media. Even the most cursory acquaintance with American campus movies will be sufficient to remind one that going to university is about having fun, playing sports, falling in love, making new friends and generally 'growing up' [7]. Such depictions about undergraduates is not only limited to the U.S. or British context; Taiwan has a similar portraits of university life, *"All you can play four years!"* This popular perception has affected engagement among undergraduates [6].

Clark and Trow proposed a typology of student orientations model [2]. In their model, undergraduates could be categorized as four types: academic, collegiate, vocational, and non-conformist. Typically, academic-oriented students attend college to pursue academic knowledge and ideas [7]; they tend to spend more time on academic studying and less on university activities or job preparation courses. Otherwise, some of students seek interaction with faculty outside of the classroom [1]. Job-oriented students may regard academic as less important as job skills. They resist intellectual demands as long as they can pass the courses [2]. These job-oriented students consider the university as an instrument for preparing them for a future career.

Some of researchers argued that many students are choosing to stay home while also acquiring an international education [11]. These new models of education include distance learning, joint and dual degrees, branch campuses, and sandwich programs involving short-term study abroad. Gray called these types as "nontraditional academic arrangements"[3]. However, most countries now view international academic mobility and educational exchanges as critical components for sharing knowledge, building intellectual capital, and remaining competitive in a globalizing world. What has changed? Does increasing numbers of students have come to realize that study abroad will enhance their career options as they enter a marketplace?

3. Methods

This study applied fuzzy statistics to demo how to transform fuzzy means, defuzzification for a fuzzy number on *R*, fuzzy correlation with related campus activities, and fuzzy correlation with intention of study abroad. This study assumed that the time spent on campus are different among gender and colleges. If they show significant correlation, the time spent may affect on the intention of study abroad. The study framework shows on Fig. 1.

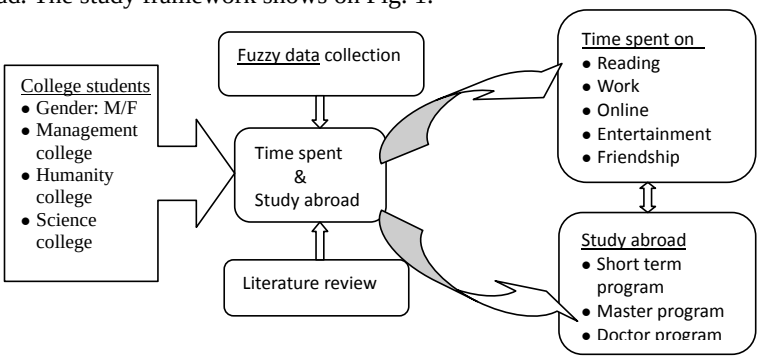


Fig. 1 A framework of analysis

The fuzzy questionnaire includes the following items: gender, college major, time spent in different activities, and intention to study abroad. The variables related to campus life in the questionnaire are including: (1) reading refers to the regular reading, reading extra books, and magazines; (2) work in terms of campus or off campus part-time job; (3) online refers to using Internet to find related information to write report or to do homework; (4) entertainment includes online games, on social networking site to talk, talking on cell phones, and travel, etc.; (5) friendship means the relationship with others, including make friends. Students' intention of study abroad are including select short term stay program, master program, or doctor program. The examples of fuzzy questionnaire are as follows:

***Direction**: The following questions related to your perceptions of time spent on the campus activities and your intention to study abroad. We need your opinions about actual time spent on selected campus activities and your intention to go abroad to study in different programs.*

*1. If your daily spent time on **reading** about 3 to 6 hours, you need circle the actual time spent with the number 3 and 6. Please cycle your perceptions in following campus activities:*

***Reading**, it refers to the regular reading, reading extra books, and magazines.*

Actual time spent

0	1	2	③	4	5	⑥	7	8	9

2. If you plan to study abroad, which program will you select? When you circle 2 to 5, it means your intention to short term program is from 2 to 5. Please start cycle your intention in the following study abroad programs:

Intention to study short term program

0	1	②	3	4	⑤	6	7	8	9

The samples were collected from a national and a private university in central and northern of Taiwan, respectively. The total valid questionnaires are 289, males are 138 and females are 151 in three different colleges, see Table 1.

Table l. The samples distribution of the target universities

Gender	Colleges			Total
	Management	Humanities	Science	
Males	55	8	75	138
Females	73	27	51	151
Total	128	35	126	289

Definition 1. Let U be the universal set and $\{Fx_i = [a_i, b_i], a_i, b_i \in R, i = 1,...,n\}$ be a sequence of random fuzzy sample on U. The fuzzy sample mean value is then defined as [5]

$$F\bar{x} = [\frac{1}{n}\sum_{i=1}^{n} a_i, \frac{1}{n}\sum_{i=1}^{n} b_i] \quad (1)$$

Definition 2. Let χ=[a, b] (a≠b) (be an interval fuzzy number on U). The defuzzification number $R_\chi$of [a, b] is then defined as [5]

$$R_x = \frac{a+b}{2} + (1 - \frac{\ln(1+|b-a|)}{|b-a|}) \quad (2)$$

Defuzzification for a fuzzy number on *R* is derived from interval fuzzy numbers. This study calculated *Rs* to rank different activities related to study abroad. The concept of interval fuzzy pattern can be defined as a well-distributed membership function with fuzzy numbers. The symbol of "[ ]" means a closed interval. If $a, b \in R$ and $a < b$, then $[a, b]$ is interval fuzzy number. We named "a" is the lower bound of $[a, b]$ and "b" as the upper bound of $[a, b]$; If $a = b$, then $[a, b] = [a, a] = [b, b] = a = b$, it is a real number a (or b).

Similarly, a real number k can be defined as $[k, k]$. If $[a, b]$ is an interval fuzzy set, we define $c_0 = \frac{a+b}{2}$, $s_0 = \frac{b-a}{2}$, it represents its "center" and "radius" respectively. We can also express an interval fuzzy number as the following way:
$[c_o; s_o] \Rightarrow [c_o + s_o, c_o - s_o] = [a, b]$ And $\ell = b - a$ is the length of the interval.
For example, we consider (x_i, y_i) as the first *i* sample value, $i = 1,2,....n$; x_i, y_i are interval fuzzy numbers; $\bar{x}, \bar{y}$ represent the sample mean respectively.

This study used the Eq. 3 to get and also to adjust a more reasonable fuzzy correlation coefficient. The study considered the *e* base of the natural logarithm function *ln* to transform the fuzzy data. Let lx_i be the length of continuous interval sample x_i, ly_i be the length of the sample interval y_i, then the corrected length of correlation coefficient is

$$\delta = 1 - \frac{\ln(1+|r_l|)}{|r_l|}\ ;\ where\ \ r_L = \frac{\sum_{i=1}^{n}(l_{x_i}-\bar{l}_x)(l_{y_i}-\bar{l}_y)}{\sqrt{\sum_{i=1}^{n}(l_{x_i}-\bar{l}_x)^2}\sqrt{\sum_{i=1}^{n}(l_{y_i}-\bar{l}_y)^2}} \quad (3)$$

Since $0 < r_l < 1$, *so the range of* δ *is* $0 < \delta < 0.3069$.

Definition 3. Let C_{xi}, C_{yi} are samples from the interval fuzzy matrix central point, l_{xi}, l_{yi} for the interval length. The r is the center of the correlation coefficient, σ is the fitter to be used to correct the length of the correlation coefficient. The relevant interval is defined as follows:

(i) $r \geq 0,\ r_l \geq 0$, (r, min(1, $r + \delta$))

(ii) $r \geq 0,\ r_l < 0$, ($r - \delta$, r)

(iii) $r < 0,\ r_l \geq 0$, (r, $r + \delta$)

(iv) $r < 0,\ r_l < 0$, (max(-1, $r - \delta$), r)

How to evaluate the significant level of fuzzy correlation coefficient? This study proposed approximate rules to test the meanings of fuzzy correlation coefficients [4].

Positive correlation:

r> .65 for the high correlation,

.35 <r <.65 for the moderate correlation,

r <.35 for the low correlation;

Negative correlation:

r> -. 35 for low negative correlation,

-. 65 <r <-. 35 for moderate negative correlation,

r <-. 65 for high negative correlation.

4. Results

This study transformed the fuzzy data with campus activities and the intent to study abroad by gender and college differences. In this section, the fuzzy statistics were used to domo how they can decode the ambiguous issues properly.

Fuzzy interval data transform by gender difference

This study found the gender had shown different styles of time spent on campus activities. The college students spent a lot of time on part-time job (*R*=3.99, *R*=4.57). Males had less engaged in reading (*R*=3.92), however, female less spent time on online activities (*R*=3.00). The ranking of time spent on activities showed on Table 2. Typically, males spent more time on campus activities than do females.

Table 2. Gender difference of time spent compared by fuzzy mean and *R*

	Males			Females		
	Fuzzy mean	*R*	Ranking	Fuzzy mean	*R*	Ranking
Reading	[2.15,4.98]	3.92	**5**	[1.68,4.17]	3.32	3
Work	[2.62,5.86]	4.57	**1**	[2.19,5.02]	3.99	**1**
Online	[2.27,5.17]	4.08	3	[1.34,3.72]	3.00	**5**
Entertainment	[2.17,5.27]	4.03	4	[1.54,4.29]	3.33	2
Friendship	[2.51,4.95]	4.43	2	[1.34,3.36]	3.17	4

Notes. N= 289, males=138, females=151; *R*=defuzzification.

Fuzzy data transform by college difference

Students in college of management reflected that they spent more time on part-time jobs [1.95,4.88], but the online activity took them relatively low time [0.77,3.27] than did other activities. Students in college of humanities also showed spent more time on part-time jobs [2.94,5.74]. However, students

in college of science tended to spend more time on their friendships [3.21,5.98], *R*=5.08. Compared with other college students, science major students had also shown spent more time on other activities. The results of fuzzy means and Defuzzification (*R*) showed on Table 3.

Table 3. Students' time spent on campus activities among different colleges

\Colleges	Management		Humanities		Science	
Activities	Fuzzy mean	*R*	Fuzzy mean	*R*	Fuzzy mean	*R*
Reading	[0.92,3.27]	2.55	[2.49.4.92]	4.04	[2.76,5.80]	4.60
Work	[1.95,4.88]	**3.78**	[2.94,5.74]	**4.61**	[2.74,5.93]	4.71
Online	[0.77,3.27]	**2.50**	[1.63,3.71]	3.11	[2.86,5.78]	4.67
Entertainment	[1.18,4.00]	2.96	[2.17,4.57]	3.75	[2.42,5.59]	**4.37**
Friendship	[0.88,3.53]	2.85	[0.86,2.34]	**2.47**	[3.21,5.98]	**5.08**

Notes. N=289, college of management=128, college of humanities=35, college of science=126; *R*=defuzzification; RK=Ranking.

The fuzzy mean and defuzzification (*R*) reflected the different intention to study abroad in different student groups. The short term stay and master program have been found more preferred by the college students. Currently, study abroad for doctoral program seems too far for the college students. Typically, female students' intention to study abroad is higher than that of male students in terms of comparing their defuzzification (*R)*. This study also found that the students in college of humanities have higher intention to study abroad. The results of fuzzy transform showed on Table 4.

Table 4. Students' intention to study abroad by gender and college majors

	Short term program		Master Program		Doctor Program	
	Fuzzy mean	*R*	Fuzzy mean	*R*	Fuzzy mean	*R*
Gender						
Male	[1.58,4.23]	3.42	[1.62,4.29]	3.51	[1.38,4.41]	3.29
Female	[2.15,5.07]	3.95	[1.91,4.68]	3.69	[1.70,4.62]	3.49
Colleges						
Management	[1.81,5.02]	3.70	[1.80,4.81]	3.65	[1.49,4.51]	3.31
Humanities	[2.91,5.69]	4.62	[2.29,4.77]	3.91	[2.20,4.97]	3.85
Science	[1.68,4.18]	3.47	[1.64,4.32]	3.52	[1.35,4.47]	3.33

Fuzzy correlation with campuses activities

Among the campus activities, reading, work on part-time, online activities, entertainment, and friendship have shown positive correlation with fuzzy computing. The result revealed that reading with online activities had shown high fuzzy correlation (.72,.94). Entertainment and online activities showed high fuzzy correlation (.68,.94). Friendship with online activities also showed high fuzzy correlation (.65,.84), see Table 5. This study found that the online activity played an important role to make the student engagement in campus more meaningful.

Table 5. College students engaged in five campus activities explained by fuzzy correlation

Activities	Reading	Work	Online	Entertainment	Friendship
Reading	1				
Work	(.57, .78)	1			
Online	**(.72, .94)**	(.46, .65)	1		
Entertainment	(.62, .84)	(.61, .83)	(**.68, .94**)	1	
Friendship	(.54, .70)	(.26, .41)	**(.65, .84)**	(.44, .61)	1

Fuzzy correlation with intention to study abroad by different programs

Will students successfully engage in campus activities affect their intention to study abroad? In this study, we cannot find satisfied evidences to support this argument. We found the students' time

spent on different activities in campus is only showed low correlation with the intention to join the different study programs in abroad, see Table 6.

Table 6. Fuzzy correlation with activities engaged and intention to study abroad

Fuzzy Correlation	Reading	Work	Online
Sort term program	(.05, .12)	(.10, .21)	(.05, .15)
Master program	(.06, .16)	(.15, .28)	(.02, .14)
Doctor program	(.14, .25)	(.20, 33)	(.09, .20)

5. Conclusion

Studying in domestic campus or intention to study abroad is not just a question can be simply responded "yes" or "not". This study found fuzzy questionnaire is more practical way to used to collect data from ambiguous issues. Defuzzification is an optional way to explain the fuzzy interval data in their traditional meanings. The fuzzy mean and fuzzy correlation can be used to interpret the data properly.

For further development, the research idea could be extended to more practical utility for tackling the other issues in campus or in other fields by way of designing a fuzzy measurement. Fuzzy measurement could be implemented in web page to collect and transform dynamic fuzzy data. This study will suggest using Excel to create a menu-driven system, especially take advantage of applying the macro commands to solve the complex fuzzy data transformation.

References

[1] A. Kirkwood, R. Branyan: *Difficult Situations in the Learning Environment: A Resolution Process*, NADE Conference Selected Paper, 7, 41-44 (2001).

[2] B.R. Clark, M. Trow in: *College Peer Groups: Problems and Prospects for Research,* edited by T.M Newcomb and E.K. Wilson, 17-70, Aldine, Chicago (1966).

[3] D. Gray: *Global Engagement in a Virtual World*. Paper Presented at the Assuring a Globally Engaged Science and Engineering Workforce Workshop of National Science Foundation, Washington, DC (September 20-22, 2006).

[4] H. Hsu, B. Wu: *An innovative Approach on Fuzzy Correlation Coefficient with Interval Data*. IJICIC, 6, 3(A), 1049-1058 (2010).

[5] H. Nguyen, B. Wu: *Fundamentals of Statistics with Fuzzy Data*. Springer, Heidelberg (2006).

[6] I. Chen: *Undergraduates Are Not Engaged in Learning*, The Liberty Times, http://www.libertytimes.com.tw/index.htm

[7] J. Brennan, R. Edmunds, M. Houston, et al.: *Improving What is Learned at University*, Routledge, London (2010).

[8] J.F. Chang: *Fuzzy Inference for Assessing Process Lifetime Performance*, IJICIC, 3, 6(B), 1729-1742 (2007).

[9] L.A. Zadeh: *Fuzzy Sets*. Information and Control, 8, 3, 338-358 (1968).

[10] M. Cooper: *Differences in Personality Among the Clark-Trow Student Types*, 497-507, http://aare.edu.au

[11] P. Blumenthal in: *The Virtual Challenge to International Cooperation in Higher Education,* edited by B. Wachter, Lemmens, Bonn, Germany (2002).

[12] S.W. Wang, D.F. Chang, B. Wu: *Does Technologies Really Help Digital Natives? A Fuzzy Statistical Analysis and Evaluation of Students' Learning Achievement*, IJIMIP, 1, 1, 18-30 (2010).

Using Soft Computing to Assess the Issue of Time Management

Dian-Fu Chang[1, a] and Wen-Ching Chou[2, b]
[1]Graduate Institute of Educational Policy and Leadership, Tamkang University, Taiwan
[2]Dept. of Education Policy and Administration, National Chi Nan University, Taiwan
[a]140626@mail.tku.edu.tw, [b]julinchou@yahoo.com.tw

Keywords: time management, fuzzy statistics, fuzzy distance, index of efficiency

Abstract. This study applied the fuzzy statistics to analyze time management issues in campus. The self-designed fuzzy questionnaire was used to collect time management data from 330 college students in Taiwan. This study addressed how fuzzy statistics can be applied for tackling ambiguity questions of time management issues. The results revealed that male and female students showed similar strategies of time management. However, the college differences showed varied time management styles among those students. This study demonstrated fuzzy distance and index of efficiency (*IOE*) provided more specifically information to interpret what college students intent to do and what they have done. By way of fuzzy statistics, the study can decode the students' time management related data in different campus more properly.

Introduction

Time management has become a hot issue in many organizations. Recently, students' time management issues in campus also has become public concerns. Harson and his colleagues surveyed college students in a study day, they found most of personal time is spent on face to face conversation, talking on cell phones, and the use of social networking sites and various forms of human communication [3]. Tanner and his colleagues found business school students in higher grades might manage their time more effective when they face to select reading books, watching TV, and doing other project[11]. The study showed that over 70% of the students in campus did not spend time on education related activities. There are 75.7% of the students only took 1-5 hours in the learning [8]. Reviewing these studies, we know many students did not manage properly their time during study in colleges, or even ignore the importance of time management.

The concept of fuzzy set proposed by Zadeh and applied to fuzzy measurement to cope with the dynamic environment [13]. Recently, the concept of fuzzy set has given a more reasonable description for different purposes of data transform [2,10,12]. Many articles displayed that the fuzzy statistics has become a useful tool for measuring ambiguous concepts in science and social science [2,4,9]. Their argument is that the traditional numerical model cannot explain complex and ambiguous human and social phenomena properly. Therefore, they encourage to use the soft computing to solve the complex and ambiguity phenomena [7]. Using fuzzy data transform, this study create a new way to assess the efficiency of time management issues.

When undergraduates attend university, their expectation about college life is based on popular culture, the media, or their idea? Various studies focus on the college learning issues, some of them tackle the engagement issues and trying to better define what students engaged in campus [1,5,6]. This study designed a fuzzy questionnaire to collect the college students' perceptions about their time management. We applied the fuzzy data transform to further interpret the campus life in terms of their time management in different daily activities. Given the purposes of study, the research questions are as follows:

(1) Do they have any difference of time management in gender?

(2) Do they have any difference of time management in colleges?
(3) What kind of time management shows by daily campus life among these students?
(4) Which type of time management efficiency shows in these activities?
(5) Can the fuzzy index of efficiency be used to interpret time management issues properly?

The framework of study

This study applied fuzzy statistics to demo how to transform fuzzy means, defuzzification for a fuzzy number on *R*, fuzzy distance in campus activities, and fuzzy index of efficiency related to time spent in campus. In this study, the definitions of fuzzy statistics show on definition 1, definition 2, definition 3, and definition 4. The Excel's spreadsheet was used to code and decode the fuzzy data. The study framework shows on Fig. 1. This study assumed the students' time spent on campus activities are different from gender and colleges. If the students show the intent time spent and actual time spent are different, then the fuzzy distance will exist in those activities respectively. The index of efficiency is used to interpret the efficiency of time management among those students.

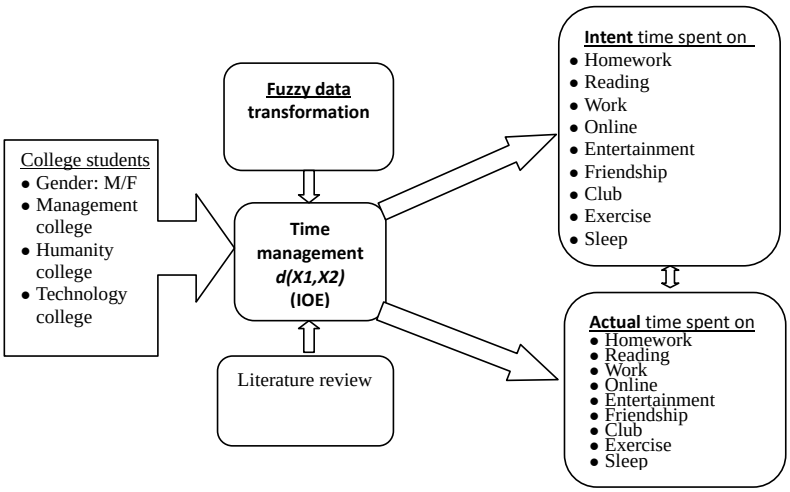


Fig. 1 A framework of analysis

Fuzzy questionnaire and samples

In this study, the time management has taken account of two dimensions: actual time spent and intent time spent on campus activities. The fuzzy questionnaire includes the following items: gender, college major, time spent in different activities with actual and intent time spent. The variables related to campus life in the questionnaire are as follows:

(1) Homework refers to doing homework and writing reports;
(2) Reading refers to reading magazines, other books which are not related to current major;
(3) Work in terms of campus or off campus part-time job;
(4) Online means online homework, it refers to using Internet to find related information to write report or to do homework;
(5) Entertainment includes online games, on social networking site to talk, talking on cell phones, travel;
(6) Friendship means the relationship with others, including make friends;
(7) Club refers to student's extra curriculum activities;
(8) Exercise refers to do exercise after classes.
(9) Sleep refers to how many hours spent on sleep per day in actual and intent to do.

This study focused on students' perception and intention of time management, the data were

collected by using fuzzy questionnaire. The examples of completing fuzzy questionnaire are as follows:

***Direction**: The following questions are related to your perceptions of your actual time spent on the campus activities and your intention of time spent on these activities. We need your opinions about actual time spent on selected campuses activities and your perception of ideal time spent on these different activities.*

*1. If your daily spent on **reading** is about 3 to 6 hours, you need circle the actual time spent with the number 3 and 6 on the scale 0-9.*

Please start cycle your perceptions in following activities:

Reading refers to reading magazine, other books which are not related to your current major.

Actual your time spent

0	1	2	③	4	5	⑥	7	8	9

2. If you feel the time spent on the activity should be 2-5 hours, it means your ideal time spent on the activity is from 2 to 5 hours on the scale of 0-9. You need circle the number 2 to 5 on the scale. Please start cycle your intention of time spent in the following activities:

Reading refers to reading magazine, other books which are not related to your current major.

lntention of time spent

0	1	②	3	4	⑤	6	7	8	9

The samples were collected from two national and one private universities in central and southern of Taiwan in October, 2011. The total valid questionnaires are 330, males are 138 and females are 151 in three different major colleges, see Table 1.

Table l. The samples distribution of the target colleges.

Gender	Colleges			Total
	Management	Humanities	Technology	
Males	56	24	77	157
Females	85	34	54	173
Total	141	58	131	330

Methods

Definition 1. Let U be the universal set and $\{Fx_i = [a_i, b_i], a_i, b_i \in R, i = 1,...,n\}$ be a sequence of random fuzzy sample on U. The fuzzy mean value is then defined as [7]

$$F\bar{x} = [\frac{1}{n}\sum_{i=1}^{n} a_i, \frac{1}{n}\sum_{i=1}^{n} b_i] \quad (1)$$

Definition 2. Let χ=[a, b] (a≠b) (be an interval fuzzy number on U). The defuzzification number R_χ of [a, b] is then defined as [7]

$$R_x = \frac{a+b}{2} + (1 - \frac{\ln(1+|b-a|)}{|b-a|}) \quad (2)$$

Definition 3. Let U be the universe of discourse. Let $\{\chi_i$=[a, b], i=1,2$\}$ be two samples from U, with center $c_i = \frac{ai+bi}{2}$, and radius $r_i = \frac{ai-bi}{2}$, the distance between the two samples χ_1 and χ_2 is defined as

$$d(\chi 1,\chi 2) = \left|C_i - C_j\right| + \left|\frac{\ln(1+|bi-ai|)}{|bi-ai|} - \frac{\ln(1+|bj-aj|)}{|bj-aj|}\right| \quad (3)$$

The *d(χ1, χ2)* is used to compare the difference of actual and intent time spent. This study

interpreted the formula as *d(I, A)* to follow the research purposes, which *I* refers to intent of time spent and *A* represents actual time spent.

Definition 4. Let U be the universe of discourse. Let $OI=(c_o; a_o)$ be the observed data and $EI=(c_e; a_e)$ be the expected data from U. The index of efficiency (*IOE*) with fuzzy distance is defined as

$$IOE=e^{-\left(\left|\frac{c_o-c_e}{c_e}\right|+\ln\left(1+\left|\frac{a_o-a_e}{a_e}\right|\right)\right)} \quad (4)$$

Where c_o and c_e represent the center of the observed and expected values, a_o and a_e represent the area of the observed and expected values with fuzzy membership function.

Results

This study measured gender and college major to compare students' time management using fuzzy means, defuzzification *R*, fuzzy distance, and index of efficiency. The fuzzy statistics were used to domo how they can be used to decode the ambiguous issues properly.

Fuzzy mean and defuzzification *R*

Table 2 shows the fuzzy mean and defuzzification *R* of students' time management in campus. The time spent on sleep listed in first, then followed the entertainment, online, work, and homework. But the reading listed in last one. The students also expressed the intention of time spent, the result revealed beside the sleep, homework listed in number two, and then followed work, reading, and friendship. The result shows a gap between actual time spent and intent time spent in campus. The percentage of time management in terms of 24 hours in a day has shown on Fig. 2. The result revealed that over 65% of time spent on campus activities did not link to education purposes.

Table 2. Fuzzy mean and Defuzzification *R of* time management of activities in campus.

Activities	Actual time Spent			Intent time spent		
	Fuzzy mean	*R*	Ranking	Fuzzy mean	*R*	Ranking
Homework	[1.80,4.42]	3.44	**5**	[2.33,5.33]	4.18	2
Reading	[1.64,4.27]	3.28	**7**	[1.73,4.60]	3.51	4
Work	[1.78,5.01]	3.76	**4**	[1.70,4.90]	3.66	**3**
Online	[2.01,4.90]	3.80	3	[1.62,4.32]	3.30	**7**
Entertainment	[2.37,5.60]	4.34	2	[1.71,4.54]	3.43	6
Friendship	[1.54,4.65]	3.44	5	[1.56,4.63]	3.44	5
Club	[0.30,0.88]	0.70	8	[0.29,0.76]	0.62	9
Exercise	[0.34,0.75]	0.64	9	[0.37,0.84]	0.70	8
Sleep	[3.99,7.82]	6.30	1	[4.40,8.45]	6.82	1

Notes. N= 330, males=157, females=173; *R*=defuzzification.

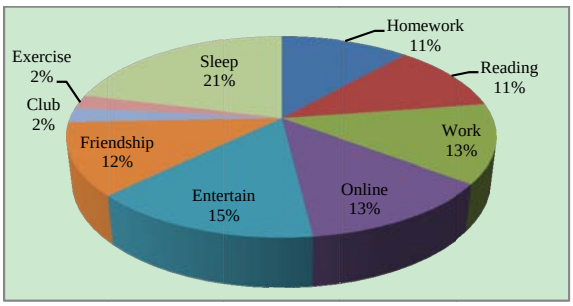


Fig. 2 College students' time management in 24 hours by percentage

Transformation of fuzzy data by gender

This study found that the gender has shown very similar style of time spent on campus activities. They spent much of time on sleep, entertainment, and part-time job, both of them had less engaged in homework and reading, see Table 3. Time spent on club and doing exercise is very limited among those students. Decoding these data, the study found that online related activities, such as Internet, entertainment, and chat to friends with cell phone, has shared a relative part of students' time spent in a day.

Table 3. Gender difference of actual time spent on campus activities with fuzzy mean and *R*.

Activities	Male			Female		
	Fuzzy mean	*R*	Ranking	Fuzzy mean	*R*	Ranking
Homework	[1.97,4.73]	3.69	**7**	[1.64,4.14]	3.21	5
Reading	[2.05,4.82]	3.77	**6**	[1.27,3.77]	2.83	7
Work	[2.28,5.66]	4.34	**3**	[1.34,4.42]	3.23	**4**
Online	[2.37,5.35]	4.21	4	[1.69,4.50]	3.54	**3**
Entertainment	[2.98,6.32]	5.02	2	[1.82,4.94]	3.73	2
Friendship	[2.03,5.33]	4.05	5	[1.09,4.03]	2.89	6
Club	[0.37,0.96]	0.78	8	[0.24,0.81]	0.64	8
Exercise	[0.44,0.91]	0.77	9	[0.26,0.62]	0.60	9
Sleep	[4.27,8.20]	6.64	1	[3.73,7.47]	5.99	1

Notes. N= 330, males=157, females=173; *R*=defuzzification.

Assess fuzzy distance and index of efficiency

In this study, *d(I,A)* reflected fuzzy distance of the intent and actual time spent, the bigger number means the larger distance between them. However, *IOE* revealed the bigger index is more efficiency in this activity in terms of their time management. The result demonstrate the *IOE* of sleep, club, and work listed in the first three which can be interpreted more efficiency in their time management in these activities. Conversely, when college students face to do homework, reading, and entertainment which showed less efficiency that do other activities. The result also showed the time management patterns of male or female students are very similar, see Table 4.

Table 4. The distances and indices of efficiency (*IOE*) in time management by gender.

Activities	Male *d(I,A)*	Male *IOE*	Female *d(I,A)*	Female *IOE*	Total *d(I,A)*	Total *IOE*
Homework	1.66	0.90	1.53	0.88	1.589	0.890
Reading	1.33	0.90	1.13	0.89	1.226	0.894
Work	1.79	0.96	1.18	1.00	1.470	0.977
Online	1.40	0.90	1.00	0.86	1.192	0.879
Entertainment	1.66	0.83	1.19	0.83	1.412	0.829
Friendship	1.57	0.98	0.94	0.90	1.237	0.941
Club	0.21	1.02	0.19	1.01	0.196	1.019
Exercise	0.20	0.98	0.18	0.97	0.188	0.971
Sleep	1.08	1.13	1.15	1.14	1.115	1.135

Table 5 shows the efficiency of time management in different colleges. The *IOE* in this table reflected on the time management efficiency among those students. The result revealed the students in different colleges also shared different time management styles.

Table 5. The distance and indices of efficiency (*IOE*) in time management by colleges.

Activities	**Management**		**Humanities**		**Technology**	
	d(I,A)	*IOE*	*d(I,A)*	*IOE*	*d(I,A)*	*IOE*
Homework	1.52	0.86	1.36	0.88	1.76	0.93
Reading	1.07	0.87	0.97	0.91	1.51	0.92
Work	1.22	0.99	1.33	1.12	1.80	0.90
Online	1.03	0.85	0.73	0.87	1.57	0.92
Entertainment	1.30	0.82	1.25	0.78	1.60	0.86
Friendship	1.05	0.92	1.13	1.04	1.49	0.93
Club	0.18	0.98	0.20	0.99	0.21	1.08
Exercise	0.18	0.90	0.17	1.09	0.20	1.00
Sleep	1.16	1.16	0.91	0.99	1.16	1.17

Conclusions

Studying on the perception of time management is not just a question which can be simply responded “yes” or “not”. This study found the fuzzy questionnaire is more reasonable way to used to collect data from ambiguous issues. The data transform can be used to different formats to interpret the time management issues. Defuzzification is a way to explain the fuzzy interval data in their traditional meanings. The fuzzy mean, fuzzy distance, and index of efficiency (*IOE*) can be used to interpret the data properly. For additional development, this research idea could be extended to more practical utility to tackle related issues of other fields by designing a fuzzy assessment.

References

[1] J. Brennan, R. Edmunds, M. Houston, et al.: *Improving What is Learned at University*. Routledge, London (2010).

[2] J.F. Chang: *Fuzzy Inference for Assessing Process Lifetime Performance*, IJICIC, 3, 6(B), 1729-1742 (2007).

[3] T.L. Hanson, K. Drumheller, J. Mallard, C. Mckee, P. Schlegel: *Cell Phones, Text Messaging, and Facebook: Competing Time Demands of Today's College Students, College Teaching*, 59, 23 (2011).

[7] H. Hsu, B. Wu: *An innovative Approach on Fuzzy Correlation Coefficient with Interval Data*, IJICIC, 6, 3(A), 1049-1058 (2010).

[8] A. Kirkwood, R. Branyan: *Difficult Situations in the Learning Environment: A Resolution Process*, NADE Conference Selected Paper, 7, 41-44 (2001).

[9] H. Lin, H. Chang, Y. Chen: *Lin Huo Wang’s Criticism: Undergraduates at National Taiwan University only Concerned about Their Own Future*, The Liberty Times. http://www.libertytimes.com.tw/2010/new /jan/22/today-life7.htm

[10] H. Nguyen, B. Wu: *Fundamentals of Statistics with Fuzzy Data.* Springer, Heidelberg (2006).

[11] O B. O.gonor, M. Nwadiani: *An Analysis of Non-instructional Time Management of Undergraduates in Southern Nigeria,* College Student Journal, 40, 204-218 (2006).

[12] T. Samatsu, K. Tachikawa, Y. Shi: *Image Processing for Car Shapes in the Fuzzy Retrieval System*, IJICIC Express Letters B, 1, 1, 1-7 (2010).

[13] C.M. Sun, B. Wu: *New Statistical Approaches for Fuzzy Data*, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 15, 2, 89-106 (2007).

[14] J.R. Tanner, G. Stewart, G.M. Maples, M.W. Totaro: *How Business Student Spend Their Time-Do They Really Know?* Research in Higher Education Journal, 3, 1-9 (2009).

[15] S.W. Wang, D.F. Chang, B. Wu: *Does Technologies Really Help Digital Natives? A Fuzzy Statistical Analysis and Evaluation of Students’ Learning Achievement*, IJIMIP, 1, 1, 18-30 (2010).

[16] L.A. Zadeh: *Fuzzy Sets*, Information and Control, 8, 3, 338-358 (1968).

Predicting Relapse of Hepatocyte Cancer by Combing Regression and Classification Using SVM

Kazuhiro Nakada[1, a], Hayato Ohwada[2,b] and Hiroyuki Nishiyama[2,c]

[1]2641, Yamazaki, Noda-City, Chiba, 278-8510, Japan

[2]2641, Yamazaki, Noda-City, Chiba, 278-8510, Japan

[a]nakata@ohwada-lab.net, [b]ohwada@ia.noda.tus.ac.jp, [c]hiroyuki@rs.noda.tus.ac.jp

Keywords: SVM (Support Vector Machine), SVR(Support Vector Regression), Linear interpolation, Tumor market, Hepatocyte cancer

Abstract. This paper explains how to predict relapse of hepatocyte cancer by combining regression and classification using a support vector machine (SVM) [1]. The method focuses on a particular tumor marker that seems to be an important factor of cancer cells, and predicts the transition of the marker based on linear interpolation and SVM regression. Relapse of hepatocyte cancer is then predicted using the SVM classification function where the predicted transition is incorporated into training data. An experiment demonstrated that our method is superior to a single use of SVM.

1. Introduction

We propose how to predict relapse of hepatocyte cancer by using machine learning in this paper. The feature of this study is treating time-line data by using a classification and regression of machine learning, and this system is able to predict the time of relapse. The study which used machine learning for diagnosis and detection of hepatocyte cancer exists mostly. But, since most of those studies are judging in the stage after relapse, discovery of a relapse is slow. However, this system is discovering earlier hepatocyte cancer, because it predict relapse in the stage before relapse. In this way, it leads to the support for early detection and selecting inspection day after an operation.

This prediction task captures postoperative progress by watching a particular tumor marker that seems to be an important factor of cancer cells, and makes an overall decision from observation data and predicted state transitions of the marker. The transition is predicted using both linear interpolation of marker values and SVM regression. Relapse of hepatocyte cancer is then predicted using the SVM classification function where the predicted transition is incorporated into training data. Figure 1 presents an outline of this study.

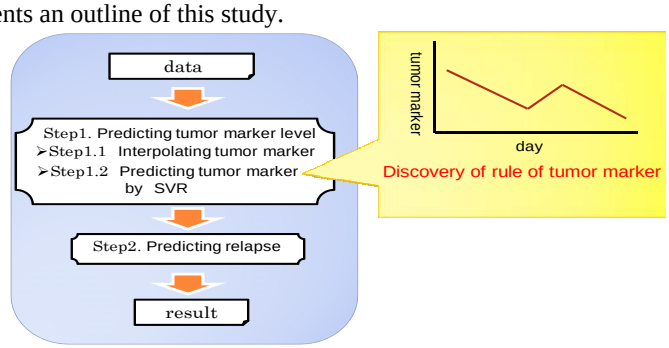


Fig.1 Schema of approach

A procedure until it predicts relapse of hepatocyte cancer is introduced.

```
1 :  y = Log(10)
2 :  x = (Log(Cells(i + 1, 6)) / y - Log(Cells(i + 1, 4)) / y) / (Cells(i + 1, 7) - Cells(i + 1, 5))
3 :  b = Log(Cells(i + 1, 4)) / y - x * Cells(i + 1, 5)
4 :  z = x * Cells(1, 28) + b
5 :  Cells(i + 1, 28) = 10 ^ z
```

Fig.2 Program of interpolating

1st line: Set the variable y to log (10).
2nd line: Ask for inclination of a straight line. (Increase of x / increase of y)
3rd line: Ask for linear y section.
4th line: Calculate the interpolation value of days to obtain.
5th line: Return the account of a table of logarithms.

2. Interpolating tumor marker

A standard medical examination interval for cancer has yet to be established, so the data used in this paper does not have a constant examination interval either. However, in order to predict the days until relapse, those data must be inspected on the same day. This paper therefore uses linear interpolation in order to synchronize the examination day. The date was interpolated using visual basics for applications (VBA) as depicted in Fig. 2. Figure 3 presents an outline of alignment interpolation.

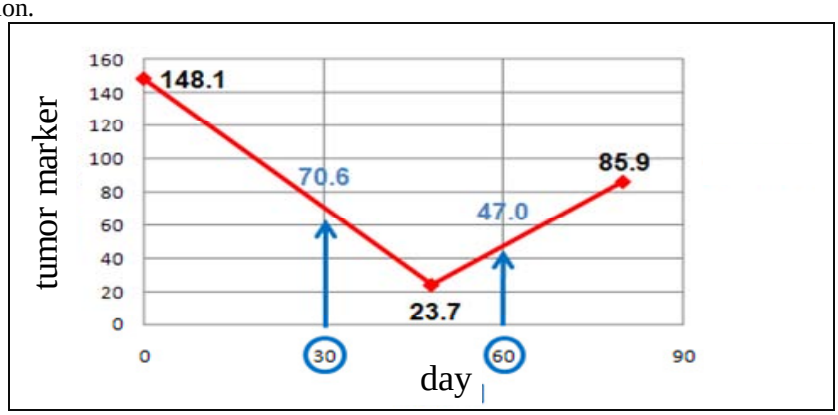


Fig.3 Schema of interpolating

The tumor marker level of the 30th day and of the 60th day is calculated using linear interpolation.

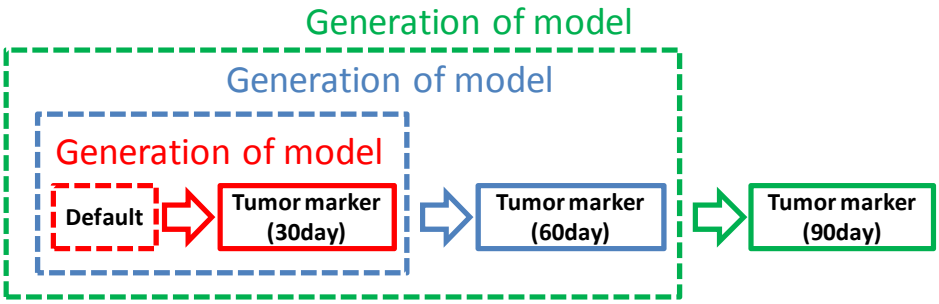


Fig.4 Schema of predicting tumor marker

This shows in detail the method of generation of the model which predicts a tumor marker level.

3. Predicting tumor marker by SVR

The model for predicting the tumor marker is generated using interpolated values as training data. Figure 4 illustrates the procedure until it predicts the value of the tumor marker on the 90th day. First, a model that predicts the value of the tumor marker on the 30th day is created by using postoperative moment data or pre-operation data as training data. Next, a model that predicts the value of the tumor marker on the 60th day is created by combining the value of the tumor marker on the 30th day and postoperative moment data or pre-operation data. Thus, a model that predicts the value of the tumor marker at intervals of 30 days is created.

4. Predicting relapse by SVM

Next, a relapse is predicted using the predicted value of the tumor marker. A model that predicts a relapse on the 30th day is generated by treating the combined value of the tumor marker on the 30th day and postoperative moment data or pre-operation data as training data. Next, a model that predicts the relapse on the 60th day is generated by combining the value of the tumor marker on the 60th day and the data used when generating the model on the 30th day. A subsequent model is also generated similarly at intervals of 30 days.

5. Experiment

An experiment which using two tumor markers (AFP and PIVKA) was conducted in this study using three patterns combining two with each case. Nu-SVR and epsilon-SVR were used to predict the transition of the tumor marker. C-SVM and nu-SVM were used to predict the relapse of hepatocyte cancer. A Gaussian kernel was used, and 154 patients participated in this study. One hundred and one patients had relapses, and 53 did not. There were only 154 cases in this study. Since, there was so little data, the experiment was conducted using the effective appraisal method (leave-one-out method).

First, the predicted transition of the tumor marker is presented. Figure 5 compares the actual transition with the transition in AFP value predicted using nu-SVR and epsilon-SVR. The result of epsilon-SVR was better than the result of nu-SVR, because the result of epsilon-SVR was closer to the actual transition.

Next, the recurrence predicted using epsilon-SVR is presented. Table 1 shows the result of using epsilon-SVR to predict the transition of a tumor marker, and the result of using C-SVM to predict relapse. We conducted experiment every 120 days and every 240 days. In every 120th days, it turns out that sensitivity is errors. As a result, there are no relapses, so no prediction of relapses is issued. Since there was a moderate number of a negative example when predicting every 240 days, a relapse was predicted. Table 2 presents the result when four prediction methods-(nu-SVR, epsilon-SVR, C-SVM, and nu-SVM)-, are combined every 240 days. The result of epsilon-SVR was better than the result of nu-SVR. The percentage of correct answers when predicting that sensitivity does not relapse is expressed. The percentage of correct answers when predicting that specificity relapse is expressed. Finally sensitivity and specificity were measured by the case of the approach and using initial data. The approach employs the result when using nu-SVR. Figures 6 and 7 indicate that the approach has excellent sensitivity and specificity. There is a large difference in sensitivity on the 960th day and a large difference in specificity on the 720th day. There indicating that the tumor marker becomes more important as days pass.

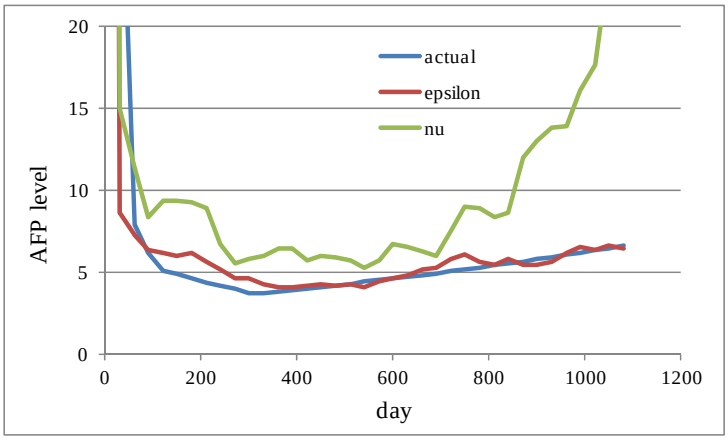


Fig.5 Result of epsilon-SVR and nu-SVR

The vertical axis expresses AFP level. The horizontal axis expresses day.

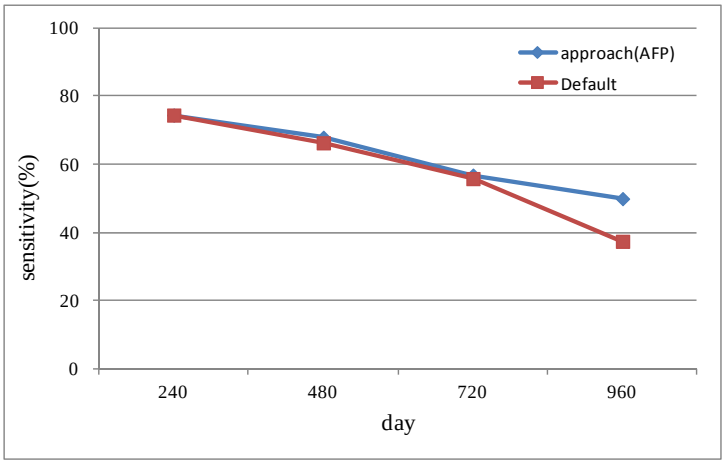


Fig.6 Relation between day and specificity

The vertical axis expresses specificity. The horizontal axis expresses day.

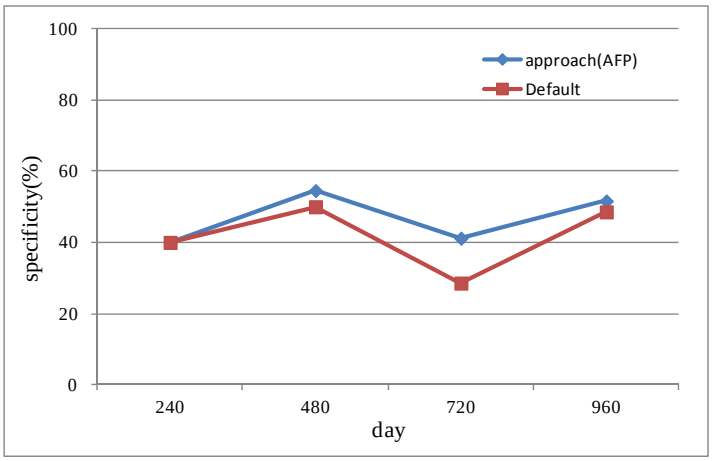


Fig.7 Relation between day and sensitivity
The vertical axis expresses sensitivity. The horizontal axis expresses day.

Table.1 The sensitivity and the specificity of each combination(Every 120th days)

	day	120	240	360	480	600	720	840	960	1080
epsilon_C	sensitivity	94.16	78.62	82.46	79.79	78.67	72.88	79.07	60.61	76.19
	specificity	error	error	error	error	error	error	error	0.00	error
nu_C	sensitivity	94.16	78.62	82.46	79.79	78.67	72.88	79.07	60.61	76.19
	specificity	error	error	error	error	error	error	error	0.00	error

Table.2 The sensitivity and the specificity of each combination (Every 240th days)

	day	240	480	720	960
epsilon_C	sensitivity	73.86	67.31	57.63	46.15
	specificity	0.00	50.00	43.75	50.00
epsilon_nu	sensitivity	74.32	67.65	57.89	45.45
	specificity	33.33	50.00	44.44	50.00
nu_C	sensitivity	74.83	67.62	56.90	46.15
	specificity	66.67	55.56	41.18	50.00
nu_nu	sensitivity	74.50	67.96	56.90	50.00
	specificity	40.00	54.55	41.18	51.61

6. Conclusion

The transition of the tumor marker was predicted and the predicted transition value was used to predict future relapse. The value of AFP, tumor marker, was used experiments in combination four prediction methods (nu-SVR, epsilon-SVR, C-SVM, and nu-SVM), respectively. The tumor marker PIVKA is used in addition to the AFP value. I think that AFP indicates the possibility of a future recurrence. Only SVM is used in this study. Better accuracy may be achieved by comparison with the result obtained using other machine learning.

The proposed method can be also regarded as a unified framework to guess future affects by applying to predict the time-line of brain waves, yielding a foundation of affective computing.

Acknowledgments

We express our gratitude to Mr. Nakatura and Mr. Nobuoka from the National Cancer Center East Hospital that supplied the data.

References

[1] V. Vapnik: The Nature of Statistical Learning Theory, Springer-Verlag, 1995.

[2] J.A. Cruz and D.S. Wishart. Applications of machine learning in cancer prediction and prognosis cancer informatics.vol.2,p.59,2006

A New Updating Strategy in Simulating Emergency Evacuation

Yugang Zhang[a], Hongjun Xue[b]

College of Aeronautics, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

[a]zhang_yu9999@163.com, [b]xuehj@nwpu.edu.cn

Keywords: emergency evacuation, algorithms, simulation, cellular automata, updating strategy

Abstract. To evaluate the total evacuation time in emergency condition accurately, the microcosmic discrete evacuation model are developed based on CA (Cellular Automata) and MAS (Multi-agent System). The model considered individual's knowledge of the environment, his realization of the emergency development, and the repulsive influence between occupants. This model differs very much from traditional models in the updating strategy, which based on "information spread". Finally, for comparing the proposed algorithm with traditional evacuation models, the author gave a thin-body civil aircraft cabin for numerical simulation. The results demonstrate preliminarily that the effectiveness and accuracy of the proposed algorithm as applied to aircraft cabin are better than those of traditional evacuation models are. The updating progress based on "information spread" is more like the whole occupants' choosing sequence in real world.

Introduction

There are several difficulties related to the 90-seconds certification demonstration when new designed aircraft be validated, which is the threat of life to the trial participants [1], the expensive cost and the long certification period [2]. Thus, the demonstration is only required one time. For the participants' stochastic behavior, the certification results cannot represent the real evacuation performance of the aircraft. The result is impossible to verify whether the plane meets the demands of safety for using and the result cannot verify whether the certified cabin configuration is optimal in terms of evacuation efficiency for design. To solve the problems mentioned and the results influencing by the stochastic behavior of participants, some researchers suggest the computer be used and emphasize the effect of simulation method in the certification demonstration application [3].

At present, there are mainly three theories for emergency evacuation. The first is building the macro-equation of passengers based on the traditional fluid dynamic theory [4], which is used to modify the evacuation approach by early designer. The second is building the kinetics micro-equation according to individual movement [5], which is used to study how the individual behavior parameters affecting the evacuation ability. The third is building rules on the discrete grids by computer [6], which belongs to the micro method. Macro-equation method cannot explain the self-organization phenomenon, though it can describe the statistic attributes. The micro-equation can explain the self-organization phenomenon but for the complicated equation. While the model with rule based on discrete grids is able to simulate the movement of passengers in complicated environment, thus the model is used to study emergency evacuation theory abroad. The method namely Cellular Automata is presented firstly by Von Neumann and associated scientists at the end of 1940s for researching of life science [7]. The famous mathematics, physics and computer scientist of Stephen Wolfram defined the method [8]. Schadschneider and Burstedde [9, 10] present the "field potential" to build up the two-dimension movement CA model for simulating movement of

passengers, which is base of the present emergency evacuation model. However, there are still some difficulties, for instance, is the parallel updating schemes can be applied in any scene?

In order to research the issues outlined above we proposed a new model. Our contribution is to extend CA model's updating strategy based on "information spread". The method considers the dynamic influence of individual nature for evacuation process by using multi-agent technology. In the next section, we will simulate the emergency evacuation of a typical cabin, and present the results and analyses.

Emergency evacuation model

The model based on CA and MAS is presented to simulate the emergency evacuation process of civil aircraft. The logical figure is shown in Fig. 1.

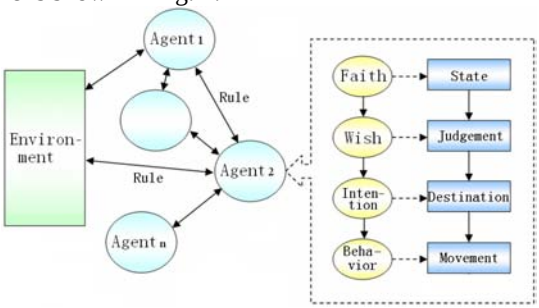


Figure 1. The logical of the model.

An agent has three attributes, which is faith, wish and intention. The faith describes understanding of environment by agent and represents the possible state, for example, the static state of enclosure and the dynamic state of occupants around, which will be acquired instantaneously when simulating. The wish coming from faith describes judgment for the scene, which will be described by following direction-choosing method. The intention coming from wish is the component of destination, which restricts the agent. In this simulation, the destination is how to escape from the opened exits. All the behavior of agents is accomplished by the interaction of environment and the agents. The interaction of agents or the agent and environment is linked and restricted by CA method.

There are mainly two processes of the model, which is choosing the evacuation direction and updating position.

Direction-choosing method. The agent can choose eight directions. The agent's final direction is affected by the local environment factors around. The probability of every direction (Fig. 2) is calculated as,

$$p_{ij} = Nn_{ij}m_{ij} / \exp(k_s S_{ij} + \sum_k k_{dk} D_{ijk}) \quad (1)$$

where N is normalization factor, which makes the sum of p_{ij} equal 1.0. n_{ij} is obstacle factor such as seats, gallery, etc. $n_{ij} = \begin{cases} 0 & \text{Obstacle} \\ 1 & \text{Have no obstacle} \end{cases}$. m_{ij} is occupant hold factor. $m_{ij} = \begin{cases} 0 & \text{hold but not himself} \\ 1 & \text{none or himself} \end{cases}$. k_s is static factor of experiential level with the scene, $k_s \in [0, \infty)$. That k_s equal 0 means agent (occupant) totally strange with the scene. The agent more familiar with the scene chooses moving direction more reasonably. S_{ij} is static field value, which represents the route distance from current position to exit. k_{dk} is dynamic factor with scene, like k_s. D_{ijk} is dynamic field value. The dynamic factors such as relation between people, fire and so on could affect the final direction.

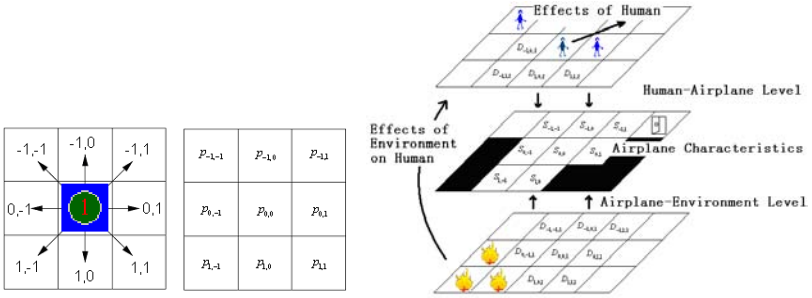


Figure 2. A particle's possible transitions and the associated matrix of preference.

Position updating strategy. Fig. 3 is the snapshot of A380 certification trial video. From the snapshot we can know that there had no interspaces, when the front passengers moved forward, the followers followed at the time. The traditional CA utilized parallel updating strategy. If someone else blocks the egress route, the agent will wait until the grid is empty. Therefore, the parallel updating strategy cannot consider the following movement phenomenon.

Figure 3. A380 certification trial video snapshot.

The occupant's evacuation is a decision process of occupant, which costs for a time and the relationship between the time and the updating time step affects the evacuation pattern. To simulate this instance, we built the updating strategy based on "information spread".

"Information spread" updating strategy is defined that when the origin(exit) point begin to stimulate a "vacant" signal, it firstly spread to its neighbour cells(Fig. 4) which tagged "1", and sequentially spread to their neighbor cells which tagged "2" and so on. The spreading process will take for some time. The time can be regard as the beings single reaction time. Murakami E. et al. has measured the time necessary to process the visual cue information and send motor command to the muscles varied from 0.15s to 0.29s in single reaction time experiment [11]. When the time delay accumulates an updating time step, the blocked people will not move, and wait to next circle. The updating process spreads around, and it will update the whole scene gradually.

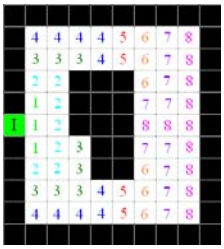
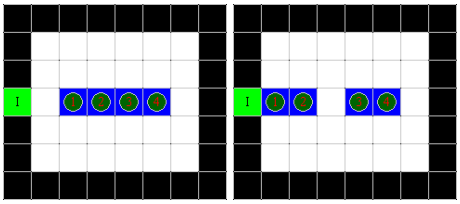


Figure 4. Updating strategy based on "information spread".

Shown in Fig. 5(a), we assume that the response time is 0.25s and time step is 0.5s. The information

spread speed is 2 grids/time step, then the position is showed in Fig. 5(b) after one time step. Apparently, when the updating speed is 1 grid/time step, it turn into parallel updating. When the updating speed is ∞ grids/time step, it can simulate “quick march” phenomenon.



(a) The origin scene (b) The scene after one time step
Figure 5. Updating strategy based on information spread.

Numerical simulation of narrow-bodied aircraft

We choose a cabin configuration shown in figure 6 to simulate our model. The aircraft had 159 seats but seated 149 passengers and had three pairs of exits in total. Type-C exits were positioned at either end of the passenger cabin sections. These exits have been labeled R1 and R3. A further pair of Type-III over wing exits accessible over seating was located at approximately the centre of the cabin section, approximately in line with the wing. This exit was designated as R2. Seat rows were in a 3-3 configuration with a central separating aisle (see Fig. 6). The evacuation results are that TET (Total Evacuation Time) is 64.1s, OPS (Optimal Performance Statistic) is 0.02 [12].

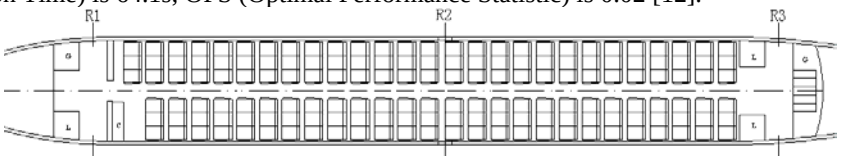


Figure 6. The cabin configuration.

The predictions for the cabin are presented in Fig. 7 (1000 simulations). From Fig. 7 it can be seen that airEXODUS generated TETs range from 62.9 to 89.8 seconds with an average TET of 70.5 seconds. Our models generated TETs range from 60.9 to 77.5 seconds with an average TET of 68.5 seconds. The airEXODUS average TETs is 10% (6.4 seconds) longer than the single measured evacuation time and our models is 7% (4.4 seconds) longer.

Examination of the generated TET frequency distribution demonstrates that the majority (94.8%) of simulations were longer than the evacuation time measured during the certification trial (see Fig. 7). This suggests that the evacuation time measured during the certification trial is positioned towards the extreme low tail of TETs generated by airEXODUS and our model.

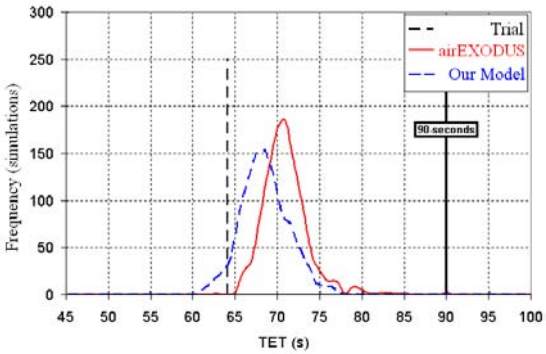


Figure 7. Frequency distribution of TETs.

To summarise, our model's predictions is similar to airEXODUS. It is capable of generating a range of TETs that include the single evacuation time measured during the certification trial.

Conclusions

A key component of our evacuation model is the use of the "information spread" updating strategy. The generalised data is a statistical composite of all available data from previous certification trials. The simulation results indicate preliminarily that the effectiveness and accuracy of our algorithm as applied to aircraft are better than those of traditional CA models are. It was also shown that the model is able to reliably predict the likely evolution of the evacuation from its start to its completion.

The analysis has also highlighted the inability of the current certification process to meaningfully rank aircraft performance, based on a single trial result due to the probabilistic nature of the evacuation process. In order to rank aircraft performance it is necessary to undertake repeated evacuation trials. Alternatively, computer simulation could be used to generate the total evacuation time probability distribution and base a ranking system on the statistical information provided by such a distribution.

References

[1] McCauley H W. *MD11 - Emergency Evacuation Demonstration - Occupant Factors Analysis*. MDC 92K1206, 1992.

[2] OTA. *Aircraft Evacuation Testing: Research and Technology Issues - Background Paper*. TR OTA-BP-SET-121, 1993.

[3] Galea E R, Owen M, Lawrence P J, et al. *Computer Based Simulation of Aircraft Evacuation and its Application to Aircraft Safety*. Proceedings 1998 International Aircraft Fire and Cabin Safety Research Conference, 1998, pp. 8-15.

[4] Henderson L F. *The statistics of crowd fluids*. Nature, 1971, pp.229-381.

[5] Helbing D, Farkas L, Vicsek T. *Simulating dynamical features of escape panic*. Nature, 2000, 407(28), pp. 487-490.

[6] Schadschneider A, Kirchner A, Nishinari K. *CA Approach to Collective Phenomena in Pedestrian Dynamics*. ACRI 2002, LNCS 2493, pp.239-248.

[7] Von Neumann J. *The general and logical theory of automata*. In L.A. Jeffress, editor, Cerebral Mechanisms in Behavior, John Wiley, New York, 1948, pp.1-4.

[8] Wolfram S. *A new kind of science*. Wolfram Media Inc. U.S.A. 2002, pp.23-50.

[9] Schadschneider A. *Cellular automaton approach to pedestrian dynamics - theory*. Pedestrian and evacuation dynamics, Springer, New York: 2002, pp.75.

[10] Burstedde C, Kirchner A, Klauck K. *Cellular automaton approach to pedestrian dynamics - applications*. Springer, New York: 2002, pp.87.

[11] Murakami E, Matsui T. *Human Control Modeling Based on Multimodal Sensory Feedback Information*. Schmorrow D.D. et al. (Eds.): Augmented Cognition, HCII 2009, LNAI 5638, 2009, pp.192-201.

[12] Galea E R, Blake S J, Lawrence P J. *Report on the Testing and Systematic Evaluation of the airEXODUS Aircraft Evacuation Model*. CAA Paper 2004/05, 2005, pp.50-52.

A New Literature Search System with Thesaurus for Biomedical Literatures

Kazuhiro Tanaka[1, a], and Hayato Ohwada[2, b]
[1, 2]2641, Yamazaki, Noda-city, Chiba, 278-8510 Japan
[a]tanakakazuhirodesita@gmail.com, [b]ohwada@ohwada-lab.net

Keywords: Literature search system, Biomedical, PubMed, MeSH, LSD

Abstract. This paper proposes a new literature search system that provides a Web-service in which implicit relationships between concepts can be extracted to support biomedical researchers. This system works on PubMed and includes thesauruses such as MeSH and LSD. It enables us to realize a search function for literature that does not include the keyword in a query. Furthermore, the concept category in a thesaurus can be used to interactively find user-desired literature. The advantage of these functions is demonstrated through life science examples.

Introduction

Research literature is increasing daily in the field of biomedicine, and it is difficult for researchers to read and understand a large number of works. A service has been developed to extract relevant information efficiently from a large amount of literatures. PubMed provides such a Web service for the research on biomedicine. It utilizes MEDLINE, which is the most commonly used biomedical database in the world.

Here, the researcher who wants to examine the literature and research papers about influenza inputs “influenza” into a text box and submits it. PubMed searches for research literature that includes the word “influenza.” However, in some cases, a researcher may want to find literature that is implicitly related to “influenza.”

In this paper, we utilize PubMed to provide a new service in which implicit relationships between concepts can be extracted, yielding support for developing researchers’ ideas. For this purpose, we have developed a literature search system using thesaurus such as MeSH and LSD [1,2]. This system enables us to realize a search function for literature that does not include keywords in a query. Furthermore, the concept category in a thesaurus can be used to interactively find user-desired literature.

This paper is organized as follows. Section 2 describes our system function. Section 3 describes output results of the new literature search system. Section 4 compares the performance of the proposed system and that of the traditional system. Section 5 provides conclusions.

Organization of the Text

Configuration. This paper, develops a literature search system using thesauruses such as MeSH and LSD. We compare MeSH to LSD by using 100 often used biomedical terms [3], and develop a database that organizes MeSH and LSD. We then develop a system in which the processing result from the database is returned to a client. The system uses Tomcat (the servlet container) and the Java servlet.

The literature search system is built as depicted Fig. 1.

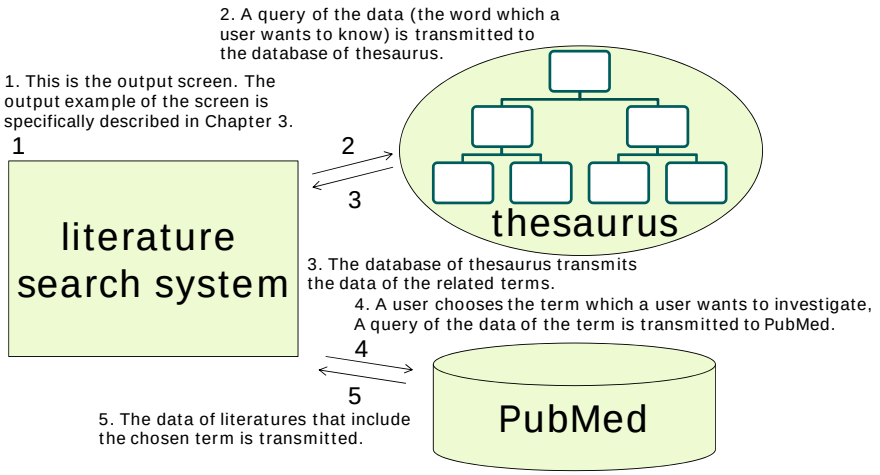


Fig. 1. Flow chart of the proposed literature search system.

In addition, we bring in the category search in output of a thesaurus screen for the user's convenience. Category search is an information search technique that gradually narrows the purpose by choosing the hierarchical classification prepared beforehand. For example, when “influenza” is input into a text box and transmitted, synonyms of “influenza” are not output immediately, but rather the category of the biomedical term previously determined is output, so synonyms of “influenza” are easier to discover. For this purpose, biomedical terms have been divided into 18 categories (Fig. 2).

Anatomy Category, Organisms Category, Diseases Category, Chemicals and Drugs Category, Analytical, Diagnostic and Therapeutic Techniques and Equipment Category, Psychiatry and Psychology Category, Phenomena and Processes Category, Disciplines and Occupations Category, Anthropology, Education, Sociology and Social Phenomena Category, Technology and Food and Beverages Category, Humanities Category, Information Science Category, Persons Category, Health Care Category, Pharmacological Actions Category, Publication Type Category, Check Tags Category, Subheadings Category, Geographical Locations Category

Fig.2. The category of the biomedical term.

A general literature search system uses almost a full-text search, which explores only the literature containing a search keyword [4]. With the proposed system, a user who does not have a keyword can refer to a thesaurus to explore related literature.

data set. We compare MeSH to LSD and develop a database that organizes MeSH and LSD. The database includes features of both MeSH and LSD, and can thus extract more related words. At present, it investigates 100 often used biomedical terms (Fig. 3).

influenza, atopic, rhinitis, asthma, antidote, antipyretic, antineoplastic agent, respiratory distress, hypnotic, fomentation, esophagitis, myocardial, heart failure, urticaria, tranquilizer, enteritis, gout, epilepsy, diabetes mellitus, cerebral infarction, cerebral hemorrhage, pneumonia, lung cancer, obesity, peritonitis, arrhythmia, cystitis, dysgeusia, tinnitus, dizziness, sweat, verruca, dentures, ataxia, malmutrition, hyperopia, inflammation, vomiting, nausea, overeating, arthralgia, wound, chest pain, myopia, myalgia, convulsion, hematuria, belching, thirst, odor, stomatitis, high fever, dyspnea, pigmentation, periodontosis, tongue, toothache, syncope, eczema, gum, numbness, bleeding, delivery, indigestion, headache, cough, malaise, stridor, pallor, weight loss, weight gain, lightheadedness, alopecia, weakness, sputum, cyanosis, disturbance, hematemesis, sore throat, urination, nausea, epistaxis, rhinorrhea, swelling, dermatitis, erosion, fatigue, sleeplessness, tremor, tonsillitis, constipation, taste, caries, heartburn, emaciation, lumbago, drooling, strabismus, presbyopia, tracheal

Fig. 3. 100 often used biomedical terms.

First, it is necessary to input the word to search in MeSH and then in LSD, in order to investigate the thesaurus word. The results of an investigation of MeSH and LSD are summarized in a table. We examine the difference between the two, and develop a database that organizes MeSH and LSD.

The comparison results of the number of thesaurus terms between MeSH and LSD are presented in Tables 1 and 2.

For example, for the thesaurus word “influenza,” 69 words were discovered in MeSH and 38 were discovered in LSD. Twenty-five thesaurus words were discovered in both MeSH and LSD. From that, a total of 82 thesaurus words were found.

The results for 100 often used biomedical terms are presented in Table 2. Here, 1210 words were discovered by MeSH, and 651 were discovered by LSD. A total of 446 thesaurus words were in both MeSH and LSD. From that, a total of 1416 thesaurus words were discovered.

Table 1. Part of comparison result of the number of thesaurus terms

	MeSH	LSD	common	only MeSH	only LSD	total
influenza	69	38	25	44	13	82
atopic	24	4	3	21	1	25
rhinits	7	7	7	0	0	7
asthma	5	10	4	1	6	11
antidote	10	1	1	9	0	10
antipyrestic	3	2	1	2	1	4
antineoplastic agent	107	6	6	101	0	107
respiratory distress	25	4	3	22	1	26
hypnotic	13	1	1	12	0	13
fomentation	2	0	0	2	0	2
esophagitis	6	3	3	3	0	6
myocardial	32	28	23	9	5	37
heart failure	3	3	3	0	0	3
urticaria	7	5	5	2	0	7
tranquilizer	16	3	3	13	0	16
enteritis	11	18	10	1	8	19
gout	4	9	2	2	7	11
epilepsy	49	22	21	28	1	50
diabetes mellitus	11	7	7	4	0	11
cerebral infaction	3	2	2	1	0	3
cerebral hemorrhage	14	3	3	11	0	14
pneumonia	33	45	24	9	21	54
lung cancer	11	3	1	10	2	13
obesity	17	7	6	11	1	18
peritonitis	7	6	5	2	1	8

Table 2. Comparison of results with MeSH and with LSD for 100 biomedical terms. The vertical axis is the number of words in the thesaurus.

	MeSH	LSD	common	only MeSH	only LSD	total
total	1210	651	446	764	206	1416

The output result of a new literature search system

The literature search system was constructed using the flow depicted in Fig. 1. An example of the output screen when searching for about “influenza” is presented in Fig. 4.

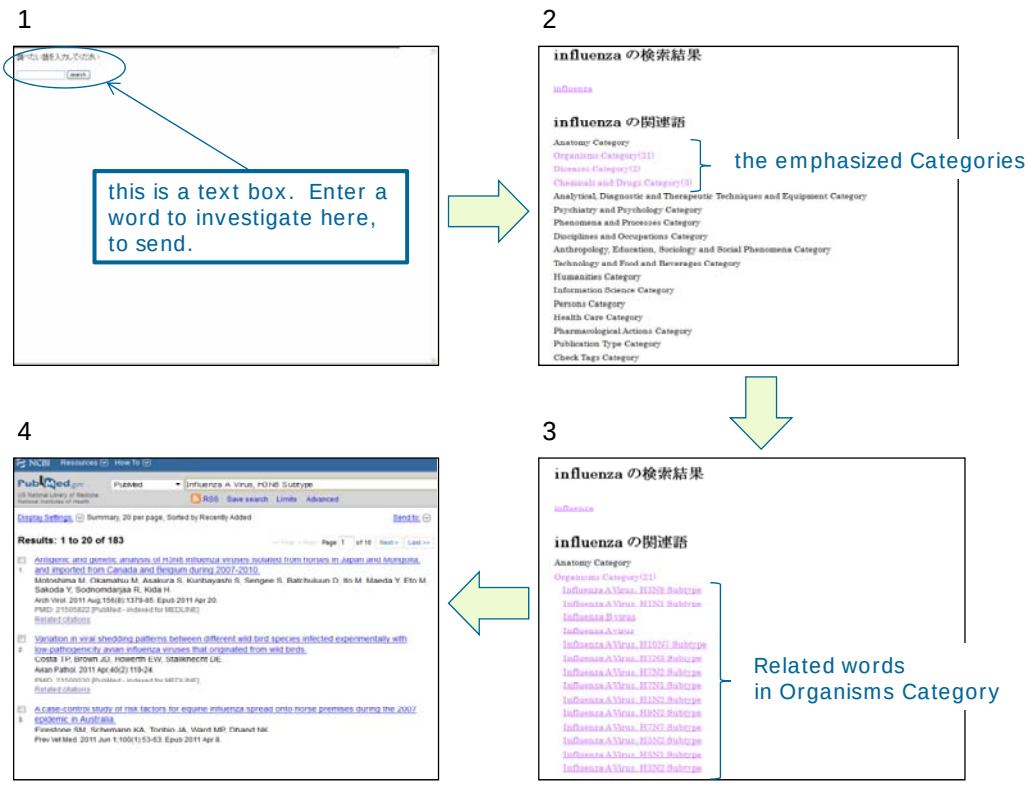


Fig. 4. Example of output screen when searching the literature for “influenza”.
1. A user inputs the word that he wants to investigate into the text box.
2. The category in which words related the input word exist is emphasized.
3. When the user selects the category, related words appear.
4. When the user selects the word, the output screen showing related research literature appears.

Experiment and Consideration

Using the proposed system, even the user who does not have a keyword can explore related literatures (Fig. 1). We experimented to determine how much we could explore the literature using this system, and compared it to the traditional system. Table 3 presents the comparison results regarding the number of studies.

Using “influenza” as an example, the number of works that could be found using the traditional system was 64574; using the proposed system, it was 99695. Therefore, 35121 more works have been extracted in this system. For example, the term “influenza” itself, as well as other words, is included in the thesaurus database. Therefore, 35121 works related to “influenza” were found without using the word “influenza.”

Table 4 presents the total results of a comparison of this system and the traditional system. The number of works that could be found using the traditional system was 6433205; however, 17160056 could be found using the proposed system. Therefore, 10726851 works were related to the keyword without using the actual keyword.

Table 3. Part of comparison result of the number of works

	traditional system	this system	difference
influenza	64574	99695	35121
atopic	27052	195136	168084
rhinitis	29637	33242	3605
asthma	123071	124913	1842
antidote	50857	429965	379108
antipyretic	50072	71479	21407
antineoplastic agent	765675	2779913	2014238
respiratory distress	36339	162433	126094
hypnotic	105120	246020	140900
fermentation	38149	38149	0
esophagitis	13865	17124	3259
myocardial	416445	639037	222592
heart failure	117470	117470	0
urticaria	16470	16524	54
tranquilizer	194598	360106	165508
enteritis	15321	64944	49623
gout	11908	65370	53462
epilepsy	129099	136589	7490
diabetes mellitus	257042	300995	43953
cerebral infaction	21911	69143	47232
cerebral hemorrhage	26682	239491	212809
pneumonia	102302	203281	100979
lung cancer	73839	166639	92800
obesity	153752	226623	72871
peritonitis	31455	34628	3173

Table 4. Comparison of the proposed system and the the traditional system.

	traditional system	this system	difference
total	6,433,205	17,160,056	10,726,851

Conclusion

This paper proposed a literature search technique utilizing thesauruses. By using a database that organizes MeSH and LSD, this system can search the literature related to a search keyword without including the actual keyword. Also, this system aims to increase user convenience by including a category search. With this system, related words are easy to discover by outputting the category of the biomedical term. We experimented to determine how much we could explore the literature in this system and compared it to the traditional system. The results indicated that this system was able to find more literature than the traditional system. We hope the proposed technique can be used for interdisciplinary research such as affective computing.

References

[1] MeSH: http://www.ncbi.nlm.nih.gov/mesh

[2] LSD: http://lsd.pharm.kyoto-u.ac.jp/ja/index.html

[3] Pharmaceutical Society of Tokai 4 prefectures: http://topnet.gr.jp/index.html

[4] Syuji Kaneko, Nobuyuki Fujita: evaluation of the bilingual thesaurus based on the analysis of literature information, Medical informatics 2006

Alternative Methodology of Complex Social System: Determining the Level of Agency and its Relations

Bogart YailMárquez, José Sergio Magdaleno-Palencia
Miguel López, Arnulfo Alanis Garza

Keywords: Complex Social Systems, Methodology, Data Mining, Neuro-Fuzzy, Distributed Agencies, Levels, Multi-agents.

Abstract. The objective of this study is to develop an alternative methodology using different mathematical and computational theories that are not conventionally used in the social sciencesAs the main purpose determining the level of agency and its relations of a complex social system using different theories.

1. Introduction

Social simulation consists in generating artificial social worlds with the capacity to produce results similar or approximate to those found in the real world. These worlds can be controlled, reused, and altered, abilities not easily found in mathematical models [1]. Social simulation is a wide and multidisciplinary field that is still growing. As simulation has been broadening its horizons in the research of traditional systems, there are a series of epistemological issues that have yet to be resolved [2].

On the other hand, the study of the social sciences faces limitations that go beyond their information processing capabilities. Due to this, representations have been simplified into deterministic or stochastic models. The use of computer agents can provide researchers with techniques and methodologies that would allow constructing and supporting more detailed and complex theories [3].

Agent technology consists of computer programs which are used to represent social actors, be it people, organizations, or companies. These programs are designed to react to their environment, where the environment is a model of the real world conditions in which the social actors operate. The use of agent technology in several scientific fields has gained acceptance in recent years and as such, it has started being implemented as a simulation technique in the social sciences [4].

A fundamental characteristic of agent based models is the ability of agents to interact, that is to say, they are able to transmit informative messages to other agents and are able to act based on what is learned from received messages. Messages can represent a spoken dialog between people or other indirect forms of communication, such as observation of other agents or detecting the actions of other agents. The ability of modeling the agent and the interactions in can have with other agents is one of the main differences between agent based models and other types of computational models [4].

The use of this technique becomes increasingly complicated as the number of agents increases. Even though it has been mentioned in the multi-agent community the need to develop and implement new methodologies, surprisingly, little has been proposed, leaving many science fields without viable options on implementing this resource.

The methodology deals with the ways reality and knowledge can be studied, not concerning itself with what has already been accepted as truth by the scientific community but only seeking strategies that will expand on the knowledge [5]. This work is motivated by the need to establish a

methodology that deals with the study of complex social systems. Unlike previews works, this methodology intends to cover adaptive and emergent matters, situations in which conventional analysis does not satisfactorily describe the complexity of real social phenomena and social actors. The described methodology generally involves the use of various computational techniques and interdisciplinary theories. The growing consensus must be able to describe every aspect of social life, as well as serve as a common language in which different theories can be contrasted.

The development of a methodology for complex systems can involve different techniques, theories, and points of view. The exclusive use of a single statistical methodology in the social sciences is insufficient for analyzing a complex system due to the great number of variables used in measuring.

The use of statistics in the social sciences is widely used and has great precision when using a large number of variables; however, it is only valid when analyzing at a single level and the majority of social problems encompass a complex system. It is difficult to generalize a methodology to span every complex system, not only for the nearly infinite number of variables that can be used for measurement, but also for the non-linear dynamics, the emergent properties, the self-organizing properties, and the multi-level, multi-dimensional interactions. Due to this nature of complex systems, it is necessary to analyze them using a multi-level focus..

2. Complexity

Existing articles mention the ambiguous definition of complexity as being a problematic aspect yet to be resolved[6], [7]. Warren Weayer, in his article science and complexity [8], tries to address the issue of defining from a simplicity problem perspective. He mentions that in the past, the physical sciences have learned to simplify variables, which lead to the creation of many technologies that are known today. However, in the case of biology and medicine, the analysis of problems differs due to the difficulty of finding any constants. Living beings are more prone to displaying highly interconnected situations and which important quantities are not quantitative. For these reasons, biological and medical problems frequently involve the consideration of a more complexly organized set [8]. Other authors such as ChristophAdami[6] state that definitions of complexity exist in dynamic systems and for biological organisms, but have conceptual or practical disadvantages where mathematical models are needed in order to demonstrate these theories.

The definition given by Heylighen[7] of what is complexity is extracted from the Latin root complexo, which roughly translated means entangled, intertwined, to encompass. This can be interpreted as, in order to have something complex, it is necessary for two or more components that are connected in a manner that is difficult to separate, to produce emergent properties. A system becomes increasingly complex with the growing number of distinct components, states, or aspects; and the number of relations or connections. A problem with this definition is that it does not give a single number or degree that would allow the objective measurement of the complexity in a phenomenon. The reason for this is that the number of distinct components/states/aspects and their connections are not objectively given in easily quantifiable entities as they exist at different levels, dimensions, and types. Despite this limitation, this definition holds favorable characteristics such as being simple and intuitive [7]. Therefore, the first step in this work's proposed methodology is to determine the different levels and relationships found in the complex system that is to be modeled.

A complex adaptive system (CAS) is situated in an environment which always has a greater degree of complexity than the very system it holds and therefore, can never be entirely predictable to the system. However, the system depends on whatever constancy the environment can offer in order to maintain a steady supply of energy to maintain its structures and internal processes. It this way the system can regulate what can enter or be extracted from the environment in a consistent way and not be severely affected by random aspects that may appear [9]. Another definition for a CAS consists on the non-homogeneous interactions of adaptive agents, adaptive being defined as having the ability to learn [10]. For the author Brownlee, a CAS is a dynamic network formed by many agents that can represent cells, species, individuals, corporations, or nations acting in parallel, continuously, and reacting to the actions taken by other agents. Controlling a CAS tends to be

highly decentralized, if there is any coherent behavior in the system, it grows based on competition and cooperation among agents. The final result of the system is given by the enormous amount of decisions made at some point by the many individual agents [11]. The study of complex adaptive systems, which are a subgroup of dynamic non-linear systems, has become an important focus point in interdisciplinary research such as in the social and natural sciences [12].

The study of CAS is the study of high levels of natural abstractions and artificial systems that are generally implementable in traditional analysis techniques. Microscopic patterns emerge from non-linear dynamics and the lower-level (microscopic) system interactions of adaptive agents. The emerging patters are more than the sum of the existing parts; therefore, the traditional non-reductionist methodology describes how the microscopic patterns emerge. On the contrary, holistic and totalistic methodologies research the application of motions that make use of simple rules and adaptive agent interactions that lead to emergence from a bottom-up point of view .

Generally, CAS examples have been extracted mainly from systems studied in biology, sociology, and economy. Frequently cited examples include: embryo development, the adaptive immune system, ecology, genetic evolution, reasoning and learning of the brain, meteorological systems, market economies, commerce systems, social systems, cultures, politics, traffic systems, insect swarms, flock of birds, the application of new ideas, scientific theory testing, and the resistance of bacteria to antibiotics. Computer simulation models play an important role in CAS research, where the system is reduced to its simplest essential components. These same simulation models exhibit aspects of complex adaptive systems and so provide fertile ground for controlled experimentation. A few modeling methods developed and used for this purpose include cellular automata (CA), agent-based models (ABM), artificial neural networks (ANN), genetic algorithms (GA), and learning classification systems (LCS) [11].

3. Modeling a complex social system

When referring to social systems, certain basic characteristics in the organizations must be present. One of these characteristics is that the consequents of social systems are probabilistic and non-deterministic. Moreover, human behavior can never be entirely predictable as people are complex beings; consequently, management cannot wait for consumers, providers, regulatory agencies and others to have a predictable behavior [13].

Organizations are seen as systems within systems. Said systems are complex, producing a whole that cannot be understood by analyzing the individual parts. The organization should be studies as a system characterized by all the essential properties for any social system[14].

Component independence: A change in one of the parts of the system will have an effect on the others. The internal and external interactions of the system reflect the different levels of control and autonomy.

Homeostasis or firm state: The organization can achieve a firm state only when two requirements are met, unidirectionality and progress. Unidirectionality means that regardless of changes in the organization, the same established results or conditions are met. Progress leading to the desired outcome is a process degree found within the limits established as tolerable.

Borders or limits: It is the line that defines what is within and what is outside the system, it need not be physical. A border consists in a closed line around selected variables among those that possess greater exchange (of energy, information) with the system.

Morphogenesis: the organizational system, different from other mechanical systems and even from biological systems, has the capacity to modify its basic structures. This is identified by Buckley [15] as its main distinctive characteristic.

On the part of social behavior, it tends to conform to the members of the group, producing cooperation and self-organization [16]. An organization, according to Ross Ashby [17], spans a multitude of definitions due to its corresponding undertakings, in relation to computation and the brain, and it is becoming increasingly important. The central concept is to provide “conditionality”.

In this manner, organizational theory is a co-existing part of functional theory of more than one variable.

Contrasting with the "conditional" is the "non-conditional", therefore the opposite of the "organization" must be, as the mathematical theory shows, the concept of "reduction potential" (also called "separability"). This occurs, in mathematical form, when what appears to be a multi-variable function is demonstrated, on closer inspection, that the parts' actions are not conditioned by the value of other parts. In mechanical form (hardware), it presents itself when what appears to be a test machine composed of two or more sub-machines, each sub-machine acts independent of each other. The matter of "conditionality" and its inverse "reducibility" can be approached by a number of mathematical and logical methods. In its generality and ideal applicability to complex behavior lies the fact that it is applicable to anything defined in the states. Its application does not require linearity, continuity, metrics, or a specified order. In this calculation, the degree of conditionality can be measured and analyzed; it is distributed between the factors and interactions in a way that is parallel to Fisher's method in variance analysis but without needing metrics in the variables, only the frequency in which the many combinations occur. Just as Fisher's conception in variance analysis outputs complex relations that can exist between the variations in a metric, McGIll and Garner's conception of uncertainty analysis provide a better understanding on dealing with the complexities of the relations when the variables are not measurable [17].

In a social system there exists the questions of what are the behaviors of individuals in agglomerations (cities, groups or networks) and why do people show such behavior. In a published study by Helbing[18], these types of behaviors are explained. It is stated that cooperation is crucial to society since it permits the creation of common benefits that one isolated individual could establish on their own. These common benefits are shared infrastructures to the social institutions. Notwithstanding, the contribution of public goods establishes a dilemma, as these benefits are shared by everyone, there exists the temptation to do as little as possible and reap the most rewards. In consequence, it has brought about the pollution and exploitation of nature and the problem of sustainable social benefits [18].

4. Proposed Methodology

The proposed methodology intends to be an alternative to describing social phenomena models as close to reality as possible. This is achieved by utilizing different mathematical-computational theories, which are not conventionally used in the social sciences and a new approach for the creation of computer simulation architectures. Furthermore, the integration of models in terms of distributed agencies provides an integral view and new ideas can be added to the singular focuses of multi-agent systems.

This research is oriented to the study of elemental hues related to the integration of groups, defining relationships, actions, and norms that are generated in the real world; proposing an agent based model that represents individuals, a methodology for the integration of groups, and the possibility of representing a social group as a single agent at a superior level. A viable methodology for complex social simulation must be constituted by the specific needs of the social problem that is to be modeled [19]. Nevertheless, every social problem adheres to certain complex system characteristics:.

5. Determining the level of agency and its relations

Mitchelle and Newman mention that a "complex systems" is a group or organization that is constituted by the interactions of many parts. Examples of complex systems include global climate, economics and the immune system. In those systems, the individual parts called "components" or "agents" and their interactions often lead to behavior that is not easily predictable based solely on the individual agents' behavior [20]. This is why it is crucial not to lose sight of the relations between the components in the same or distinct levels.

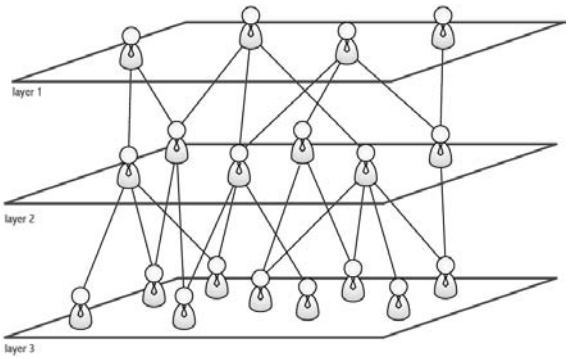


Figure 1. Levels agents represented in layers

5.1 Determining the levels

In traditional approaches of multi-agent systems and utility maximizing, an actor chooses the best alternative given the set of possibilities that is found in each level. The mean difference with this works proposed approach is that the phase space includes the transformations made by an upper level. Furthermore, an agent is composed of subcomponents of lower levels that can possess their own agencies. It is the agent's responsibility to present its subcomponents with individual phase space with optimal solutions that are acceptable to the upper-level agent to whom it belongs to. That is to say, the subcomponent agents optimize the phase space in which they are located, while the upper level must consider the manipulation of this set of possibilities that will make the desired global behavior. In this sense, if a company is taken as an agent, this level is composed of the subdivisions that form the company—which are directed by a group of people. The company is also located in an upper level that includes all the possible norms for the industry, and this level is encapsulated in a specific society.

The methodology to be implemented represents a new approach for the creation of a simulation architecture. It represents a general theory of the collective conduct and the formation of structures. The approach of Distributed Agencies (DA) treats agents as something that can be agent-like, as opposed to a traditional approach in which entities are either agents or not.

In this frame, agents are considered as intermediate, only possessing a certain degree of agency that is taken into account by the upper-level agent to whom it belongs to, and it is obtained from the lower level agents that compose it.

The proposed methodology redefines the levels of agency in two ways, the being that there are no obvious agents.

The distributed agency language expresses the observed behavior as the result of agents that maximize their objective functions; taking the concept of representation (AR) as a descriptive reference position in which all behavior is the result of the optimization of individuals, intertwined and potentially spanning multiple dimensions. An AR describes how a system was developed, how any agent makes decisions in an environment that implicitly considers the future effects of its own actions, family members and other members of the group.

5.2 Determining relations

In traditional approaches of multi-agent systems and utility maximizing, an actor chooses the best alternative given the set of possibilities that is found in each level. The mean difference with this works proposed approach is that the phase space includes the transformations made by an upper level. Furthermore, an agent is composed of subcomponents of lower levels that can possess their own agencies. It is the agent's responsibility to present its subcomponents with individual phase space with optimal solutions that are acceptable to the upper-level agent to whom it belongs to. That is to say, the subcomponent agents optimize the phase space in which they are located, while the

upper level must consider the manipulation of this set of possibilities that will make the desired global behavior. In this sense, if a company is taken as an agent, this level is composed of the subdivisions that form the company—which are directed by a group of people. The company is also located in an upper level that includes all the possible norms for the industry, and this level is encapsulated in a specific society.

The methodology to be implemented represents a new approach for the creation of a simulation architecture. It represents a general theory of the collective conduct and the formation of structures. The approach of Distributed Agencies (DA) treats agents as something that can be agent-like, as opposed to a traditional approach in which entities are either agents or not.

In this frame, agents are considered as intermediate, only possessing a certain degree of agency that is taken into account by the upper-level agent to whom it belongs to, and it is obtained from the lower level agents that compose it.

The proposed methodology redefines the levels of agency in two ways, the being that there are no obvious agents.

The distributed agency language expresses the observed behavior as the result of agents that maximize their objective functions; taking the concept of representation (AR) as a descriptive reference position in which all behavior is the result of the optimization of individuals, intertwined and potentially spanning multiple dimensions. An AR describes how a system was developed, how any agent makes decisions in an environment that implicitly considers the future effects of its own actions, family members and other members of the group.

6. Conclusions

Our intention is to create a system that is composed of agents where each agent represents anlevel whose adaptation is the result of complex interactions in nonlinear dynamics, emergent phenomena which arise in the system, and to compare reality with the artificial system and observe the properties, processes and relationships by using different computational methods.The methodology is developed in a holistic manner with an analysis of the different levels in dimension and time, it attempts to apply the dynamics of non-linearity, emergent properties, self-organization processes, and the interaction between multiple levels and dimensions. The field is interdisciplinary and this work's main objective is to find a connection between the different social theories and computational theories related to the science of complexity; the result of the methodology provides a powerful alternative to complement, substitute, and/or widen the traditional approaches in the social sciences from the perspective of a complex system when developing the different proposed techniques.

References

[1] Gershenson, C., Philosophical Ideas on the Simulation of Social Behaviour. Journal of Artificial Societies and Social Simulation, 2002. 5.

[2] David, N., J.C. Caldas, and H. Coelho, Epistemological Perspectives on Simulation III. Journal of Artificial Societies and Social Simulation, 2010. 13(1): p. 14.

[3] Davidsson, P., Agent Based Social Simulation: A Computer Science View Journal of Artificial Societies and Social Simulation 2002. 5.

[4] Gilbert, N., Computational social science: Agent-based social simulation, in Agent-based modelling and simulation, D. Phan and F. Amblard, Editors. 2007, Bardwell: Oxford. p. 115-134.

[5] Márquez, B.Y., et al., Methodology for the Modeling of Complex Social System Using Neuro-Fuzzy and Distributed Agencies. Journal of Selected Areas in Software Engineering (JSSE), 2011.

[6] Adami, C., What is complexity? BioEssays, 2002. 24(12): p. 1085–1094.

[7] Heylighen, F., Five Questions on Complexity. 2008.

[8] Weaver, W., Science And Complexity, in "Science and Complexity", American Scientist. 1948, 1948: New York City.

[9] Jost, J., External and internal complexity of complex adaptive systems in Theory in Biosciences. 2004, Springer Berlin / Heidelberg: Berlin. p. 69-88.

[10] Ahmed, E., A.S. Elgazzar, and A.S. Hegazi, An Overview of Complex Adaptive Systems. Mansoura, 2005.

[11] Brownlee, J., Complex Adaptive Systems. 2007, Complex Intelligent Systems Laboratory, Centre for Information Technology Research, Faculty of Information Communication Technology, Swinburne University of Technology: Melbourne, Australia.

[12] Lansing, S. (2003) Complex Adaptive Systems. DOI: 10.1146.

[13] Suarez, E.D., A. Rodríguez-Díaz, and M. Castañón-Puga, Fuzzy Agents, in Soft Computing for Hybrid Intelligent Systems, O. Castillo, et al., Editors. 2007, Springer: Berlin. p. 269-293.

[14] Yolles, M., Organizations as Complex Systems: An Introduction to Knowledge Cybernetics, in Managing the Complex. 2006, Information Age Publishing: Greenwich, Connecticut, USA. p. 866.

[15] Boulding, K., General Systems Theory The Skeleton of Science Management Science, 1956. 6(3): p. 127-139.

[16] Jaffe, K. and L. Zaballa, Co-Operative Punishment Cements Social Cohesion. Journal of Artificial Societies and Social Simulation, 2010. 13: p. 4.

[17] Ashby, R., Principles of the self-organizing system. E:CO Special Double Issue, 2004. 6: p. 102-126.

[18] Helbing, D., W. Yu, and H. Rauhut, Self-organization and emergence in social systems. Modeling the coevolution of social environments and cooperative behavior. 2009.

[19] Cioffi-Revilla, C., A Methodology for Complex Social Simulations. Journal of Artificial Societies and Social Simulation, 2010. 13(1): p. 7.

[20] Mitchell, M. and M. Newman, Complex Systems Theory and Evolution. In Encyclopedia of Evolution (M. Pagel, editor), 2002.

Estimate the Intrinsic Dimension of a Metric Space Using the Eigenvalues of the Pair-wise Distance Matrix

Xi Liu, Houjun Tang, Zhao Jiang, Pang Yue, Ye Cai,
Haijun Lei, Hong Zhou, Rui Mao[1]

National High Performance Computing Center at Shenzhen

College of Computer Science and Software Engineering, Shenzhen University

3688 Nanhai Road, Shenzhen, 518060, China

xii.liu@hotmail.com, houj.tang@gmail.com,

{2110230103, 2110230101}@mail.szu.edu.cn, {caiye, lhj, hzhou, mao}@szu.edu.cn

Keywords: Similarity search, metric space indexing, intrinsic dimension, eigenvalue, pair-wise matrix.

Abstract. One of the important properties of a metric space is the intrinsic dimension, which relies solely on the given space. The intrinsic dimension is a key factor in metric space indexing for nearest-neighbor search and range search. Therefore, there has been several studies of how to estimate it accurately and effectively. In this paper, we propose a simple and effective method to estimate the intrinsic dimension of a metric space, using the eigenvalues of the pair-wise distance matrix of the metric space. Three criteria are compared to find the best one that can most accurately determine the intrinsic dimension. Empirical results and comparison with other method show that this method can be used to reliably measure the intrinsic dimension of a metric space.

Introduction

Metric space indexing, or distance-based indexing, is a technique used for similarity search in complicated unstructured data. It performs on the model with a distance function which obeys the triangle inequality, while the object set is called a metric space [1]. The characteristic of a metric space will influence greatly in the performance of similarity searching. Therefore, the research for the properties of metric space has been taken in order to improve the process of indexing.

There is an interesting property of the metric space, the intrinsic dimension, which may be possible to enhance the performance of the similarity search [6].

The intrinsic dimensions, which are the dimensions that rely only on specific metric space, instead of the outer space that contains the given space, are important for metric space. For it shows the indexability of a set of data. Further, it could be a guidance in the process of nearest-neighbor search or range query search in order to get a better performance.

The intrinsic dimension influences the difficulty of the indexing. Therefore, the estimation of the intrinsic dimension has been discussed widely. There're many related work to it, and several methods has been proposed to do the estimation.

The following are some existing methods.

Method 1: Chavez et al. [1] use the formula $\rho = \mu^2 / 2\sigma^2$ to estimate the intrinsic dimension. Here, μ and σ^2 represent the mean and variance of the pair-wise distances respectively.

[1]The correspondence author

Method 2: Mao et al. [8] seek the relationship between the number of points in a hyperball and the radius of it. They propose that slope coefficient could be used to estimate the intrinsic dimension. The slope coefficient is calculated by linear regression on the logarithm of the range query radius and the average number of range query results of a hyperball.

Method 3: Mao et al. [9] use another method to determine the intrinsic dimension and select pivot in the pivot space model, the PCA method. Firstly, the PCA method calculates the eigenvalues of the covariance matrix of the original matrix. Secondly, the method determines the intrinsic dimension by comparing the value of those eigenvalues which called PC (Principal Component).

Method 1 is easy to execute. Method 2, which in fact is a variation of the Box Dimension, is limited by r and the distribution of the data. Method 3 is an application of the PCA algorithm, which Mao et al. [9] also used in dimension reduction for distance-based indexing. Method 3 is easy to understand and apply, and more accurate than the first two methods [9].

Compared to the PCA method, our method needs less calculation and the yields similar accuracy in the result. Our method only needs to calculate the eigenvalues of the original distance matrix, instead of the covariance matrix of it.

As with the criteria for determine the intrinsic dimension according to eigenvalues, we propose and compare three methods. The absolute values (λ_i) of all the eigenvalues of the pair-wise distance matrix are listed in descending order. For method 1, we find the biggest ratio of $\lambda_i / \lambda_{i+1}$, and it will be the intrinsic dimension. As for method 2, the difference between λ_i and λ_{i+1} will be counted, to determine the intrinsic dimension. Finally, for method 3, the accumulative percentage is defined, and we will decide the intrinsic dimension according to the accumulative percentage.

Eigenvalues for Estimating Intrinsic Dimension

We now introduce a fourth method to estimate the intrinsic dimension based on the eigenvalues of the pair-wise distance matrix.

The description of method 4 is shown below:

Method 4: Let M be the pair-wise distance matrix derived from the metric space. Let $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}$ be the eigenvalues of the pair-wise distance matrix, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$.

```
EigenvaluesCalculation()
{
    //get data from the I/O interface
    1.   data = getData(input file);

    //generate the distance matrix
    2.   matrix = countPair-wiseMatrix(data);

    //calculate eigenvalues of the matrix
    3.   EigenSET =countEigenvalues(matrix);

    //convert the eigenvalues into absolute value
    4.   for each eigenvalues ∈ EigenSET
         Eigen = |Eigen|;

    //output the eigenvalues in descending order
    5.   sort(EigenSET);
         output(EigenSET);
}
```

Fig. 1. Algorithm for eigenvalues calculation

The figure 1 shown above is the algorithm of calculating the eigenvalues of a pair-wise distance matrix. After the program output the absolute eigenvalues in descending order, we use 3 criteria to determine the intrinsic dimension of the data.

Note that we have convert the eigenvalues into absolute value, so they're all positive numbers. And because that the first eigenvalue is much bigger than the others, 50% of the sum of all the values, therefore, we ignore it in the comparison while add it in the final result, so that it won't affect the comparison. Here we propose three criteria to estimate the intrinsic dimension: (1) Find the max ratio between λ_i and λ_{i+1} , that is, $\hat{d} = \arg\max_i (\lambda_i / \lambda_{i+1}),\ \ i=2,\ 3,\ ...\ n\text{-}1$. (2) Find the maximum difference between λ_i and λ_{i+1}, that is, $\hat{d} = \arg\max_i (\lambda_i - \lambda_{i+1}),\ i=2,\ 3,\ ...\ n\text{-}1$. (3) Calculate the percentage of each eigenvalue, $p_i = \lambda_i / sum(\lambda) \times 100\%$, $i=2,\ 3,\ ...\ n\text{-}1$. Find the last i which fit the formula $\sum_{j=1}^{i} p_j \geq 70\%,\ i=2,\ 3\ ,\ ...\ n\text{-}1$.

The criterion (1) indicates that the eigenvalues decrease the most from λ_i to λ_{i+1}. Criterion (2) guarantees that the difference between the eigenvalues λ_i and λ_{i+1} is the biggest. Criterion (3) ensures that the sum of the selected dimension will be over a certain percentage.

Empirical results

Here we use the empirical result to illustrate the performance of the eigenvalue method.

Organization of Empirical Study. We evaluate our method in the Molecular Biological Information System(MoBIoS), which is a java-based library developed by University of Texas, Austin, utilizing metric space indexing techniques to perform faster retrieval of sequences and mass spectra signatures data [7].

The test suite we use is similar to the one Mao et al. [9] used to test the PCA method. It consists of real vector data, biological data, an image dataset, and synthetic vector data [7]. The real vector data consists of the US boundary of the states Hawaii and Texas. For the biological data, analytically determined peptide fragmentation spectra of human and E. coli proteins with a pseudo-semi-metric cosine distance is involved. The image dataset consists of images represented by 66 dimensional feature vectors with a linear combination of L1 and L2 norms. The synthetic vector consists of data of uniform distributions. The suite is summarized in Table 1.

Table 1. Summary of test suite

Workload	Total size	Distance oracle	Domain dimension
Uniform Vector	1M	L^1, L^2, L^∞ norm	1-20
Hawaii	9k		2
Texas	190k		2
Protein	100k	Weighted edit distance	6
Image	10221	L-norms	66

Comparison of dimension estimation methods. We compare the three criteria with eigenvalues and Mao et al.'s PCA-base method [9]. The eigenvalues of the distance matrix are calculated and output during the index building process, and then we estimate the intrinsic dimensions according to the eigenvalues using the three criteria and compare the result with that of Mao's PCA-base method. We only compare the intrinsic dimensions estimated by the two methods, and the cost of constructing time is ignored. When possible, we change the domain dimension to see the variation of the estimates. The estimates are shown in Table 2.

In order to make a detailed comparison, we show the result of the 3 criteria for uniform vector from dimension 1 – 10 in the Table 3. Note that here in Table 3, we show the first biggest eigenvalue, but we ignore it when using the three criteria.

Table 2. Estimate of intrinsic dimension

Workload	Domain dimension	Intrinsic dimension			
		$\lambda_i / \lambda_{i+1}$	$\lambda_i - \lambda_{i+1}$	percentage sum	Mao's method
Uniform vector	1	2	2	2	2
	2	3	3	3	3
	3	4	4	4	4
	4	5	5	5	5
	5	6	6	6	6
	6	7	7	7	7
	7	8	8	8	8
	8	9	9	9	9
	9	10	10	10	10
	10	11	11	11	11
Protein	q=6	7	7	7	7
Hawaii	2	3	2	3	3
Texas	2	3	2	4	2
Image	66	5	3	7	5

From the table 2 and table 3 we can see that for all the method in the table 2, the result for uniform vector and protein are the same. They all estimate the intrinsic dimension as d+1 (or q+1). And as for Hawaii, Texas and Image, criteria 1 is more constant, but sometimes the results (Table 3) are not so obvious. Criteria 2 although showing the results obviously, is not so constant for some kinds of data, for example, the image data. Criteria 3 is constant and obvious for uniform vector data, however, the number of the percentage is hard to define, and would vary from data to data. Therefore, for the 3 criteria, the first one is best.

Then we use the result of criteria 1 to compare with that of Mao et al.'s method. We find that the results of the two methods are alike both in accuracy and stability. Nevertheless, Mao et al.'s method has to calculate the covariance matrix of the original pair-wise distance matrix to get the eigenvalues, while our method does not. Instead, we get the eigenvalues directly from the original pair-wise distance matrix. Therefore, our method needs less computation than Mao et al.'s method.

Moreover, when analyzing the statistics of the eigenvalues of the pair-wise distance matrix (see Table 4), we also find some interesting trends. (1) The biggest eigenvalue, which is also the only negative eigenvalue, is always 50% for all the eigenvalues that have been calculated. (2) The ratio from the 2^{nd} to the $d+1^{th}$ eigenvalues are constantly 1, they are approximately the same in value. (3) In the statistics of the uniform vector, as the dimension increases, the differences between the eigenvalues (except the biggest one) are decreasing. However, the sum of the eigenvalues is increasing. (4) In the statistics of the uniform vector, as the dimension increases, we find that apart from the d+1 biggest eigenvalues, the ratio between the $d+2^{th}$ and $d+3^{th}$ eigenvalues is increasing significantly, and even become the max ratio when the dimension is bigger than 19. This trend means that the $d+2^{th}$ eigenvalues is becoming bigger and bigger. These trends may be helpful to the future work of applying eigenvalues to the metric space indexing.

Table 3. Results of eigenvalues for uniform vector

dim	1			2			3			4		
i	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	1.7	1158	0.50	3.5	3095	0.50	5.4	4412	0.50	7.3	5457	0.50
2	3.2	1126	0.79	1.0	7	0.64	1.0	11	0.59	1.0	13	0.56
3	2.8	332	0.88	3.7	876	0.78	1.0	11	0.68	1.0	3	0.63
4	1.6	73	0.91	1.2	70	0.81	3.9	726	0.77	1.0	13	0.70
5	1.6	41	0.93	1.5	90	0.84	1.5	87	0.79	4.1	633	0.76
6	1.4	18	0.94	1.7	69	0.86	1.0	2	0.80	1.8	90	0.78
7	1.3	12	0.95	1.0	1	0.87	1.0	3	0.82	1.0	1	0.79
dim	5			6			7			8		
i	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
1	9.1	6333	0.50	11	7123	0.50	12.8	7825	0.50	14.8	8482	0.50
2	1.0	12	0.55	1.0	3	0.54	1.0	6	0.53	1.0	6	0.53
3	1.0	10	0.60	1.0	11	0.59	1.0	11	0.57	1.0	8	0.56
4	1.0	11	0.66	1.0	8	0.63	1.0	4	0.61	1.0	6	0.60
5	1.0	18	0.71	1.0	14	0.67	1.0	12	0.65	1.0	1	0.63
6	4.1	549	0.76	1.0	17	0.72	1.0	4	0.68	1.0	11	0.66
7	2.1	93	0.77	4.1	494	0.76	1.0	18	0.72	1.0	3	0.69
8	1.0	2	0.78	2.5	94	0.77	4.2	458	0.76	1.0	17	0.72
9	1.0	0	0.78	1.0	0	0.77	2.7	90	0.76	4.3	429	0.76
10	1.0	0	0.79	1.0	0	0.78	1.0	0	0.77	3.0	87	0.76
dim	9 (dim from 1-5)			10 (dim from 1-5)			9 (dim from 6-10)			10 (dim from 6-10)		
i	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
1/6	16.6	9074	0.50	18.5	9643	0.50	1.0	4	0.64	1.0	0	0.63
2/7	1.0	5	0.53	1.0	4	0.52	1.0	9	0.67	1.0	5	0.65
3/8	1.0	7	0.56	1.0	5	0.55	1.0	16	0.70	1.0	8	0.68
4/9	1.0	4	0.58	1.0	2	0.58	1.0	1	0.73	1.0	16	0.70
5/10	1.0	5	0.61	1.0	9	0.60	4.3	404	0.75	1.0	5	0.73

(Notation: c1 for $\lambda_i / \lambda_{i+1}$, floored to one decimal place, c2 for $\lambda_i - \lambda_{i+1}$ floored to integers, c3 for percentage sum, floored to two decimal places.)

Conclusions and future work

There are many works related to the intrinsic dimension of a metric space, because it is an important property of a metric space and may help improve the performance of the similarity search over it. For the past work, Mao et al.'s PCA method may yields more consistent and stable performances [9]. We use the eigenvalues of the pair-wise distance matrix to estimate the intrinsic dimension. Compared to Mao et al.'s PCA method, our method yields a similar performance. Therefore, our method could be more consistent and stable than other works. Moreover, because our method uses the eigenvalues of the original pair-wise distance matrix, while Mao et al.'s PCA method uses the eigenvalues of the covariance matrix of the original pair-wise distance matrix. Our method is one step less than Mao et al.'s, indicating less calculation. In other words, our method yields similar accuracy to Mao et al.'s with less time cost.

The eigenvalue method is a powerful tool in linear algebra. It has many applications in physics and mathematics, and can be applied in areas of metric space indexing and yields a good performance. The PCA algorithm in pivot selection and dimension estimation is an example [9].

In this paper, we apply the eigenvalue method in the estimation of intrinsic dimension of metric space, and prove it to be accurate and simple. The result of this application shows that eigenvalues can be a useful tool in metric space indexing. And we show three trends concluded from the analysis of the eigenvalues for the pair-wise distance matrix. The next step of research should focus on the analysis of these four trends, and more work on the application of eigenvalues is expected.

Acknowledgements

This research was supported by the following grants: NSF-China: 61033009, 61003272, 61170076; China NSF-GD grant: 10351806001000000; a grant from the Computer Architecture Key Lab of Chinese Academy of Sciences: “Transportation and optimization of Hadoop and GeDBIT on Loongson based platforms”; Shenzhen Foundational Research Project: JC201005280408A, JC200903120046A; a grant from the Shenzhen-Hongkong Innovation Circle Project: ZYB200907060012A; a SZU Research Course Project: 0000132373.

References

[1] Bustos, B., G. Navarro, and E. Chavez, Pivot selection techniques for proximity searching in metric spaces. Pattern Recogn. Lett., 2003. 24(14): p. 2357-2366.

[2] Chavez, E., G. Navarro, R. Baeza-Yates, and J. Marroqu, Searching in metric spaces. ACM Computing Surveys, 2001. 33(3): p. 273-321.

[3] Strang, Gilbert (2006), Linear algebra and its applications, Thomson, Brooks/Cole, Belmont, CA, ISBN 0-030-10567-6 .

[4] Golub, Gene F.; van der Vorst, Henk A. (2000), "Eigenvalue computation in the 20th century", Journal of Computational and Applied Mathematics 123: 35–65, doi:10.1016/S0377-0427(00)00413-1 .

[5] Aldrich, John (2006), "Eigenvalue, eigenfunction, eigenvector, and related terms", in Jeff Miller (Editor), Earliest Known Uses of Some of the Words of Mathematics, http://jeff560.tripod.com/e.html, retrieved 2006-08-22

[6] Clarkson K.L., Nearest-neighbor searching and metric space dimensions, In the Nearest-Neighbor Methods for Learning and Vision: Theory and Practice, MIT Press, 2006, pp. 15-59.,

[7] MoBIoS test suite: http://aug.csres.utexas.edu/mobios-workload/

[8] Mao, R., W. Xu, S. Ramakrishnan, G. Nuckolls, and D.P. Miranker. On Optimizing Distance-Based Similarity Search for Biological Databases. in the 2005 IEEE Computational Systems Bioinformatics Conference (CSB 2005). 2005.

[9] Mao,R., Miranker, W. and Miranker, D.P., Dimension Reduction for Distance-Based Indexing, in the Proceedings of the Third International Conference on SImilarity Search and APplications (SISAP2010), page 25-32, Istanbul, Turkey, September 18 - 19, 2010.

Improved Usability of Object Structure and Error Location Analysis

Keiji Takiguchi[1, a] and Hayato Ohwada[2,b]

[1,2]2641, Yamazaki, Noda-City, Chiba, 278-8510, Japan

[a]takiguti@ohwada-lab.net, [b] ohwada@ia.noda.tus.ac.jp

Keywords: Usability improvement, Menu structure, Error location

Abstract. This paper proposes a usability improvement method that involves analysis of the object structure and the user's error location. Analysis of the object structure makes a usability test more effective. An improver tends to analyze only the visible location of a user interface for improving usability. However,a improver must not neglect invisible locations like the menu structure. By considering both object structure and error location, the improver can achieve a more practical and suitable interface. The usability of video recorder operation is assessed to determine the effectiveness of the proposed method.

1. Introduction

Recently, the spread of multifunctional and highly efficient digital equipment has been increasing. However, their operation has become more complicated with the increase number of functions. Thus, it is necessary to improve their usability.

In this study, the features of multifunctional equipment are classified into visible components (e.g., Graphic User Interface (GUI)) and invisible components (e.g., menu structure). In the improvement process, the latter tends to be ignored because problems are difficult to analyze. In order to improve both components, the focus of this study is "object structure" and error location". By conducting a usability test after analyzing the object structure, we can discover additional problems. Their Classification of visible and invisible components leads to more appropriate improvement.

The paper is organized as follows. Section 2 describes the usability improvement method we propose. Section 3 discusses usability assessment of HDD recorder operation to demonstrate the advantage of the proposed method. Section 4 provides conclusions.

2. The proposed method

This section explains the proposed usability improvement method preliminary analysis, usability test and classification and error analysis.

2.1 Preliminary Analysis

Analyzing an object structure is necessary for creating a task in a usability test. Object features include the input-and-output form and the menu structure

Various kinds of the input-and-output forms exist. For example, with a touch panel input and an output are performed by the same device. In contrast, with a television, input is performed by a remote control, and output is performed by the display. Thus, the features of input-and-output structures differ. For example, if the menu structures of HDD recorders with exactly the same

function differ, it is possible that the operation processes of recording or playback also differ. Of course, it is assumed that usability is affected by the difference in the system structure. Thus, it is very important to investigate the system structure [2].

Next, System structures are divided into large shallow structures and deep narrow structures (Fig. 1). The large shallow structure requires little operation of the target item, and there are many choices for one item. In contrast, although a deep narrow structure has few choices for each item, much operation is required for the target item. It is very important to consider the depth and width of a structure when building a system.

Considering the type of structure in the analysis of a task in a usability test can help an improver determine the practical operation and difficulty for each function

Fig. 1 Large shallow structure (left) and deep narrow structure (right)

2.2 Usability Test

In a usability test, the subject carries out the task that the improver has prepared. Such a test is necessary in determining the fault of the usability of an object [3].

Although the form of a usability test is not specified, we should perform a test that can find many causes which make it more difficult for the user to use. Therefore, we should be conscious of conducting a usability test in the environment.

Careful selection of the task required of a subject is also necessary. The task cannot be too easy or too difficult. Therefore, we must consider the menu structure, the required minimum number of operations, and the number of changes on a screen.

2.3 Classification and Analysis of Errors

In order to understand the location at which an object cannot be used easily and why, it is necessary to classify types of error.

First, we must determine the stage at which the user falters in the object operation. Next, we must determine the location of the user's error.

According to D. A. Norman, human behavior involves seven stages of action [1].

1. Forming the goal
2. Forming the intention
3. Specifying an action
4. Executing the action
5. Perceiving the state of the world
6. Interpreting the state of the world
7. Evaluating the outcome

For analysis of error location, the location at which a user makes an error in object operation is divided into a visible and the invisible components. Visible components are an input, an output, and

language expression. For example, for an HDD recorder, the remote control is the input and the display is the output. Error in language expression occurs when a user cannot understand the language in the manual or on the display. Menu structure determines the process of operation to achieve a user's purpose. Generally, the invisible location is neglected. Therefore, it is important that menu structure be considered in the analysis of error location.

We propose that the cause and location of an error could be understood in detail by using the classification table (Table 1)

Table1. Error classification checklist

Object Components | Seven Stage of Action	①	②	③	④	⑤	⑥	⑦	others
input								
output								
language expression								
system structure								

Because an error might occur in locations other than Norman's seven stages of an action, it may be necessary to add other items.

3. Application

This section describes the process involved in operating an HDD recorder using the proposed method.

3.1 Object Equipment

An HDD recorder SR-DVM700 made by Victor was used as the object.

The menu structure of the SR-DVM is presented in Fig. 2. Unlike a common HDD recorder, both a DVD and a miniDV can be used. A program is fundamentally selected using a cursor and a selection button: other operations are performed with other buttons.

The input (remote control) and output (television) are separated (Fig. 3). Such a separation has two disadvantages. The first is that the opportunity exists for the user's eyes to leave the output location with the input of information. Therefore, the user must alternately check the input location and the output location. The second disadvantage is that it is possible for the user to push a button that is not required for the target operation. Therefore, a user may not understand which button to push in some cases. A multifunctional HDD recorder has many remote control buttons, causing the user to become even more perplexed.

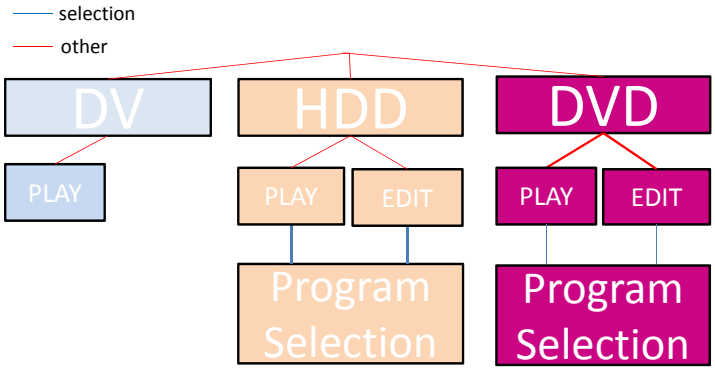


Fig. 2 Menu of the HDD recorder (SR-DVM700)

The quadrangle denotes the operation. The line indicates the button required for operation of a lower quadrangle. Blue lines denote the cursor and selection button. Red lines denote other buttons.

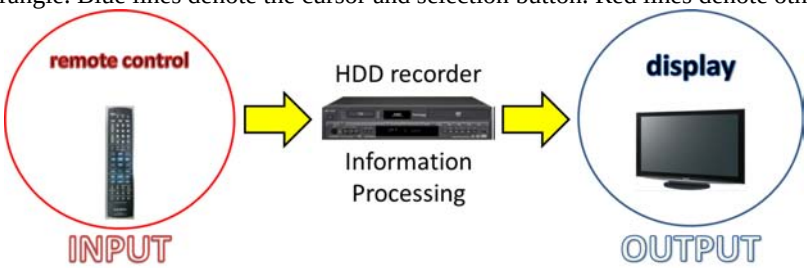


Fig. 3 Input and Output of the HDD recorder (SR-DVM700)

The information is input by the remote control, and the television perform the output (display) is,: thus, the input portion and the output portion are separated.

3.2 Usability Test

We conducted a usability test in which 20 students who had not used an HDD recorder operated an HDD recorder made by Victor (SR-DVM700).

The HDD recorders are operated consulting by two subjects sets in order to make the subjects easy to utter (Fig. 4). The subjects were stationed so that the mutual output screen of the equipment being operated was not visible to them. The subjects performed the tasks by, referring to an attached manual.

A television for image output and a digital camcorder for recording were used for the experiment. The subject performed the following three tasks. The number in parentheses is the number of operations required for task achievement.

1. Please play the specified program on HDD (6).
2. Please set the thumbnail of the specified program on the HDD as "0:30:00" and register "variety" as the genre (12).
3. Please combine the two specified programs on the HDD and dub the program to the DVD (43).

Fig. 4 Experiment setup

The subjects are stationed so that the mutual operation screen is not visible to them

3.3 Result and Analysis

The achievement ratio of tasks 1 and 2 was 95%. The achievement ratio of task 3 was 50%. The average required time was 204 sec for task1, 376 sec for task 2 and 1352 sec for task 3.

In accordance with the method of 3.2, the error classification results of the experiment were as presented in Table 2.

Table 2 Error classification result in the experiment

Object Components | Seven Stage of Action	①	②	③	④	⑤	⑥	⑦	others	Sum
input	0	0	0	3	0	0	0	2	5
output	0	0	0	3	0	0	15	6	24
language expression	1	0	4	0	0	0	0	9	14
system structure	1	0	21	0	0	0	0	0	22
sum	2	0	25	6	0	0	15	17	65

There were 65 errors, with the most errors (58%) in "3. Specifying an action and menu structure", and "others and language expression". The next most prevalent errors (15 errors) were in "7.Evaluating the outcome and an output".

3.4 Analysis and Directivity of Improvement

Many errors in "3.Specifying an action" were generated due to the complexity of the menu structure. The subject thought that operations like “playback” and “edit” were to be chosen after program selection (Fig. 5). Possible improvement might involve changing the input device to a touch-panel remote control form that displays only the required button.

The error for "7.Evaluating the outcome and output" occurred because there was no clear display (e.g., “complete") when a certain operation was finished. Therefore, although the subject had successfully performed the operation, he was perplexed. Since the subject was not able to understand the location on the menu structure, the error of "others and an output" occurred.

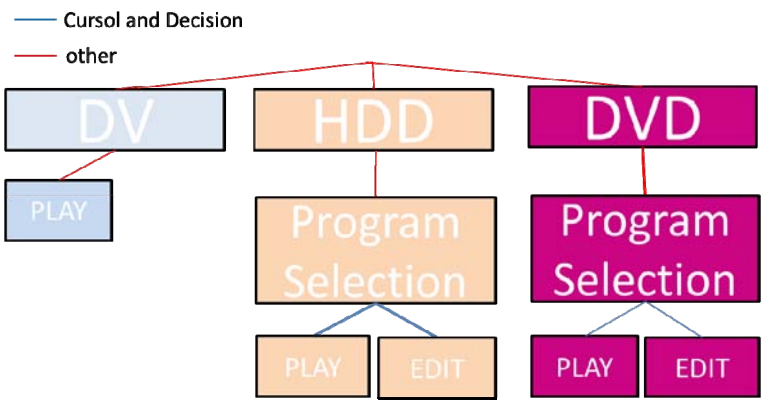


Fig. 5 Proposed menu structure

4. Conclusion

In this paper, we proposed a method for improve usability focusing on object features and error location in order to improve the usability of an object. We applied the method to an HDD recorder,

and demonstrated the directivity of the improvement. This method can also be applied to improve computing effectiveness by analyzing brain waves and sounds.

Acknowledgment

We would like to thank to the students who cooperated in the experiment.

References

[1] Norman, DA .:The Psychology of Everyday Things, Basic Books, New York(1988)

[2] Michael E. Rakauskas.: Using Utility Theory to Evaluate IVR Menu Structure and Reduce Driving Distraction, PROCEEDINGS of the Fourth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design,2007

[3] Torkil Clemmensen :Cultural Cognition in the Thinking-Aloud Method for Usability EvaluationTwenty Ninth International Conference on Information Systems, Paris,2008

Task Merger and Spanning Tree Based Grid Tasks Rescheduling*

Tingwei Chen[a], Jingsen Wang[b], Shanjie Zhou[c]

School of Information Engineering, Liaoning University, 110036, Shenyang, China

[a]twchen@lnu.edu.cn, [b]wangj353256@163.com, [c]zhou6586@163.com

Keywords: Tasks Rescheduling, Spanning Tree, Task Merger

Abstract. Grid resources are autonomic, distributed and their status change over time, those applications scheduling are not fixed, so grid tasks need rescheduled. Rescheduling will increase response time. To solve this question, in this paper, task merger and a spanning tree based grid tasks rescheduling approach is presented to improve response time. This approach focus on reducing the reduceing the number of tasks and chooseing minimal communication cost. Minimal tasks-load are merged for reduce reschedule probability, then generate a minimal spanning tree to improve response time. Experiment show that this approach could improve response time, hence overcome deficiency of previous algorithm.

1. Introduction

Grid computing has rapidly developed specifically for complex scientific calculations of distributed computing model on the Internet. In grid computing, a large-scale application is often broken down into multiple sub-tasks, each task is assigned to different resources, so task scheduling and response time problems not only have high research value but also has high practical value.

In the tasks are executing, resource degradation, resource exit and other events that could impact on the application complete, raising resource performance and new resources for improving application performance, so application need to response to the change of such resources, that trigger frequent rescheduling. Frequent rescheduling will increase the user's response time, the occurrence probability of rescheduling is closely related with the number of tasks. In the case of frequent rescheduling, if the user's response time reduce, the scale of tasks will be necessary to reduce. Considering reducing the scale of the tasks can only ensure that reducing the probability of rescheduling, it does not guarantee that this consideration can reduce the user's response time by executing the overall tasks. So, when the scale of the tasks is large, just rely on reducing the number of tasks to reduce user's response time is not enough, it also need to consider the overall cost of implementation of application.

As the number of task gradually reducing, to the contrary,communication between tasks will also increase the overall cost of the task. Each task occupies a resource node, communication could effect completion of task and reduce the benefits. The larger amount of task lower benefit. so with some nodes should be removed to reduce the amount of communication.

Based on the above analysis, in this paper, task merger and a spanning tree based grid tasks rescheduling approach is presented to improve response time. This approach considers from reduce the number of tasks and choose minimal communication cost. First, mergering of small scale of solely task to reduce the number of tasks.Second, select the least costly communication path. Reduce the number of tasks can reduce the communication between tasks and the probability of rescheduling.

* This work is supported by Dr. start-up foundation of Liaoning Province (No.20091031)

According to the results of the combined tasks generate the minimum spanning tree to achieve the purpose of reducing user response time. To verify the effectiveness of the approach, We conducted a simulation experiment to compare existing approach with this approach. When the scale of task is larger, this approach can reduce response time and optimize utilization of resources, so this approach has better performance than others.

2. Related Works

Dealing with the problem of Scheduling, different researchers have different ideas.

First, such policies improve the prediction accuracy, reduced dynamic, more backup resources to make goal. For example, Plan Switching[1] belong to this policy, this approach build a series of activity diagrams before the application is executed. Each diagram represents a scheduling, if a diagram of the activity diagrams in the run-time become invalidation, then application will switch to another activity diagram, actually, it uses more resources backup to improve the feasibility and optimality of a static scheduling.

Second, such policies aimed to reduced dependence on the prediction accuracy and adapt to the dynamic nature .Mainly rescheduling strategy for such a plan as a means of adjusting.

SIL and MQD[2] ,could solve the static scheduling algorithm depends on the accuracy of performance prediction, but they could only be used to data-intensive and compute-intensive applications.

DAG-Man[4] is a scheduling system of Condor-G, support scheduling and rescheduling, but it be used as a fault-tolerant technology.

AHEFT[5] fully takes into account a variety of changes in resources and performance prediction accuracy for the impact on the optimal scheduling. It based on HEFT heuristic scheduling strategy to adjust the policy of application scheduling. High degree of parallelism, data-intensive applications use this policy to have better performance.

To sum up, research on tasks rescheduling at present don't consider the overall cost of the task scheduled for execution and response time. This study proposes a task-merging and rescheduling spanning tree concept and calculation method. Intended to reduce the weight by a reasonable number of tasks involved in scheduling to reduce the likelihood of re-scheduling. Through the minimum spanning tree optimization to further speed up user response time.

3. Problem Definition

There are a variety of resources in grid environment, an application take into a series of subtasks, then these subtasks are assigned to the grid resources. When these resources are changed, the tasks need to be rescheduled, this will increase the user's response time. This need to consider the overall situation for reduce response time. Overall reduction in the likelihood of rescheduling and reduce the number of the communication between tasks to achieve purpose that reduce user's response time.

3.1 Application Representation

Task diagram of grid application is a spanning tree, each node is self-management, the edge between tasks represents the relationship communication of tasks, and non-dependent tasks can be performed in parallel.

In grid computing, a grid application often requires coordination of multiple tasks to complete. A application T make up of tasks. w(tu) represent calculation weight of task tu. The minimal resource's requirements of task tu is a(tu), euv is relationship of communication from tu to tv. Show in Fig.1.

Definition 1. Task diagram. Grid application T expressed as TGT=(TT,ET). Collection of nodes TT={t1,t2,...,tn}, Set of edges ET={euv|0<u,v $\leqslant$ n;u $\neq$ v}. tu represent task; euv represent relationship of communication from tu to tv.

Definition 2. Task node. Each task is to save information of the parent node and child node,

information of interdependent nodes. When tasks are rescheduled, these information will to be changed; task node t(tu,Rid,Pid,Cid,Sid,a,w), tu is task mark. Rid is interdependent node of tu, Pid is parent node, Cid is set of child node, Sid is resource mark, a is minimum request of resource, w is weight of task. cuv is data transmission from tu to tv.

Each resource node saves a routing table for to communicate between tasks. when each task node send information to relative nodes, node based on local routing table transmit information to interdependent nodes for implement communication between tasks.

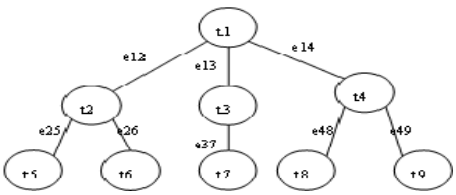


Fig.1 Grid task graph

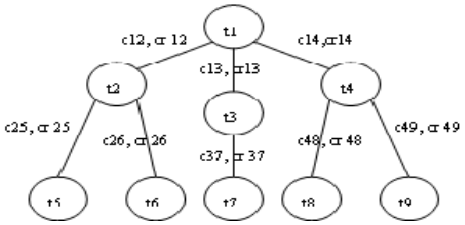


Fig.2 Grid task Minimum Spanning Tree

3.2 Grid Resource

When the application is submitted to the gird computing environment,they need to select a subset of resources from resources available to support application execution.

Definition 3. Grid Resource. All the resources registered to grid call grid resources, it is expressed as SG=(m,ws), m is resource mark, ws is capacity of resource.

Definition 4. Available Resource. a(tu) is represent as minimum required processing capacity of tu. SG=(m,ws), if ws>a(tu), m will to be a available resource of application T.

Definition 5. Resource Description of Grid. GR =(T,ST,CRT) is a three-tuple. T is grid application, ST is a collection of resources of the T, ST ={m|m ∉ T};CR =(cruv)n*n is network matrix in the ST. cruv is network connection bandwidth between resource u and resource v.

Definition 6. Communication cost between tasks. V(cuv,cruv) is data communication cost bwtween tasks. cuv is weight of data transmission from u to v. VT is cost of data communications for the application T.

Definition 7. MST(Minimum Spanning Tree).MST=(T,ST,VT). T={t1,t2,...,tn} is tasks of application T. ST ={m|m ∉ T}. VT is cost of data communications. Show in Fig.2.

4. Rescheduling

As the grid is a heterogeneous environment, computing power and connection bandwidth of resources is difference, so the different scheduling leads to different length of entire scheduling. Length of the scheduling determines the efficiency of scheduling, the longer scheduling lower efficiency of scheduling and increase response time.

Condition of trigger rescheduling: When tasks are executing,resource degradation, resource exit and other events that could impact on the application complete, resource performance raise and new resources to improve application performance.

Step of rescheduling:

1)Begin to execute application.

2)Detect status of the node during the task execution.

3) Perform rescheduling algorithm:

(1) task merger. The basic idea is reducing the number of tasks to achieve the two objectives: reduce the probability of rescheduling; reduce the number of edge in the communication between tasks. From these two aspects can reduce the user response time.

With the implementation of the tasks, weight of each task is gradually reduced. The small tasks were merged into one task, it could reduce number of total tasks. Lots of tasks could increase the probability of rescheduling , on the contrary, probability of rescheduling is smaller.

Definition 8. The initial quality of task tu is expressed as IMu, the present quality of task tu is expressed as CMu. tu and tv depend on each other. if (IMu+IMv)/2≥(CMu+CMv), then tu and tv form a new task, these new tasks to form a new set of application. TT={t1,t2,...,tk}, k≤n, n is number of previously tasks. k is present number of tasks.

(2)Search for available resources.

The minimal resources requirements of tu is expressed as a(tu).SG=(m,ws) is available resources of application T. if ws>a(tu), m is available resource of tu, it is marked as m ∉ tu. ST ={m|m ∉ T} is all of available resources of application T.

(3)Tasks TT select available resources from ST, they form a connected network that each node depended on each other. Each node maintains some information that include identification of task and identification of resource.

Merger Algorithm: According the definition 8 to merge tasks. To update information of new node ,the information include its parent and children id, status of executing or waiting.

Two tasks is combined into one task, their common and closest ancestors task as the parent task , their common and closest descendants node as the descendant task . when rescheduling is triggered, First, check whether it can be merged into the parent task, if it is not, then check whether it can be merged into the children task .if it is not, Don't merger. if the two tasks form a new task, the initial quality of the new task is IMu=CMu+CMv. If task to be combined with the new task, then according the initial quality of this new task to check whether it to be merged another task.

To combine tu and tv into a new task, then send message that tu and tv is combined to one task to other nodes that dependent on tu or tv. Such as, tu—>tv, —> be expressed as tu depends on tv, tu is parent of tv, tu be merged into the tv. For example, ti—>tu， tu—>tv , ti is parent of tu, tu be merged into tv, tu also be merged into ti, so it occurs redundant that tu be merged into two nodes. ti and tv send message that tu is mergerd be myself to parent node and child node of tu, ti is received a message and send a message, so it know tu is be merged other node, and tv has also two messages. Because ti is ancestor, tv is descendant node, so tu merged into the ti, tv abandon merger.

There are six cases occurred in the merger process:

case 1: only one parent, only one children. show in fig.3. According (IMi+IMj)/2≥(CMi+CMj) in definition 8 to decide that task tj is merged to ti or tk.

case 2: only one parent, more children. Show in fig.4. tj have more children, find out the minimal quality task from all children(tk,tu,tv) to combine to tj. if they have the same task quality, select anyone to be taken as case 1.

Fig.3 One parent and one children(case 1)

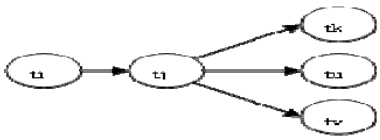


Fig.4 One parent and more children(case 2)

Case 3:more parents, only one children. show in fig.5. tj have more parents, find out the minimum quality task from all parents(ti,tu,tv) to combine to tj. If they have the same task quality, select any one to be taken as case 1.

Case 4:more parents,more children. show in fig.6. To select minimum quality task from parents(ti,tu,tv),to be taken as case 3.

Case 5: no parents. Show in fig.7.

Case 6: no children. Show in fig.8. To decide whether to merge tj by definition 8.

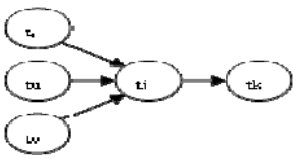


Fig.5 More parents and one children(case 3)

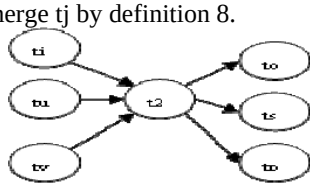


Fig.6 More parents and more children(case 4)

Fig.7 No parents(case 5)

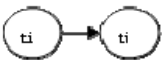


Fig.8 No children(case 6)

The formation of network connectivity graph:

Task tu select resource from available resources ST, there are dependencies between the tasks connected to form a dependency of the undirected graph RG, previous communication edge are joined in RG to form a network connectivity graph, then to form a spanning tree. Such as:

Using the new task set T and available resources ST form RG. tu and tv have a directly dependence and directly communication edge, vice versa. eou is communication edge between tu and tv in previous spanning tree before rescheduling is triggered, then eou join into RG to form a network connectivity graph. Show in fig.9.

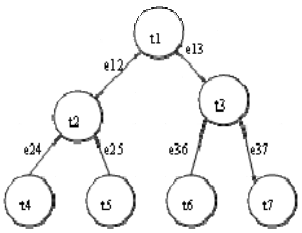


Fig.9 ST before scheduling

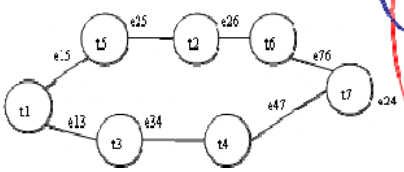


Fig.10 RG after rescheduling

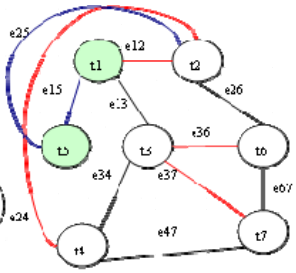


Fig.11 Connected graph

Definition 9. RG=(Tk, Ek), Ek are set of edge in RG. STG=(Tn, En) is expressed as spanning tree before rescheduling is triggered, En are set of edge. E=Ek ∪ En. E is set of edge in network connectivity graph.

Definition 10. Total cost of the data transmission between tasks is the weight of edge. it is expressed as $L(uv)=c_{uv}/cr_{uv}$, network connectivity graph CNG=(Tk,E, LT), LT is set of costs.

Figure 11 is network connectivity graph that figure 9 is combined into fig.11. Red line is previous edge. En={e12,e13,e24,e25,e36,e37} is set of edge before rescheduling. Ek={e15,e13,e34,e47,e25,e26,e76} is set of edge connectivity graph, E=En ∩ Ek.

In fig.11, t5 and t1 have dependence, if t5 and t1 accord with definition 8, t5 is combined into t1, t5 and t1 to form a new task t1. {t2,t3} is its succeed tasks. so we could delete e15 and e25,E=En ∩ Ek-{e15,e25}, Tk={t1,t2,t3,t4,t6,t7}. The consequence shows that the number of tasks and communication edge are reduced. It could reduce communication cost and probability of rescheduling.

(4)There are many algorithm to constructed spanning tree, many algorithm uses the nature of MST, to suppose N=(V,{E}) be expressed as connectivity graph, U is un-empty subset of V, if (u,v) is minimum weight edge and u∈U， v∈V-U, so there will be a minimum spanning tree that contains (u,v).

The most commonly algorithm is Prim and Kruskal algorithm, the time complexity of Prim is O(n2), edge of the graph has noting to do with Prim. The time complexity of Kruskal is O(eloge), e is number of graph, it suitable for rare edge of graph. This paper use Kruskal algorithm , because grid application have more tasks than other.

CNG=(Tk, E, LT). NO={VO, EO}, EO={},VO=Tk, each vertex connected component self-contained in the VO. Select minimum cost edge from LT, if this edge be contained different vertex in VO, this edge be added to EO, else select next edge and delete this edge. continue this algorithm, until all the VO vertex falls on the same connected components, the result is minimum spanning tree.

5. Experiment

5.1 Experimental design

This paper use grid middleware Globus Toolkit to design grid environment simulator to validate our algorithm. This simulator is divided into three layers, the upper layer is experimental simulator, the middle layer is scheduling, and the lower layer is Globus environment. Globus environment use MDS, DUROC, and GRAM services and implement scheduling interface. The scheduling layer mainly implement the scheduler and rescheduling trigger. Experimental simulation layer mainly implement the virtual task generator, task merger, task graph and data statistics. The hardware resources are 120 Lenovo PC, cpu 2.0~2.4G, memory 1G, Window2000 os, 2 giabit router.

5.2 Experimental results

In order to verify the effectiveness of the algorithm, the author compared Max-min, min-min, Max-Int algorithm with proposed algorithms in this paper. The main compare indicators are: the task response time and resource utilization. Show in fig.12, fig.13.

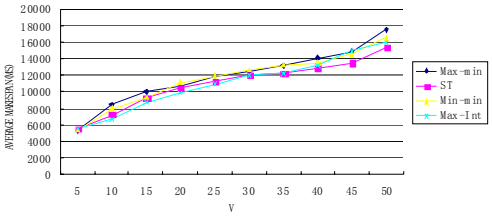


Fig.12 Comparison of response time

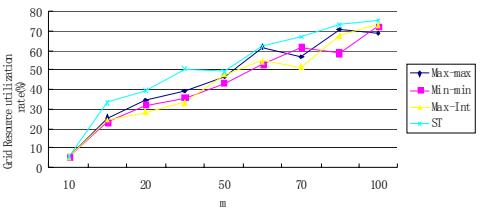


Fig.13 Comparison of resource utilization

Figure 12 shows that when there are fewer number of tasks, the response time of each algorithm is closest. With the increase of number of tasks, Communication between tasks is also increase, thus the algorithm task response time is lower than other algorithms. It can be seen that task response time of this algorithm is lower than other algorithms in larger number of task.

Figure 13 show that resource utilization of this algorithm is better than other algorithm. With reduce of number of tasks, this algorithm is also reducing the usage of the resources. Utilization of each resource will achieve the first best . So this algorithm has greater value in resource utilization .

6. Conclusion

This paper propose an algorithm to resolve the problem about user's response time and frequently rescheduling. The target of this algorithm is reduce user's response time , it reduce probability of rescheduling and use minimum spanning tree to optimize the whole task response time.

References

[1] Yu H, Marinescu D C, and et al. Plan switching: an approach to plan execution in changing environments[A]. Proceedings of the 2006 International Parallel and Distributed Processing Symposium[C], 2006:33-41.

[2] Lee Y C and Zomaya A Y. Practical Scheduling of Bag-of-Tasks Applications on Grids with Dynamic Resilience, IEEE TRANSACTIONS ON COMPUTERS, 2007,56(6): 815-825.

[3] Sakellariou R and Zhao H. A low-cost rescheduling policy for efficient mapping of workflows on grid systems[J]. Scientific Programming, 2004,12(4):253-262.

[4] Imamagic E, Radic B, Dobrenic D. An approach to grid scheduling by using condor-G matchmaking mechanism. Information Technology Interfaces[A]. Proceedings of the 28th International Conference[C], 2006:625-632.

[5] Yu Zhinfeng and Shi Weisong. An Adaptive Rescheduling Strategy for Grid Workflow Applications Parallel and Distributed Processing Symposium[A]. Proceedings of the 2007 International Parallel and Distributed Processing Symposium[C]. 2007:1-8.

2012 International Conference on Affective Computing and Intelligent Interaction
Lecture Notes in Information Technology, Vol.10

Path Planning of a Data Mule for Data Collection in the Sensor Network by Using an Improved Clustering-Based Genetic Algorithm

Ko-Ming Chiu[1, a] and Jing-Sin Liu [1,b]

[1] Institute of Information Science, Academia Sinica, Nangang, Taipei, Taiwan 115, ROC

[a]chiukoming@iis.sinica.edu.tw, [b]liu@iis.sinica.edu.tw

Keywords: Path Planning, Traveling Salesperson Problem with Neighborhoods (TSPN), Genetic Algorithms (GAs), Data Collection

Abstract. In recent years, use of a data mule for collecting data in the sensor network has become an important issue. Previous related researches showed that use of data mule indeed significantly reduces energy consumption at sensor nodes compared to commonly-used multi-hop forwarding approach in dense network. Use of data mule has a fatal drawback that increases the latency of data delivery. In order to decrease the latency of delivery, the path length of a data mule must be shortened as possible to decrease latency of data delivery and energy. Planning path of a data mule for collecting data in the sensor network with variable communication ranges can be regarded as a variant of traveling salesman problem with (variable-size) neighborhood (TSPN), hence the data collection problem is a NP-hard problem which evolutionary computation could be effectively applied to find a collection of suboptimal paths. In this paper, we design an improved clustering–based genetic algorithm by employing a CSEX crossover operator and a 2-opt mutation operator in combination with a closest waypoint algorithm to effectively reduce path length of a data mule. Simulations results confirm the effectiveness of the new design of clustering-based genetic algorithm in shortening the path that the data mule had to travel in a WSN containing sensors with variable sensing ranges.

Introduction and Related Works

Rapid advances in the technology of wireless sensor networks (WSN) have led to its incorporation in a variety of applications, such as environment monitoring, surveillance systems and unmanned space or planet exploration. Due to constrained energy of sensors, a WSN may not be fully connected or it needs a huge number of sensors to be fully connected. In recent years, effective use of one or multiple data mules for collecting data in the sensor network has become an important issue [14]. To transfer data stored in nodes to base station, one possible solution for this data collection problem is to employ one or more mobile robots/elements mounted with sensor and wireless communication devices, called data mules (or mobile sinks), that can communicate with sensor nodes to autonomously gather data from all sensors and may prevent the sink area to become a bottleneck. The data mule can follow random, predictable, or controlled mobility. Recent related researches [10], [15], [16], [18], [19] have shown that for different density of nodes the use of a data mule indeed significantly reduces data latency and energy consumption at sensor nodes compared to commonly-used multi-hop forwarding approach, thus prolonging network lifetime [17]. However, a fatal drawback due to the use of a data mule is increasing the latency of data delivery in WSN. In order to decrease the latency of data as delivery, the traveling path length of a data mule must be shortened. Use of a data mule for collecting data in the wireless sensor network involves generating a path along which the mobile robot can retrieve all data from all sensors while minimizing overall travel costs (related to latency of data delivery) subject to hard or soft constraints on path or network. It can be regarded as a special form of the Traveling Salesman Problem with Neighborhoods (TSPN) [3], a variant of the Traveling

Salesman Problem (TSP) known as the NP-hard problem [5], [7] where each neighborhood is a disk region whose center corresponds to a node and the radius is determined by the communication range [11]. First studied by Arkin and Hassin [3], TSPN solutions examine a collection of several regions in a plane, called neighborhoods, and find the shortest tour that visits all neighborhoods. For a WSN containing sensors with overlapping communication ranges, methods such as the greedy method, the approximation algorithm [4] or genetic algorithms (GAs) based design of [6] fail to effectively decrease travel costs. According to the previous researches about clustering [8] in the sensor network, the mechanism of clustering is indeed beneficial to reduction of data delivery in wireless sensor network. For data collection using a data mule, clustering is also beneficial to decrease the number of locations that must be visited by a data mule, as demonstrated by [9], so that the data mule could follow a shorter route to finish the task of data collection in sensor network: it only visits a specific location within the effective communication range of a sensor (or a cluster) to download data, and thus does not need to visit the sensor's precise location. In this paper, to decrease latency of data delivery further we propose an improved algorithmic design and implementation of clustering-based genetic algorithm of [9] for the minimization of travel cost, aiming for shortening path length that a data mule had to travel to collect the data from all sensors with variable sensing ranges.

Model and problem statement

The network consists of a data mule R and a set of stationary sensors $S = \{s_1 , s_2 , s_3, \dots , s_n\}$. All sensors are deployed over a two-dimensional plane in advance and their precise locations are also known in advance. Let the sensing range and transmission range of each sensor be represented as r_s and r_t, respectively. We assume that each sensor s_i can sense a disk region centered at s_i with radius r_s. Each sensor is equipped with omnidirectional antennae to communicate with other sensors or the data mule within a disk region centered at s_i with radius r_t . Then, the data mule R can receive data from or communicate with other sensors when the data mule is inside the sensor transmission range.
The data collection problem is that a data mule moves from a given start position (S) to collect data from each sensor with communication range (r_s) in a wireless sensor network. The data mule returns to the start position (S) after finishing the data-collecting task. The main objective (see e.g. [2], [18]) is to generate a path to guide the data mule to collect all data from all sensors and minimizes a travel cost defined by the path and sensor network. As stated previously, the data collection problem for finding a shorter route to optimize the distance the data mule had to travel can be regarded as a TSP with neighborhood (TSPN) (see e.g.[1], [17]), a variant of travelling salesman problem-- an NP-hard problem [5]. Thus, the data collection problem in WSN is suitable for GA-based approaches that are capable of finding multiple reasonably good routes.

Improved Clustering-based Genetic Algorithm

Clustering Algorithm for Variable Sensing Range

In order to decrease entire travel cost, a core focus of our study is reducing the number of visited nodes while still ensuring the data mule can precisely collect all data from all sensors. By dividing the global network into small clusters to reduce the overall complexity, the data mule travels only to the intersection areas within each of these clusters in WSN to collect the data of sensors within one cluster at the visited point of each cluster using one-hop communication, thus saving the time and energy. In [9], the clustering algorithm can make a sensor network consisting of sensors with identical sensing range be classified into several clusters $L = \{L_1, L_2, \dots\dots, L_m\}$. In this paper, we propose an improved clustering algorithm to handle the classification of sensors with variable sensing range. The improved clustering algorithm will be described as follows.

Improved Clustering algorithm: The clustering algorithm makes all sensors with variable sensing range be classified into several clusters. Sensors belonging to the identical cluster have a common intersection area.

Input: a set of sensors $\{s_1, s_2, s_3, \ldots, s_n\}$ and each sensor has own sensing radius (communication range) r_i, $1 \le i \le n$.

Output: a set of clusters $L = \{L_1, L_2, \ldots, L_m\}$.

1. Initially, $L = \emptyset$ and the corresponding Cluster IDs of all sensors are set to *null*.
2. For each sensor pair (s_i, s_j), i, j$\in\{1, 2, \ldots, n\}$ and i$\neq$j. If Euclideandistance$(s_i, s_j) \le (r_i + r_j)$, go to step 3, otherwise, directly go to step 11.
3. If s_i and s_j don't have the common Cluster IDs, go to step 4, otherwise, go to step 10.
4. If cluster IDs of s_i and s_j are *null*, go to step 5. Otherwise, go to step 6.
5. If Euclideandistance$(s_i, s_j) \le (r_i + r_j)$, the cluster IDs of s_i and s_j will be assigned to a new cluster ID and go to step 10. Otherwise, the cluster IDs of s_i and s_j will be assigned to a new cluster ID respectively and go to step 10.
6. If cluster IDs of s_i are not *null* and cluster IDs of s_j are *null*, go to step 7. Otherwise, go to step 8.
7. The cluster ID of s_j will be assigned to cluster ID of s_i when Euclideandistance$(s_p, s_j) \le (r_p + r_j)$, $s_p \in$ the cluster of s_i and go to step 10. Otherwise, the cluster ID of s_j will be assigned to a new cluster ID and go to step 10.
8. If cluster IDs of s_i and s_j are not *null*, go to step 9. Otherwise, go to step 10.
9. If Euclideandistance$(s_p, s_i) \le (r_p + r_i)$, $s_p \in$ the cluster of s_j and Euclideandistance$(s_q, s_j) \le (r_q + r_j)$, $s_q \in$ the cluster of s_i, The cluster IDs of s_j will be added to the cluster ID of s_i and go to step 10. Otherwise, directly go to step 10.
10. If all sensors are divided into a set of clusters L, return back with L, otherwise, go to step 2.

If $p = (p_1, p_2)$ and $q = (q_1, q_2)$, the function Euclideandistance(p, q) is defined as follows.

$$\text{Euclideandistance}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \quad (1)$$

By the above clustering algorithm, a set of clusters in a sensor network is obtained.

Chromosome Encoding and Generation of Initial Population

By above clustering algorithm, a set of clusters is obtained for a sensor network. Each permutation of these clusters can be regarded (or encoded) as an independent chromosome. In order to generate the initial population with high diversity (no repeated chromosomes), we adopt permutation tree based Chromosome Generation Algorithm (CGA) developed in our previous work [9] to generate the initial population consisting of a sufficient number of valid tours.

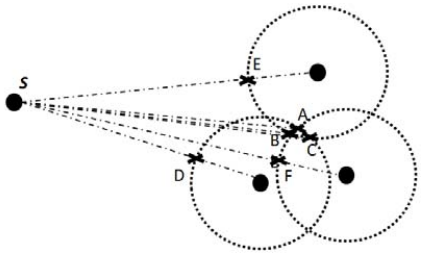


Fig. 1. Illustration of finding the closest waypoint for a cluster to a start point S.

Finding the Closest Waypoint (CW) for Each Cluster to a Start Point

The genes of each chromosome consist of clusters, and the closest waypoint for each cluster varies according to different permutation of clusters. Given permutation of each chromosome, the closest waypoint in a cluster is obtained by following algorithm (see Fig. 1 for an illustration).

CW algorithm: This algorithm can find the closest waypoint (CW) when the inputs are given.

Input: A start position S and next target cluster to visit.
Output: the coordinate of the closest waypoint (CW) within the target cluster from S.

1. Find all sensors belonging to a target cluster.
2. Calculate points of intersection according to communication range of sensors and add these points to a set of points (PC).
3. Calculate cross-points from S to the center of each sensor and add these points to PC.
4. Select a point with shorter Euclidean distance from PC, where the data mule can move to this point to collect data of all sensors belonging to this target cluster.
5. A point is selected as the closest waypoint and is returned back with CW.

Fitness Evaluation

Each chromosome in the population represents a possible path for the data mule as an ordered sequence of visited clusters. Each is evaluated by a fitness value which indicates the quality of the path. Let f_i denote the fitness value of *i*-th chromosome defined as follows

$$f_i = 1/C_i \tag{2}$$

where C_i denotes the total cost of the *i*-th chromosome defined as the sum of the Euclidean distances. C_i of a chromosome (CH_i) can be calculated as follows. First, a tour CH_i is regarded as a graph $G = (V, E)$ where $V = L = \{L_1, L_2, \ldots..L_m\}$ is a set of m clusters, E is a set of edges. Then, total cost C_i of *i*-th chromosome is obtained as

$$C_i = \sum C_{jk} \text{ and } C_{jk} \in E \in CH_i,\ 1 \le j, k \le |CH_i|, j \neq k \tag{3}$$

where the cost C_{jk} denotes Euclidean distance between position j and position k of tour CH_i defined as

$$C_{jk} = \text{ Euclideandistance}(j, k) \text{ and } C_{jk} \in E,\ j \neq k,\ 1 \le j, k \le m \tag{4}$$

A higher fitness value is given for a superior chromosome.

Selection, Crossover and Mutation operations

Generating the better posterity by evolution operations (such as crossover and mutation) is an important issue for genetic algorithm (GA). Firstly, we can array all chromosomes of population in descending order by their fitness values. Then, two chromosomes with higher fitness values are selected as two parents for crossover that exchanges information between two parents for generating descendants. The crossover operation is used to explore new permutation of chromosome and hopefully to be able to reduce the length of the current tour (i.e. a solution with better fitness value). A sub-tour based crossover tailored to TSP, complete sub-tour exchange crossover (CSEX) [12], is used as crossover operation in this paper. CSEX can guarantee that all descendants are feasible when two parents have common sub-tours. In this way, the sub-tours are preserved and exchanged from generations to generations.

It is important to diversify the search via mutation operation that occasionally creates a diversity of genes between descendants during genetic evolution process. In this paper, a 2-Opt move [13] is used as mutation operation. We find that the number of different genes before and after mutation increases, indicating that the diversity between original and mutated chromosomes gets larger.

Simulation Results

In our simulations, the sensors are randomly distributed over a two-dimensional map of size 500*500. The number of sensors is set to 100 and the sensor nodes are randomly deployed in the map.

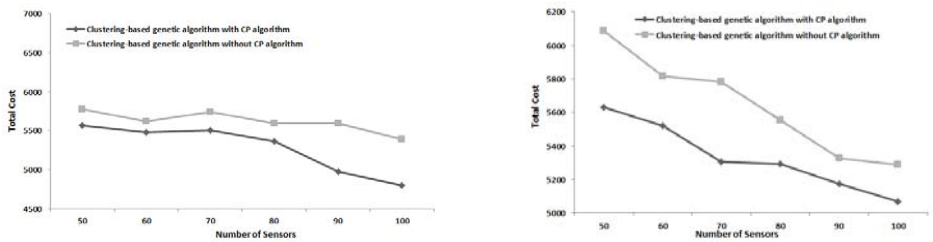
Analyses of Probability of Crossover and Mutation Operations

In this section, firstly, we will evaluate effect of crossover probability with CSEX on average total cost. From our analyses of simulations, we find that using CSEX crossover, the total cost with higher probability of crossover is better than the total cost with lower probability of crossover. Secondly, we re-examine the relation between population size and the probability of finding common sub-tours. According to our simulations, we find that the probability of finding common sub-tours and the number of found common sub-tours both increases when population size increases. In addition, we also find that convergence speed becomes fast when population size increases. Because the probability of successfully finding common sub-tours increases, the probability of successfully generating locally optimal tours also increases. Summarizing above, applying the crossover operation (CSEX) with higher crossover probability is indeed beneficial to decrease total cost (latency of data delivery).

On the other hand, the effect of mutation operation is evaluated by varying the probability of mutation in the range [0.1, 1]. According to our simulations, the total cost is better when the probability of mutation falls into the range from 0.2 to 0.4. For the lower probability of mutation (e.g. the probability of mutation = 0.1) and the probability of mutation larger than 0.5, there is no improvement on total cost. The main cause is that when the probability of mutation is too high, a descendant with higher fitness value is easy to be re-permuted and destroyed to produce new descendant with lower fitness value. On the contrary, when the probability of mutation is too low, the worse descendant with lower fitness value is not easy to be re-permuted to produce a new descendant with higher fitness value. Therefore, adaptive probability of mutation can be set a value which falls into the range from 0.2 to 0.4 to yield a better performance in the performed experiments.

Effect of CW Algorithm

Our previous work [9] demonstrated by simulations that clustering-based genetic algorithm with appropriate setting of parameters generates a shorter path. In this section, we will evaluate the effect of incorporating CW algorithm into clustering-based genetic algorithm upon further reduction of path length. Fig. 2 shows the cost comparison, in which each value is the average of 30 samples. It is seen that for the same permutation of cluster sequence, the clustering-based genetic algorithm with CW algorithm can efficiently find a shorter path by a more flexible choice of waypoint in each cluster by CW algorithm. We conclude that the clustering-based genetic algorithm with CW algorithm indeed effectively reduces the total cost of a data mule regardless of whether sensing range of sensors is fixed or variable.



(a) (b)

Fig. 2. Comparison of clustering-based genetic algorithm with and without CW algorithm [9] (population size = 300, probability of crossover = 1, probability of mutation = 0.2).

(a) fixed sensing range: the radius (r_t) of sensing range of all sensors are set to 30;

(b)for variable sensing range: the radius (r_t) of sensing range of all sensors are variable (r_t is assigned to a randomly generated value T, $10 \le T \le 50$).

Conclusions and Future Works

For the TSPN motivated from data collection problem in a WSN using a data mule with controlled mobility, this paper presents an easy-to-implement new algorithmic design and implementation of clustering-based genetic algorithm by incorporating a closest waypoint (CW) algorithm. Our design of GA employs two operations, complete sub-tour exchange crossover (CSEX) and a mutation mechanism of 2-Opt move, to enhance the evolutionary path planner performance for TSPN. From the comparative simulation results, CSEX provides a mechanism of generating more feasible tours, and 2-Opt move provides a mutation mechanism capable of generating a tour with high diversity. Comparative simulations confirm that a clustering-based genetic algorithm with closest waypoint algorithm can further shorten traveling path length of a data mule than clustering-based genetic algorithm developed in [9]. Because the connectivity and topology of wireless sensor network will change with time due to change of sensing range of sensors (the effect of energy depletion of sensors), the kind of off-line approaches, such as genetic algorithm, doesn't handle the situation of topology of wireless sensor network. A future work is to devise a path planner that could generate a path to guide a data mule for data collection in real time.

References

[1] O. Tekdas, V. Isler , J.H. Lim and A. Terzis. Using mobile robots to harvest data from sensor fields, *IEEE Wireless Communications, Special Issue on Wireless Communications in Networked Robotics*, vol. 16, no.1, 22–28. (2009)

[2] D. Bhadauria and V. Isler, Data gathering tours for mobile robots, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3868-3873. (2009)

[3] E.M. Arkin and R.Hassin, Approximation algorithms for the geometric covering salesman problem, *Discrete Applied Math*., vol. 55, no.3, 197-218. (1994)

[4] Khaled Elbassioni, Aleksei V. Fishkin, Nabil H. Mustafa and René Sitters, Approximation algorithms for Euclidean group TSP, *LNCS 3580*, 1115-1126. (2005)

[5] C.H. Papadimitriou, The Euclidean traveling salesman problem is NP-complete, *Theoretical Computer Science*, vol. 4, no. 3, 237-244. (1977)

[6] Chang Wook Ahn; Ramakrishna, R.S., A genetic algorithm for shortest path routing problem and the sizing of populations, *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 6, 1-7.(1999)

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, MIT Press, Cambridge, Massachusetts. (2001)

[8] R. V. Kulkarni, A. Förster, and G. K. Venayagamoorthy, Computational intelligence in wireless sensor networks: A survey, *IEEE Communications Surveys & Tutorials*, vol. 13,1 – 29. (2011)

[9] K. M. Chiu and J. S. Liu, Robot routing using clustering-based parallel genetic algorithm with migration, *2011 IEEE Workshop on Merging Fields of Computational Intelligence and Sensor Technology*, 42-48. (2011)

[10] Ryo Sugihara and Rajesh K. Gupta. Path planning of data mules in sensor networks, *ACM Transactions on Sensor Networks*, vol. 8, no. 1, Aug. (2011)

[11] B. Yuan, M. Orlowska, and S. Sadiq. On the optimal robot routing problem in wireless sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, 1252–1261. (2007)

[12] K. Katayama, H. Hirabayashi and H. Narihisa, Performance analysis of a new genetic crossover for the traveling salesman problem, *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E81-A, no.5, 738-750. (1998)

[13] E. Aarts and J. K. Lenstra, Local search in combinatorial optimization, Wiley Series in Discrete Mathematics & Optimization, 215-310. (1997)

[14] M. Di Francesco, S.K. Das and G. Anastasi. Data collection in wireless sensor network with mobile elements: a survey, *ACM Transactions on Sensor Networks*. (2011)

[15] L. He, Z. Chen and J.D. Xu. Optimizing data collection path in sensor networks with mobile elements, *International Journal of Automation and Computing*, vol. 8, no. 1, 69-77. (2011)

[16] T.C. Chen, T.S. Chen and P.Y. Wu. On data collection in wireless sensor networks. *IEEE Transactions on Systems, Man, an Cybernetics*, *Part B*, vol.11, no.6, 1213-1224.(2011)

[17] A. Jarry, P. Leone, S. Nokoletseas and J. Rolim. Optimal data gathering paths and energy balance mechanisms in wireless networks, *LNCS* 6131, 288-305.(2010)

[18] S. Gao, H. Zhang and S.K. Das. Efficient data collection in wireless sensor network with path constrained mobile sinks, *IEEE Transactions on Mobile Computing*, vol.10, no.5, 592-608.(2011)

[19] L. He, J. Pan and J. Xu. Reducing data collection latency in wireless sensor network with mobile elements, *INFORCOM IEEE Conference Computer Communications Workshops*,572-577.(2011)

Rule Extraction from SOM for Academic Evaluation

Sathya Ramadass[1,a] and Annamma Abhraham[2,b]

[1]Professor, Dept. of MCA, Jyoti Nivas College (Autonomous), Bangalore, India.

[2]Professor and Head, Dept. of Mathematics, B.M.S.I Technology College, Bangalore, India.

[a]sathjoe@gmail.com, [b]annamma65@gmail.com

Key words: Cluster approach, KSOM, Rule-extraction, Pattern extraction

Abstract: Cognitive approach of neural network is being used in many real world applications to solve a problem, especially wherever nonlinearity exists. Though neural network has many advantages, one major disadvantage is the knowledge learned by a neural network is difficult to interpret. But, the knowledge from neural network can be extracted in the form of symbolic rules. Likely, artificial neural network are accepted in industry and other applications, but use of neural network in education is limited. In this paper, we use the direct method to extract the symbolic rules from a Self-Organizing Network that is established to predict the students' performance during the admission to a postgraduate course and show that, rate of evaluation helps the institution in the academic process.

1. Introduction

Since the development of Neural Network (NN), it has successfully been applied to many areas like image recognition, handwriting recognition, voice and speech recognition, weather forecasting, stock prediction control etc, due to the learning, adaptation, and training and generalization capability. For example, NN can learn and analyze large amounts of data to establish patterns and characteristic in situations in lesser time and also when input information are incomplete or noisy in them. Thus, NN have the potential to provide some human characteristics to problem solving that are difficult to simulate using logical, analytical techniques and standard software technologies. Parallel-distributed architecture of the NN makes brain-style computation possible. The intelligence of a NN emerges from the collective behavior of neurons, which performs limited operation but works in parallel. The topology of NN relies on the arrangement of those neurons. Based on the interconnection scheme, a NN can be either feed-forward or recurrent. Similarly, NN models can also be classified in term of their applications: classification models, association models, optimization models and self-organizing models.

Whatever be the model, a general assumption on a NN is that though the process of knowledge acquisition is simple, it lacks in the explaining the reasons. For example, in a pattern identification process a NN model will identify the characteristics of the input attributes based on the similarity and cluster them into an output neuron that signifies that pattern. Since all the information, initial knowledge, the learning rule in terms of weight, the operation and the output patterns are in numeric representation, human cannot reason the result obtained by the NN model. To overcome this disadvantage, NN model can have a rule extraction procedure explicitly or augmented with the model when it is designed. This procedure will extract the rules or the patterns, which signifies the input attributes in a symbolic representation. The simplest form of symbolic representation is if-then rules as, *if* $x \in X_n$ *then* $C(x) = C_k$ where x is a pattern present in the subset X_n then it belongs to a cluster C_k. In other words, a rule generated from a neural network has the form, *if the premise, then the action*. The rule premise is limited to a conjunction of attributes.

In this paper, we apply direct method of rule extraction on a clustered node in a KSOM and the extracted rules are further examined for system's accuracy. This paper is the revised and extended version of our previous paper [1]. The sections are organized as follows: section two describes the rule extraction methods, section three explains the direct method of rule extraction and section four analyzes the direct method applied in the problem domain and explain the results of experimental work.

2. Rule Extraction

Rule extraction is a procedure done on the cluster nodes or classified nodes to extract the patterns or the features. It also finds the effect of an attribute of the input pattern, in the selection of a cluster node. The goal of classification patterns is to identify the actual groups of related variables or attributes. The found clusters represent the data structure only with reference to the selected attributes. This may contain both significant and/or insignificant attributes. Distinguishing these attributes in the clusters will help 1) further data analysis on the input pattern 2) identifying the components to improve the performance of the network 3) reduce the input noise 4) provides insight into the data set. By extracting rules or patterns from the cluster will help to identify the attributes of the clusters. Also by extracting rules one can improve the performance of the system because it will solve the difficulties of knowledge acquisition, and it will explain the behavior of the system.

Rules can exist in different formats. The standard format if-then rule can be identified by the rule extraction algorithm from the neural network and further can be processed for refined network knowledge. If the NN is a supervised network, the rule is valid if the target concept is activated in excess of a certain activation function level. Suppose if the function level is 0.5 then the interval from 0.5 to 1 creates a range of levels that can be associated with rules. That is, possible patterns of a concept are combination of weights greater than the threshold. This cannot be done in unsupervised NN since, there is no threshold value and the output nodes are activated in competitive way. Hence, in these networks, it is possible to combine the weights that are maxima among all cluster or output nodes. In any way, a separate process is necessary to extract the rule or knowledge in the NN.

Some Knowledge-based conceptual neural network (KBCNN) developed by Limin Fu and RuleNet developed by McMillan have special architectures and special training algorithms so that when the networks have been trained, rules are immediately readable from the NN with less or no search [2, 3]. A rule that is extracted can be single rule or multiple rules from the NN. As Gallant suggests, to form a rule, the attribute with the greatest strength among attributes will be considered and the procedure continues until the conjunction of selected attributes is suitably strong to conclude the concept [4, 5]. There is some advantage if the rules are written in Fuzzy rules. Hayashi uses this principle and he organized the input into a set of cell groups [6]. Rule search is conducted directly in the space of primary attributes without involving pattern formation and combination in the hidden layer. Hence the overall search is limited. Another rule extraction algorithm NofM developed by Towell is used in many problem domains [7]. It explicitly searches for rules of the form if N of the following M antecedents are true, then concept is true. This algorithm will work efficiently with KBCNN or any knowledge based NN. Though this extracts rules efficiently, it cannot be applied for many real world problems because, it assumes symbolic concept and relations to hidden nodes and during learning period adaptation can not happen. Other extraction algorithm developed by many researchers are RuleNeg, KT algorithm, Rulex algorithm, ID3, SUBSET etc., [8,9]. Another algorithm called LREX algorithm developed consists of two modules mRex and hRex, where mRex extracts the IF THEN rule type from the domain in the hidden neuron and hRex finds which hidden neuron shares the between classes [10].

All the above procedures are for multilayered NN. Hence, the input to any rule extraction algorithm is a representation of the trained NN with its nodes, links and data set. Here, one or more

hidden and output neurons will participate in the derivation of rules. More specifically, as each neuron in the NN has a threshold value, propagating backwards from the cluster node to the hidden node will derive the rules of a concept based on the weights of the node which is greater than the threshold. However, in the case of any unsupervised networks like Kohonen's SOM, as there is no such threshold, the rule extraction should be done either by boundaries approach using U-Matrix (unified distance matrix) or using cluster approach. In boundary approach, U-matrix is used to find the boundary between the neighboring cluster units which leads to extraction of rules that describe the discovered pattern in the input vector [11,12]. In cluster approach, rule extraction is performed from a SOM by discovering clusters instead of boundaries. It finds the most significant value for each feature and cluster and extracts the rule for those clusters.

3. Rule Extraction using Direct Method

The rule generation aspect of NN is utilized to extract significant attribute that have influence on the clusters [2]. First using any NN architecture a cluster network is formulated. Then the rule extraction procedure starts. The direct approach used to extract influence attributes of a patterns in a cluster node:

1. For each cluster node j, search for a set of attributes/patterns with each pattern p satisfying the following condition:

The summed weights of p are the maximum among all clusters for cluster j, and $W_{ji} \geq W_{ki}$ for every i Є in $A_T - A_p$ and for every $k \neq j$, where A_p be the set of attributes involved in p and A_T is the set of all attributes.

2. Refine the extracted patterns or attributes in each cluster node j, if A_p participates in more than one cluster.

The input to this procedure is a data set that is clustered by the NN algorithm. Then it finds for a given cluster node, a valid pattern which maximally activates it in all circumstances. In other word, it finds the attributes of the input set that are significant to characterize each cluster or class. To do so (step 1), consider all cluster nodes. Calculate for each cluster node the activation of each attribute summed weight and mark the highest value for each row. In order to find the attributes that are most influencing for a description of a cluster, these significant values of the attributes are normalized in percentage of the total sum of significant values of a class. Then these are arranged in decreasing order. From this, the attributes with the largest significant value in the ordered sequence are taken until the cumulative percentage equals or exceeds a given threshold value. The standard threshold value is 50%. Thus, the attributes, which involves in this cumulative calculation has high influence on that class.

Similarly, if there are any attribute which are marked as highest significant value, and not considered so far, they are considered for description of the class. The algorithm is performed on all clusters and provides for each cluster the set of significant attributes to be used in a meaningful description of that class. If an attribute is exceedingly significant than all the others are, only very few attributes are selected. On the other hand, if almost all attributes possess the same significance, considerably more attributes are taken into account. Thus, number of rules describes a class and attributes are selected by the above procedure. These rules may be too strong or too soft to describe a class. This can be generalized by finding the significant attribute A_p in the range [min, max]. Sometimes, a rule can belong to two classes (step 2). In that case, a finer description about the class boundary between the two classes is necessary or the summed weights of p in conjunction with any pattern based on the attributes can be included in the rule.

4. Applications

We tested this rule extraction on a clustering network applied to a real world application. The network predicts post-graduate MBA students performance during their admission procedure. There are many researches done in education domain using NN like predicting students' performance using supervised NN, analyzing the technical college rank etc [13-18]. This paper, which is an extension to our earlier work [1], improved their previous model and extracts the rules using the direct method that is explained above to analyze the performance of each student and their significant characteristics for further guidance to them.

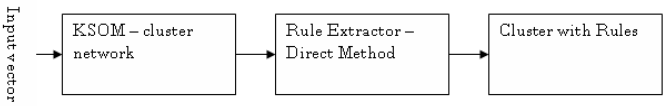


Fig.1. Process of Clustering and its Rules extraction

The process initiates with clustering network as shown in the Fig 1. Clustering of an input pattern can be done using supervised learning or unsupervised learning procedures. Since there are many works experimented with supervised NN for the specified application [15-17], this work concentrates on application of self-organizing network Kohonen's SOM. In our model the cluster network is a two layered KSOM. It is developed with 10 input neurons and 3 output neurons and implemented the pattern clustering process as presented in our earlier paper [1]. Input vector takes value between [0-1] and weight vectors are initialized as $1/\sqrt{d}$ where d is the number of input neurons and also the number of components in the weight vector. The system was processed for 200 data set and tested for 50 data. The accuracy of the system is 87%. The patterns are clustered to the 3 respective neurons first class, successful and below average. These are the performance rate of the student during their admission.

To find the significant attributes that are clustered into the node, as a second step, rule extraction algorithm is applied to each cluster nodes and the significant attributes are identified. For example, the significant attributes, which influence the selection of cluster node 1, are A_1, A_2, A_3, A_4 and A_5. That is, students who possess good academic record and eligibility criteria are clustered into node 1 that is identified as success rate above or equal to first class. Similarly, students who do not have commerce major in their undergraduate course or higher secondary and are under privileged in the status among other certain attributes are classed into cluster 3 (success rate is below average). In the same way, cluster node 2 has attribute patterns who do not posses a major degree in commerce but having work experience and comes from city than rural areas have exposure to subject with good percentage in under graduation have good success rate. In this way attribute, patterns are selected for each cluster nodes and hence the rules are extracted in the form of If A_1 and A_2 and A_3 and A_4 and A_5 then cluster C_k.

This type of rule extraction on a cluster node helps the education institution and associated faculties to adapt their methodology of teaching and other extra mentoring based on the students potentials. This is important to an institution because, 1) higher education is costly, 2) not all students apt for higher education and also, 3) competitions are high, 4) some people who join the course due to cost or inability to cope with the subject, the rate of discontinuing is more. So, if the institution predicts about the students success rate during admission, it will benefit the institution to motivate the students and train them accordingly and to students, to know their potential and compose them to work hard. Furthermore, the extracted rules are analyzed to improve the system's accuracy. It is observed that the prediction from SOM accuracy is high and can be further improved by tuning the input parameters. After the necessary refining, the network performs prediction very reliably.

5. Conclusion

In this paper, rule extraction method on a self-organizing model SOM is applied to forecast the student's efficiency during the admission into the post-graduate course MBA. The rules extracted from SOM provide information about the student's trends and their skills. Thus, it is useful for the students and institution to observe their performance rate and develop their skills according to their predicted result.

This work is an attempt to study the student's interest in post-graduate course and implementation is done for one college initially. The input parameters are taken from common admission process followed in India mainly in colleges in South India. Further, this can be implemented in other colleges and universities, which will be a beneficial to them to understand the potential of a student and educate them according to their performance.

References

[1] R. Sathya and A. Abhraham : *Application of Kohonan SOM in Prediction*, In Proc. of ICT 2010, CCIS 101, Springer-Verlag Berlin Heidelberg, p. 313–318, (2010)

[2] L. Fu : *Neural Networks in Computer Intelligence*, Tata McGraw-Hill Publishing Company Limited (2003)

[3] C. McMillan, M.C. Mozer and P. Smolensky: *The connectionist scientist game: Rule extraction and refinement in a neural network*, In Proc. of the Thirteenth Annual Conference of the Cognitive Science Society, (1991)

[4] S.I. Gallant: *Neural Network learning and Expert Systems*, The MIT Press, Cambridge, Massachusetts. (1993)

[5] K. Saito and R. Nakano: *Rule extraction from facts and neural networks*, INNC-90 Paris: in Proc. of the International Neural Network Conference. Paris, France, The Netherlands: Kluwer 1, p. 379-382 (1990)

[6] Y. Hayashi: *A neural expert system with automated extraction of fuzzy if-then rules and its application to medical diagnosis*, In advances in neural information processing systems, (1990).

[7] G. G. Towell and J.W. Shavlik: *Interpretation Of Artificial Neural Networks: Mapping Knowledge-Based Neural Networks Into Rules*, In: Advances in Neural Information Processing Systems, 4, J. Moody, S. Hanson, and R. Lippman, Eds., San Mateo, CA: Morgan Kaufmann (1992)

[8] R. Andrews and S. Geva: *Rule Extraction From Local Cluster Neural Nets*, Neurocomputing, (2000)

[9] M.T.A. Steiner, P.J.S. Neto, N.Y. Soma, T. Shimizu and J. C. Nievola: *Using Neural Network Rule Extraction for Credit-Risk Evaluation*, J. IJCSNS, Vol. 6 (5A), (2006)

[10] S.M. Kamruzzaman and Md. M. Islam: *An Algorithm to Extract Rules from Artificial Neural Networks for Medical Diagnosis Problems*, International Journal of Information Technology, Vol. 12 (8), (2006)

[11] P. Christos, M. Stylianos and S. Andreas: *Extracting Rules from Trained Self Organizing Maps*, International Conference Applied Computing, (2007)

[12] A. Ultsch, G. Guimaraes, G. Korus and H. Li: *Knowledge Extraction from Artificial Neural Networks and Applications*, In Proc. Transputer Anwender Treffen/ World Transputer Congress TAT/WTC 93 Aachen, Springer (1993)

[13] S.S. Mahapatra and M.S. Khan: *A neural network approach for assessing quality in technical education: an empirical study*, J. International Journal of Productivity and Quality Management, Vol. 2 (3), p. 287-306 (2007)

[14] E.R. Naganathan, R. Venkatesh and Uma Maheswari: *Intelligent Tutoring System: Predicting Students Results Using Neural Networks*, Journal of Convergence Information Technology, Vol. 3 (3), (2008)

[15] B. Naik and S. Ragothaman: *Using Neural Network to Predict MBA Student Success*, College Student Journal, 38, p. 143-149 (2004)

[16] P. Priti and K. Maitrei: *Forecasting Student Admission in Colleges with Neural Networks*, International Journal of Computer Science and Network Security, Vol. 7 (11), (2007)

[17] S.T. Karamouzis and A. Vrettos: *An Artificial Neural Network for Predicting Student Graduation Outcomes*, In Proc. of the World Congress on Engineering and Computer Science, (2008)

[18] V.O. Oladokun, A.T. Adebanjo and O.E. Charles-Owaba: *Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course*, The Pacific Journal of Science and Technology , 9, p. 72– 79 (2008)

EEG Analysis of Drivers under Emergency Situations

Luzheng Bi[1,a], Zhi Wang[1,b], and Xin-an Fan[1,c]

[1]School of Mechanical Engineering, Beijing Institute of Technology, South 5, Zhongguancun Street, Haidian District, 100081 Beijing, China

[a]bhxblz@bit.edu.cn, [b]20903137@bit.edu.cn, [c]3120100145@bit.edu.cn

Keywords: EEG, emergency situations, driver response, power spectrum analysis, detection model.

Abstract. This paper investigates the EEG characteristics of drivers' response to pedestrian sudden occurrence, as an example of emergency situations, to explore the feasibility of using the EEG of drivers to develop a method to address driver possible slow response for avoiding traffic accidents or reducing deaths and injuries. Nine drivers attended the experiment in a driving simulator with virtual driving environments, along with EEG signals being collected at twenty standard locations on the scalp. The power spectrum analysis was applied to capture the changes of EEG. Experiment results suggest that the power spectrum of the EEG of drivers under pedestrian sudden occurrence are significantly different with those of drivers' response to normal situations at the locations: P3, P4, P7, P8, PZ, O1, O2, OZ, T7, and T8, which can be used to help select associated information of EEG as input features of the detection model of pedestrian sudden occurrence.

Introduction

Road traffic accidents are a problem of serious concern to society killing approximately 1.3 million people and injuring tens of millions of people around the world every year [1]. Pedestrian hit by motorized vehicle is the biggest group of road fatalities [2]. Thus, it is important to develop counter measures to avoid pedestrian crashes and to mitigate pedestrian deaths and injuries.

Not taking action (braking hard or steering off) instantly is a major factor causing pedestrian accidents. One reason for not taking action is that drivers are not aware of pedestrians at all. Another reason is that drivers respond slowly due to the inherent response characteristic of drivers or other causes (like drowsiness and alcohol), although they attend pedestrians.

To address this challenge, some researchers have developed pedestrian detection systems to take place of drivers to detect pedestrians. These pedestrian detection systems apply various kinds of sensors and image processing techniques to detect pedestrians and to prevent accidents by warning drivers or triggering autonomous braking [3]-[9]. Although the kind of methods has made valuable findings to pedestrian detection, they have limitations in system performances, especially reaction time, which is especially critical under emergency situations such as pedestrian sudden occurrence.

According to the analysis above, it is necessary to explore an alternative method, as a complement to existing methods, to address the challenge of avoiding pedestrian accidents or reducing pedestrian injuries under pedestrian sudden occurrence in front of drivers when drivers are aware of this situation.

In this paper, we focus on an experimental investigation on electroencephalogram (EEG) characteristics of drivers' response to pedestrian sudden occurrence, as a case of emergency situations, so as to explore the feasibility of using EEG of drivers to predict emergency situations and develop a corresponding method for avoiding traffic accidents or reducing deaths and injuries, given drivers are aware of these situations. The reminder of this paper is organized as follows. In

section Ⅱ, we introduce the details of the experiment. In Section Ⅲ, we describe the data analysis and results. Our conclusion and a discussion of our future work are presented in Section IV.

Experiment

Experiment Design. A one-way, related samples design was used in this experiment to test the EEG changes of drivers between the two situations of pedestrian sudden occurrence and normal driving. The independent variable of the experiment was the variable of the two driving situations of pedestrian sudden occurrence and normal driving, and the dependent variable was the EEG change of drivers, which is measured by the mean of the power spectrum of EEG signals collected at the specified positions on the scalp**.**

Participants. Nine male drivers (aged 20 to 22) attended the experiment. All participants are physically healthy, and neither have any history of brain diseases nor eat any drugs before the experiment.

Data Collection. EEG potentials, referenced to the average of the left and right ear lobes, were recorded by using a 192- channels digital brain wave measurement system of Japan NEURO company with sintered Ag/AgCl electrodes, at the twenty standard locations (FP1, FP2, F3, F4, C3, C4, P3, P4, O1, O2, P7, P8, PZ, CZ, FZ, OZ, T7, T8, F7 and F8) based on an international 10–20 system [10]. The ground electrode was positioned on the forehead (FPz). The EEG was amplified and digitalized with a sampling rate of 500 Hz and a power-line notch filter to remove the line noise as well as a band-pass filter between 0.53 and 60 Hz to remove other high-frequency noise for further analysis. Before data collection, the contact impedance between EEG electrodes and scalp was calibrated to be below 10kΩ.

In addition, we recorded the time when a pedestrian occurred in order to extract the corresponding segment of EEG signals of driver response to pedestrian sudden occurrence.

Procedure. The experiment was conducted in a driving simulator, locating in a laboratory, with the EEG acquisition system being set well including good placements of electrodes at the corresponding locations of the scalp of participants well. The experimental session began with some practice until participants learned the driving task well. Participants were asked to maintain the vehicle at the center of the lane where the vehicle was traveling and brake hard as quickly as possible to avoid pedestrian hit when they were aware of pedestrian occurrence in front of the vehicle. To avoid the introduction of the effect of any possible prediction to pedestrian occurrence of the participants, pedestrian occurrence was randomly assigned during the whole process of driving.

Data Analysis and Results

Power Spectrum Analysis. Since analysis of changes in the power spectrum of EEG can characterize the EEG signals [11], we applied the power spectrum to capture the changes of EEG of driver response between the two situations of pedestrian sudden occurrence and normal driving situations. The two-seconds twenty-locations EEG signals of each subject associated with two situations were first extracted, respectively.

Then, every 250 ms, we calculated the power spectrum , $\tilde{P}(f)$, over the one-second data (i.e., N, the length of a sample, $x_N(n)$, is equal to 500 and the number of the samples is five with 75% overlap between two consecutive samples) at each location under each driving situation, using the Welch method with a hamming window, $d(n)$, of 500 ms and the 20% overlap between the two consecutive sections ($x_N^i(n)$ is the *i*th section and L represents the number of the sections and is equal to 6 here), according to the following Eq. 1 and Eq. 2 [12].

$$\hat{P}^i(f) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x_N^i(n)\,d(n)\,e^{-j2\pi fn}\right|^2, 1 \le i \le L, \quad (1)$$

where $\hat{P}^{i}(f)$ represents the power spectrum of the ith section, $U=\frac{1}{M}\sum_{n=0}^{M-1}d^{2}(n)$, and M is the length of hamming window and equals to 250.

$$\tilde{P}(f)=\frac{1}{L}\sum_{i=1}^{L}\hat{P}^{i}(f)=\frac{1}{LMU}\sum_{i=1}^{L}\left|\sum_{n=0}^{M-1}x_{N}^{i}(n)d(n)e^{-j2\pi fn}\right|^{2} \quad (2)$$

Finally, the mean of the power spectrum (in the band of 4-50 Hz) of each sample was calculated as the measure of the EEG, P.

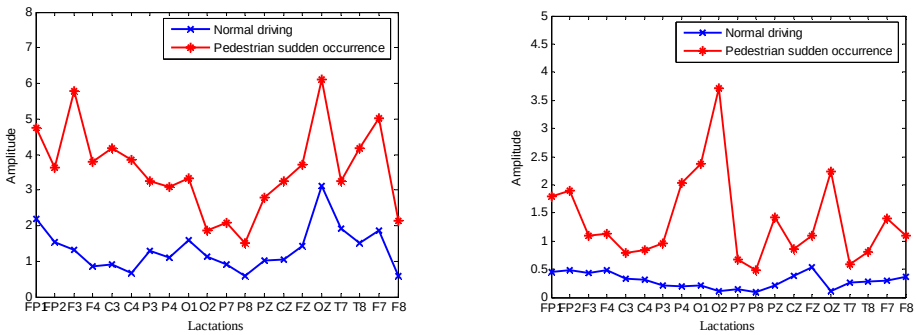
Results. Figs. 1 (a) and (b) show the comparisons of the average of P values of five samples at each selected location of subjects A and B between the two driving situations, respectively. From Fig. 1, we can see that there is likely significant difference in the power spectrum and the magnitude range between subjects is different.

Due to the individual difference in magnitude range, the average of P values, $\bar{P}$, was normalized according to the Eq. 3 so as to further examine the changes of EEG between two situations by using variance analysis.

$$\tilde{P}=\frac{\bar{P}-\bar{P}_{\min}}{\bar{P}_{\max}-\bar{P}_{\min}}, \quad (3)$$

where $\tilde{P}$ is the normalized result of $\bar{P}$.

Table 1 shows the results of a one-factor, repeated measures analysis of variance analysis (ANOVA) of $\tilde{P}$. As shown in Table 1, the power spectrum of the EEG at the locations including P3, P4, P7, P8, PZ, O1, O2, OZ, T7, T8 are statistically significantly different between the two situations of pedestrian sudden occurrence and normal driving, which can likely be used to select features of the to-be-developed detection model.



(a) Subject A (b) Subject B

Fig. 1. Comparisons of the average of P values of five samples at each location of (a) Subject A and (b) Subject B between the two driving situations.

Table 1. Results of ANOVA of $\tilde{P}$ of the EEG at all selected locations

Lactation	FP1	FP2	F3	F4	C3	C4	P3	P4	O1	O2
P value	0.309	0.531	0.305	0.201	0.256	0.066	*0.002*	*0.003*	*0.001*	*0*
Lactation	P7	P8	PZ	CZ	FZ	OZ	T7	T8	F7	F8
P value	*0.024*	*0*	*0.038*	0.209	0.325	*0.001*	*0.007*	*0.001*	0.583	0.282

Conclusion

In this paper, we investigated the EEG characteristics of driver response under pedestrian sudden occurrence, as a case of emergency situations, so as to explore the feasibility of using EEG of drivers to develop corresponding driver assistant systems to lower pedestrian hit. The experimental results suggest that the power spectrum of EEG signals at the locations: P3, P4, P7, P8, PZ, O1, O2, OZ, T7, and T8 are significant different between the two driving situations of pedestrian sudden occurrence and normal driving, which can likely be used to help select corresponding information of EEG potentials as input features of recognition models to detect emergency situations.

Our future work focuses on further validating our finding by using other emergency situations and developing corresponding models based on EEG to recognize emergency situations.

Acknowledgments. This paper is funded by National Science Foundation of China (NSFC) under grant 61004114.

References

[1] World Health Organization, http://www.who.int/violence_injury_prevention/road_safety_status /en/index.html

[2] Mohan, D.: Traffic safety and health in Indian cities. J. Trans. Infrastructure. 9, 79--94(2002)

[3] McCall, J. C., Trivedi, M. M.: Driver Behavior and Situation Aware Brake Assistance for Intelligent Vehicles. Proceedings of the IEEE. 95, 374--387(2007)

[4] Zhao, L., Thorpe, C.: Stereo and Neural Network-Based Pedestrian Detection. IEEE Trans. Intelligent Transportation Systems. 1, 148--154 (2000)

[5] Bertozzi, M., Broggi, A., Fascioli, A., Graf, T., Meinecke, M. M.: Pedestrian Detection for Driver Assistance Using Multiresolution Infrared Vision. IEEE Trans. Vehicular Technology. 53, 1666--1678 (2004)

[6] Broggi, A., Cerri, P., Ghidoni, S., Grisleri, P., Jung, H. G.: A New Approach to Urban Pedestrian Detection for Automatic Braking. IEEE Transactions on Intelligent Transportation Systems. 10, 594-605 (2009)

[7] Bi, L., Tsimhoni, O., Liu, Y.: Using image-based metrics to model pedestrian detection performance with night-vision systems. IEEE Trans. Intell. Transp. Syst.. 10, 155–164 (2009)

[8] Bi, L., Tsimhoni, O., Liu, Y.: Using the Support Vector Regression Approach to Model Human Performance. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans. 41, 410-417 (2011)

[9] Gerónimo, D., López, A. M., Sappa, A. D., Graf, T.: Survey of Pedestrian Detection for Advanced Driver Assistance Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence. 32, 1239-1258 (2010)

[10] Sharbrough, F., Chatrian, G. E., Lesser, R. P., Lers, H., Nuwer, M., Picton, T. W.: American Electroencephalographic Society Guidelines for Standard Electrode Position Nomenclature. J. Clin. Neurophysiol. 8, 200-202(1991)

[11] Lin, C. T., Wu, R. C., Liang, S. F., Chao, W. H., Chen, Y. J., Jung, T. P.: EEG-based drowsiness estimation for safety driving using independent component analysis. IEEE Transactions on Circuits and Systems Ⅰ: Regular Papers. 52, 2726-2738 (2005)

[12] Proakis, J. G., Manolakis, D. G.: Digital signal processing. Publishing House of Electronics Industry, Beijng (2007)

Emotion Image Retrieval Based on SOFM

Yang Tan[1, a], Guowen Wu[2,b] and Xin Luo[3,c]

[1]School of Computer Science and Technology

Donghua University

Shanghai,China

[a]yyjay2@163.com, [b]gwwu_dh@163.com, [c]rashin.lx@gmail.com

Keywords: SOFM image retrieval algorithm; Users' emotion space; Image emotion retrieval based on SOFM

Abstract. Image contains a wide range of emotions, and image low level feature can not fully meet the needs of users'. In this paper, a method adds users' emotion space based on SOFM retrieval algorithm is proposed, it used support vector machines to map images from the low level feature space to the high level emotion space, and automatically annotate unevaluated images. Thus the image retrieval is implemented by low level feature and users' emotion feature. From the experiment of 200 images, the retrieval system significantly improved the accuracy and fully satisfied users' requirements.

Introduction

With the development of computer technology, images, video files and audio files are increasing rapidly. How to quickly and accurately query images in huge database is the critical problem.

Content-based image retrieval (CBIR) is break the limitation of the keywords retrieval, directly analyze the images and the retrieval is getting close to the images entity. There are various kinds of methods in feature extraction, such as the feature extraction of shape, color, texture, and contour, according to these methods to measure the similarity of images, and finally realize the image retrieval. But in fact, such low level feature retrieval can not satisfy the precision of image retrieval.

Aiming at this problem, we present an approach as emotion image retrieval based on SOFM. Owning to accuracy and high efficiency of SOFM algorithm, we combine the users' emotion space, and SVM (Support Vector Machines) are used to map images from low lever feature to users' emotion space, and then automatically annotate unvalued images based on users' emotion. When users want to inquire images, firstly, similarity comparison in emotion space; secondly, use SOFM algorithm; finally, combine the results.

Related Work

CBIR system usually extracts color, texture, shape feature to express image low lever feature, and compare feature vector of query image with all vector in images database to measure the similarity of them. By the traditional method such as the SSA (Sequential Scan Algorithm), which is compared with each vector and the retrieval of huge database will surely exert a significant influence on the efficiency of retrieval, because of the scale of the database.

Image feature vector generally contain high-dimension, and the high-dimension index technology also be the hot topics of retrieval fields. Multidimensional index technology includes R-tree, SR-tree in Euclidian space, and the method in ordinary space includes VP tree, MVP tree, M tree, but these methods are difficult to solve the complicated boundary index problem, both the complex sample

distribution. The experimental result shows that the efficiency of all methods we mentioned will fall rapidly when dimension over 20, almost be equivalent to order query method [1]. This phenomenon is called curse of dimensionality.

In this paper, we adopt SOFM (Self-Organizing Feature Map, SOFM) [2] to accomplish the neighbors' retrieval in image database, better in both efficiency and precision.

Various emotions are aroused by various images, and different backgrounds may express different emotions in emotion feature fields. Thus image semantic analysis is essential in the image retrieval research and also feasible. But the present image retrieval takes no account of people's perception of image. Consequently, we bring in users' emotion and are able to identify the emotional information of image, enhance harmonious of the Man-Machine Interaction and efficiency of retrieval.

The present adjective of image emotion retrieval includes: an indefinite adjective space and a definite adjective space. The former method definite users' own emotion or expression by subjective estimation and decision directly[3], that is to say, the users feel free to add adjectives, but the method is different to analyses the adjectives and will have redundancies. Unlike the former method, the latter builds a definite space. Firstly, select the adjective which express users' emotion from the database. Additional, according to the questionnaire surveys, we gather the data from the images which valued from the subjects by SD (Semantic Differential), and the user's emotion database was made of the results. Finally, by using the multivariate analysis method to analyses the data, at last, emotional space is built [4]. The latter method ensures orthogonality of emotional space, and can measure the similarity in emotional space.

In this paper, we take the second method, and carefully select 18 pairs of representative adjectives to describe uses' emotion.

Proposed Method

SOFM image retrieval algorithm is based on Self-Organizing feature map, subject to high-dimension reservation, image feature mapped into one dimension space, and retrieval accomplished in low-dimension.

The process of this algorithm:

• By mapping high-dimension feature to one dimension using SOFM algorithm, then feature value is divided to several clusters and the cluster numbers= neuron numbers.

• The feature value of each cluster is pointed to dynamically allocated memory location, and each neuron located on one dimension space point to memory location pointer of its classes.

• When user requests images, at the first to search the optimum matching neuron .

• Calculating the distance from query image to neuron use high-dimension data of memory location.

• The results output in order of increasing according to distance.

The advantage of this algorithm is: on the basis of neighbor's relationship of high-dimension space, mapping high-dimension feature to one dimension space, using one dimension self-organizing map to implement high-dimension neighbor's retrieval in huge database, and show the great improvement in both efficiency and precision.

SOFM is based on image low level feature, but semantic in a image more than expression from low level feature (color, texture, etc) [5]. The image retrieval for our ideal is computer following users' subjective feeling and the way of understanding to image retrieval, it is more reasonable and humane. The system automatic extract image high lever feature, that is to say, the feature extracted from low lever feature has a great difference in users' understand [6].

Aiming at this problem, we proposed combine users' emotion to SOFM algorithm to overcome the semantic problem.

Users' Emotion Space

Image emotion feature defined as feelings, expresses, and emotions by image excited, or even various

subjective experiences. Besides semantic (color, texture, shape) in a image, there are more content of emotions. Ancient China has seven affections and six desires, and modern psychology has another way of saying it, but most approved of one point, that is human emotion including primary emotions and complex emotions. Primary emotions are the elements of emotions, and they may consist of complex emotions, that are described in emotional psychology.

1) We carefully consider 18 pairs of adjectives which listed in Table 1.

Table 1 18pairs of adjectives

beautiful — ugly	bright — dull
romantic — unromantic	like — dislike
changeable—monotone	vast —narrow
warm color—cold color	tidy — untidy
Harmonious—absonant	clean — dim
Carefree — depressive	afeared—calm
Animate — inanimate	happy—sad
Enthusiastic — cold	soft—unsoft
Solemn — playful	angry—quiet

2) For According to user's impression, he assigns each of given images a scale. There are 5 scales for him, "very like, a litter like, OK, dislike, strongly dislike", and the value is 0, 0.25, 0.5, 0.75, 1, respectively. 15 graduate students are elected to evaluate 30 sample images which are randomly selected in 200 images.

3) Factor analysis (FA) is used to analyze the data which evaluated by graduate students, then to build emotion space.

Let is the evaluation value of user to the sample image using the adjective, according to Eq.1, we have a mean, and matrix X is obtained by standardizing the mean using Eq.2.

$$y_{mn} = \frac{1}{k}\sum_{k=1}^{k} z_{mnk} \quad . \tag{1}$$

$$x_{mn} = \frac{y_{mn} - \overline{y_n}}{s_n} . \tag{2}$$

Where

$$\overline{y_n} = \frac{1}{M}\sum_{m=1}^{M} y_{mn} \quad . \quad s_n^2 = \sum_{m=1}^{M} \left(y_{mn} - \overline{y_n}\right)^2 \quad .$$

$$X = FA' + UD \quad . \tag{3}$$

F is called common factor matrix, A is loading matrix, U is unique matrix, and D is weight of unique factor. By PCA (principle component analysis), the dimension space is reduced to $L-$dimension space and called orthogonal emotion space. Additional, the $m-th$ row of matrix F ($f_m = (f_{m1}, f_{m2}, \cdots, f_{mL})$) is corresponding to the coordinate of simple image m in emotion space, and $n-th$ row of matrix A ($a_n = (a_{n1}, a_{n2}, \cdots, a_{nL})$) is indicates coordinate of adjective n also in emotion space.

Emotion Annotation

To build up mapping relationship on image low level feature to uses' emotion feature is defined as image emotion annotation, by learning users' emotion feature to automatically annotate every unevaluated image [7]. Analysis of factor analysis, L-dimension emotion space can be obtained.

The image database consists of 200 images, and there are 170 images besides the samples, which are needed to be mapped to users' emotion space.

In this paper, SVM (Support Vector Machine) is used to annotate images. (1) Construct the mapping image low level space to emotion space. (2) Learn the characteristic of users' emotion. (3) Automatically annotate every unevaluated image. After annotation, each image and adjective can be viewed as a vector in space.

Emotion Retrieval Model Based On SOFM

To meet the need of retrieval in emotion space and image low level space, the system combines the methods of image low level feature and emotion feature. According to different user's semantic, the system my flexible retrieval.

1) The retrieval of image low level feature. SOFM algorithm is only rely on image low level feature, then chosen the top 8 images as the results.

2) When users have emotion requirements, according to the chosen adjective, the system calculate similarity of the adjective of each image and show the top 8 images with the closest similarity to the chosen adjective as the results.

The similarity from image m to n adjective (d_{mn}):

$$d_{mn} = \frac{a_n \cdot f_m}{|a_n \| f_m|} \tag{4}$$

In fact, the defined similarity is proved, and be nicely able to satisfy the retrieval.

3) Combining users' emotion and image low level feature. Firstly, calculate the similarity of the adjective of each image and show the top 3 images named 1,2,3 with the closest similarity. Additional, select image 1 to use SOFM retrieval to show the top 3 images named 11,12,13; in the same way to image 2, and show the top 2 images named 21,22; the same to image 3 and choose the closest similarity image named 31. Finally, the retrieval result is: image 11, 12, 13,21,22,31.

Introduction

In this paper, our database consists of 200 representative images, and 30 images were taken out randomly.

Analysis of retrieval result of "warm color" and "romantic" is showed in Table 2, 3.

Table 2 the result of different retrieval requirement (warm color)

	SOFM retrieval	Emotion semantic retrieval	Image emotion retrieval based on SOFM
Average times	0	4.12	3.08
Accuracy rate	86%	85%	90%
Succeed rate	90%	90%	93%

Table 3 the result of different retrieval requirement (romantic)

	SOFM retrieval	Emotion semantic retrieval	Image emotion retrieval based on SOFM
Average times	0	4.34	3.67
Accuracy rate	85%	85%	90%
Succeed rate	90%	90%	93%

Accuracy rate [6] and success rate [7] are defined as following:

$$Accuracy\ rate = \frac{SN}{AN}. \tag{5}$$

$$Success\ rate = \frac{SUM}{UN}. \tag{6}$$

Where, SN is the number of images which users are satisfied; AN is the number of all retrieved images; SUM is the number of images who find he wanted and UN is the number of all users.

It is showed that according to different users' semantic, the SOFM algorithm has a satisfactory result under no average times; at the same time, emotion semantic retrieval is nicely meet the users' requirements under the acceptable average times (4 times); finally, image emotion retrieval based on SOFM combines the advantages of both SOFM and emotion semantic methods, higher accuracy rate and success rate is achieved. And the average times (3.5 times) is totally acceptable.

Conclusions

The resent retrieval based on image low level feature is not well take into account emotions, hobby, and subjectivity. However, image retrieval based on emotion is from the perspective of users, looking into the relations of images low level feature and users' emotion, finally a good emotion model is constructed to satisfy users' emotion requirement. In this paper, combining users' emotion and low level feature, the image retrieval system has a satisfactory result. However, the problem of semantic gap still existing between users' emotion and low level feature, and needs further research in the mass.

References

[1] Bohm C, Berchtold S and Keim D, in: Searching in High-Dimensional Spaces-Index Structures for Improving the Performance of Multi-media Databases, ACM Computing Surveys, p. 322.(2001)

[2] T. Kohonen, in: Self-OrganizingMaps, Third Edition, Brelin / Heidelberg, Springer.(2001)

[3] Lee Joo Young, Cho Sung Bae, in: Interacitve genetic algorithm with wavelet coefficients for emotional image retrieval, ethodologies for the Conception, Design and Application of Soft Computing,Proceedings of IIZUKA'98[C].p.829 (1989)

[4] Shibata.T, Kato.T, in: 'Kansei' image retrieval system for street landscape：discrimination and graphical parameters based on correlation of twoimages, IEEE International Conference on System Man and Cybernetic, Tokyo, Japan.vol. 6,p. 247(1999)

[5] Datta.R, Li Jia, Wang.J.Z, in: Content-based image retrieval-approaches and trends of the new Age, Proceedings of the 7th International Workshop on Multimedia Information Retrieval, in conjunction with ACM International Conference on Multimedia.p.253–262(2005)

[6] Rui.Y, Huang.T.S and Chang.S.F, in: Image retrieval: current techniques, promising directions, and open issues, Journal of Visual Communication and Image Representation.vol.10,p. 39(1999)

[7] WANG Shang-fei,WANG Xu-fa, in: A Double-Level Emotion Image Retrieval Model, Journal Of System Simulation.vol.16(2004)

Image Retrieval by Optimal Distance Measure based on Metric Matrix Learning Algorithm

Xin Luo[1, a], Yang Tan[1, b] and Guowen Wu[1, c]

[1]School of Computer Science and Technology

Donghua University

Shanghai, China

[a]xluo@dhu.edu.cn, [b]yyjay2@163.com, [c]gwwu@dhu.edu.cn

Keywords: CBIR; Euclidean distance; Distance Measure; Metric Matrix Learning.

Abstract. Much CBIR processing depends on the Euclidean distance or Mahalanobis distance function between two feature vectors, but this has problems with regard to feature weightings and feature correlations. Understanding the relationship among different distance measures is helpful in choosing a proper one for a particular application. This paper proposes an optimal metric matrix algorithm to improve image retrieval systems. This metric is optimal in the sense of global quadratic minimization, and can be obtained from the clusters in the training data in a supervised fashion. We compare two commonly used distance measures in CBIR system, namely, Euclidean distance and Mahalanobis distance, for nearest neighbor queries in high dimensional data spaces. Experimental results show that our approach is effective in improving the performance of content-based image retrieval systems.

1. Introduction

CBIR (Content-Based Image Retrieval) is often done by computing the distance from a query image to images in the database, followed by the retrieval of nearest neighbors. The retrieval performance mainly depends on two related components: the image representation and the distance function used. Given a specific image representation, the quality of the distance function used is the main key to a successful system. However, the distance function so far has been largely defined, usually by a weighting scheme and a simple cosine similarity, equivalently, a Euclidean dot product.

Distance metric learning aims to learn a distance metric from the training data that tries to maintain the class information of examples by their distances, i.e., examples sharing the same class are close to each other while examples from different classes are separated by a large distance. During the past few years, a large number of studies are devoted to distance metric learning [1]. In many machine learning problems, the distance metric used over the input data has critical impact on the success of a learning algorithm. For instance, k-Nearest Neighbor classification, and clustering algorithms such as k-means rely on if an appropriate distance metric is used to faithfully model the underlying relationships between the input data points [2]. Much research effort has been spent on learning a Mahalanobis distance metric from labeled data [3, 4].

In this paper, we propose an optimal distance function that is parameterized by a global metric matrix. This metric is optimal in the sense of global quadratic minimization, and can be learned from the given clusters in the training data. These clusters are often attributable with many forms, such as paragraphs, documents or document collections, as long as the items in the training data are not completely independent.

2. Distance Measure

Distances and metrics are now an important problem in information retrieval, machine learning and pattern recognition. The performance of algorithms for data classification and clustering often depends heavily on the availability of a good metric. In addition, distances and metrics have found application in a number of real-world problems, including face recognition, visual object recognition, and automated speech recognition.

In CBIR, the space of features is a vector space, but it is not obvious how to introduce a norm because of the incommensurability of the components. Similarity between descriptors is usually computed with either the Euclidean or the Mahalanobis distance measure.

A Euclidean distance matrix is one in which the (x, y) entry specifies the squared distance between particle x and particle y [5],

$$d_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{1}$$

However, there are two main problems with this distance[6]:

(1) The correlation between features is ignored;

(2) Feature weighting is inevitably arbitrary.

In those cases, the simple Euclidean distance is not an appropriate measure, while the Mahalanobis distance will adequately account for the correlations. According to Definition, Mahalanobis distance between two points $x = (x_1 \ldots .x_p)^T$ and $y = (y_1 \ldots \ldots y_p)^T$ in the p dimensional space R is defined as

$$d_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1}(x-y)} \tag{2}$$

The Mahalanobis distance gives better matching results but has 3 disadvantages, viz [7]

(1) it requires a covariance matrix estimated from training data;

(2) performance results will depend on the training set used;

(3) it is a global measure, not optimally adapted to the local structure at any feature point of interest.

Based on these problems, we propose a more generic measure similar to the metric matrix and takes into account the local structure at the feature points of interest. In this case the covariance matrix is obtained directly from the differential structure at each interest point. The matrix can be obtained in analytical form and reflects the actual behavior of the descriptor due to small perturbations. In the next section we present the details of this approach.

3. Optimal Metric Matrix

3.1 Ellipsoid Distance Function

As stated above, vectors in the same cluster must have a small distance between each other in the ideal geometry. When we measure an L_2-distance between x and y by a Mahalanobis distance parameterized by M :

$$D^2(x,y) = (x-y)^T M (x-y) \tag{3}$$

where symmetric metric matrix M gives both corresponding feature weights and feature correlations.

Generally, when setting M to be an identity matrix, the distance in Equation(3) becomes the common Euclidean distance.

For any symmetric matrix M , the following equation holds:

$$a_{ij} = a_{ji} \qquad (i, j = 1,2,..., p) \quad \text{and} \quad M^T = M$$

M is also a positive definite matrix, then it is easy to show that $x \neq 0$, $x^T M x > 0$.

By the learning matrix $M = [m_{ij}]$, Equation (2) can be rewritten as follows:

$$D^2(x,y) = \sum_{i}^{p}\sum_{j}^{p} m_{ij}(x_i - y_i)^T(x_j - y_j) \tag{4}$$

Where learning matrix M denotes the weight of any component and the relationship among components.

Because M is a symmetric matrix, then:

$$D^2(x,y) = (M^{1/2}(x-y)^T)(M^{1/2}(x-y)) \tag{5}$$

Note that this distance is global, and different from the ordinary Mahalanobis distance in pattern recognition that is defined for each cluster one by one, using a clusterspecific covariance matrix.

Therefore, we require an optimization over all the clusters in the training data. Generally, data in the clusters are distributed as in Fig.1(a), comprising ellipsoidal forms that have high covariances for some dimensions and low covariances for other dimensions.

Further, the cluster is not usually aligned to the axes of coordinates. When we find a global metric matrix M that minimizes the cluster distortions, namely, one that reduces high covariances and expands low covariances for the data to make a spherical form as good as possible in the $M^{1/2}$-mapped space (Fig.1(b)), we can expect it to capture necessary and unnecessary covariations and correlations on the features, combining information from many clusters to produce a more reliable metric that is not locally optimal. We will find this optimal M below.

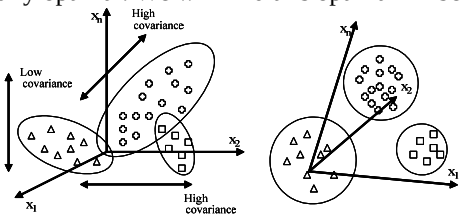


Fig.1. Cluster Geometry of Feature Space

3.2 Optimum Metric Matrix Learning

As we define the distance among classes, gravity method was chosen to perform the sum of simples in each class.

Suppose that R^n is divided into K clusters ($X_1,...,X_K$) for each cluster X_i , cluster centric c_i is calculated as

$$c_i = \frac{1}{X_i}\sum_{x \in X_i} x \qquad (x = 1,2,3,...,K) \tag{6}$$

Where X_i denotes the number of data in X .

We seek the metric matrix M that minimizes the distance between each data x and the cluster centroid c_i , $d_M(x,c_i)$ for all clusters X ($x \in X_i$). Mathematically, this is formulated as a quadratic minimization problem:

$$M = \arg\min_M \sum_{i=1}^{K}\sum_{x_j \in X_i} d_M(x_j,c_i)^2 \quad = \arg\min_M \sum_{i=1}^{K}\sum_{x_j \in X_i}(x_j - c_i)^T M(x_j - c_i) \tag{7}$$

Subject to a constraint:

$$\det(M) = 1 \tag{8}$$

The constraints are shown in [8]. Without any constraints, the zero matrixes would give the minimum.

We expand the Equation (7). Then, we have

$$\sum_{i=1}^{K}\sum_{x_j \in X_i}(x_j - c_i)^T M(x_j - c_i) = \sum_{i=1}^{K}\sum_{x_j \in X_i}\left[\sum_{k=1}^{n}\sum_{l=1}^{n}(x_{jk} - c_{ik})m_{kl}(x_{jl} - c_{il})\right] \tag{9}$$

where $x_j = [x_{jk}]$, $c_i = [c_{ik}]$ and n is the dimension.

And from the constraint (8), for all k

$$\sum_{l=1}^{n} (-1)^{k+l} m_{kl} \det(M_{kl}) = 1 \tag{10}$$

Therefore

$$\sum_{k=1}^{n} \sum_{l=1}^{n} (-1)^{k+l} m_{kl} \det(M_{kl}) = n \tag{11}$$

where M_{kl} denotes an adjuvant matrix of m_{kl}.

By introducing the Lagrange multiplier λ, we have

$$L = \sum_{i=1}^{K} \sum_{x_j \in X_i} \left[\sum_{k=1}^{n} \sum_{l=1}^{n} (x_{jk} - c_{ik}) m_{kl} (x_{jl} - c_{il}) \right] - \lambda \left[\sum_{k=1}^{n} \sum_{l=1}^{n} (-1)^{k+l} m_{kl} \det(M_{kl}) - n \right]$$

Differentiating by m_{kl} and setting to zero, we obtain

$$\frac{\partial L}{\partial m_{kl}} = \sum_{i=1}^{K} \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il}) - \lambda (-1)^{k+l} \det(M_{kl})$$

Let us define $\frac{\partial L}{\partial m_{kl}} = 0$ .Then,

$$\sum_{i=1}^{K} \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il}) = \lambda (-1)^{k+l} \det(M_{kl})$$

Therefore,

$$\det(M_{kl}) = \frac{\sum_{i=1}^{K} \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il})}{\lambda (-1)^{k+l}}$$

The inverse matrix $M^{-1} = [m_{kl}^{-1}]$ can be represented as:

$$\begin{aligned} m_{kl}^{-1} &= \frac{(-1)^{k+l} \det(M_{kl})}{\det(M)} = (-1)^{k+l} \det(M_{kl}) \\ &= (-1)^{k+l} \frac{\sum_{i=1}^{N} \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il})}{\lambda (-1)^{k+l}} \\ &= \frac{\sum_{i=1}^{N} \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il})}{\lambda} \end{aligned} \tag{12}$$

Let $A = [a_{kl}]$ be the matrix $a_{kl} = \sum_{x_j \in X_i} (x_{jk} - c_{ik})(x_{jl} - c_{il})$.

Form Equation (12), we have $A = \lambda M^{-1}$, then, $\det(A) = \lambda^n \det(M^{-1}) = \lambda^n \lambda = [\det(A)]^{\frac{1}{n}}$.

Therefore

$$M = \lambda A^{-1} = [\det(A)]^{\frac{1}{n}} A^{-1} \tag{13}$$

4. Experiments

4.1 Experimental Setup

Based on our proposal, a working system for CBIR has been established. To date, we have tested our CBIR system on a general purpose image database with 51,138 images from COREL CDs. These

images have 60 categories with 100 images in each category. Every category represents a different semantic topic, such as building, mountain, beach, dog, and horse, etc. Two image features, color and texture, are used for image retrieval. The color feature is a 48-dimensional vector generated by QBIC [9]. The texture feature is a 48-dimensional vector generated using the algorithms proposed in [10].With these two types of features, 96-dimensional feature database is used for retrieval in our experiment.

After then, we chose 369 images of the most representative from the CDs. And these images with complex background and composition. For example, image of snow mountain and a lake, image of grasslands, bronze statue and trees, etc. In general, training cluster doesn't take into account the cluster structure in the features in the extraction of images, where neither extra weighted, nor does it consider the diagonal relationship of features, the positive and negative weighted are usually even distributions. Actually, diagonal element is able to reflect the main body from features. The learning clster of "Fireworks" shown in Fig.2, which indicates the distribution of the four image feature. In addition, from the distribution diagram, concentration of feature can faithfully reflect the description and subject of image of the class.

Fig.3 shows that one of the learning claster of metric matrix with the algorithm, and by diagonal element, the main body of feature and the relationship of claster can be enhanced.

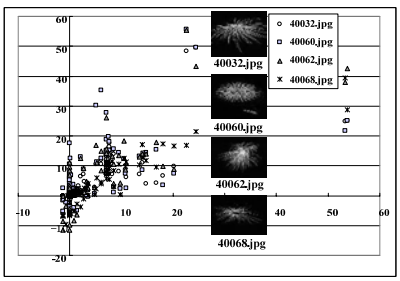


Fig.2. Date distribution of "fireworks" cluster

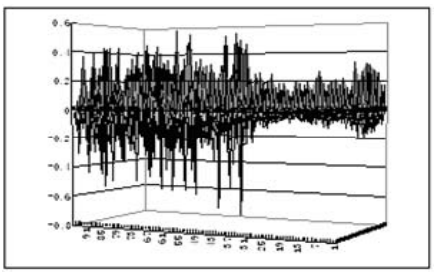


Fig.3 .the diagonal distribution of feature of one cluster

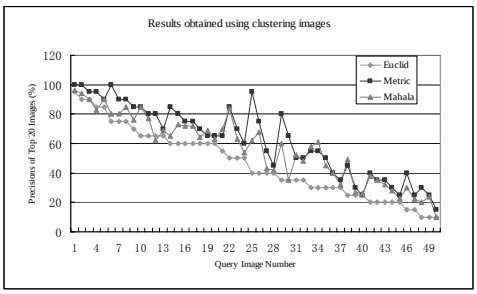


Fig.4 Precisions of in-class at various distance metric

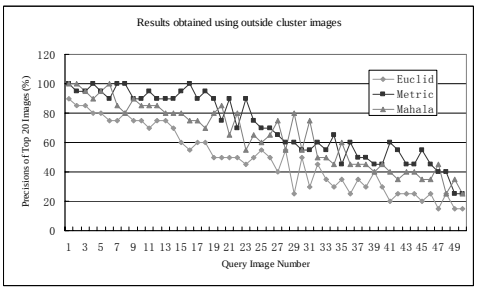


Fig. 5 Precisions of out-of-class at various distance metric

4.2 Experimental Results

For evaluating the performance of the algorithm, 100 images of randomly selected are chosen from the database as the test query images due to their clear semantic meaning, including 50 images in learning class and 50 images in out-of-class. The query image was selected from the database or the test query images and it would be the first image in the result list. Other images in the result list were retrieved and ranked based on the similarity to the query image. Based on precision and recall, we adopt Euclidean distance, Mahalanobis distance and presented algorithm to a query, respectively. The performance of combining methods are similar, though "Rank Euclidean" is a little behind. Fig. 4 shows that comparative precisions in aspect of learning class. Figure 4 shows that the comparison of precisions to qery of out-of-class. From comparison of answer set, it is indicated that retrieval

accuracy of the classified image is higher than out-of-class. More important, the retrieval results using learning metric matrix are better than unused.

Thus, we find in our learning set those images that are similar to the focal image according to at least one elementary distance measure. For each of the M metric matrix measures, we find the top K closest images. For example, “apple”, ”red rose”, “dinosaur”, “playing cards”, “surfing” ,etc. in the learning class images and “red flower”, “sun”, “man”, “aircraft”, “bud”,etc. in the out-of-class images. If all K images are in-class, then we find the closest out-of-class image according to that distance measure and make K triplets with one out-of-class image and the K similar images. Especially these query images with complicated background, general methods tend not to be adequate in retrieval, but our algorithm's performance in this test to emphasize more.

5. Conclusions

We proposed a metric matrix that is useful for image retrieval where Euclidean distance and mahanalobis distance has been used. Through our theoretical analysis and experimental results, we conclude that EUD and CAD are similar when applied to high dimensional NN queries. This distance is optimal in the sense of quadratic minimization over all the clusters in the training data. Through our theoretical analysis and experimental results, we conclude that improvements over Euclidean distance and mahananobis distance, with a significant refinement with tight training clusters in image retrieval.

In future work, we plan to extend our geometrical model to analyze other distance measures, such as the Manhattan distance.

References

[1] S. C. H. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, Learning distance metrics with contextual constraints for image retrieval, *In Proc. Computer Vision and Pattern Recognition*, 2006.

[2] N. Shental, T. Hertz, D. Weinshall, and M. Pavel, Adjustment learning and relevant component analysis, *In ECCV*, pp. 776-792, 2002.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra and I. S. Dhillon, Information theoretic metric learning, *in Proc. Int. Conf. Mach. Learn., Corvalis, Oregon*, pp. 209-216, 2007.

[4] L. Yang, R. Sukthankar, and S. C. H. Hoi, A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32(1), 2010.

[5] Jie Chen, Ruiping Wang, Shiguang Shan, Xilin Chen, Wen Gao, Isomap Based on the Image Euclidean Distance, *The IEEE 7th international conference on pattern recognition (ICPR2006).*, pp. 1110-1113, 2006.

[6] Liwei Wang, Yan Zhang, Jufu Feng, On the Euclidean Distance of Images, *IEEE Transactions On Pattern Analysis and Matching Intelligence*, Vol.27(8), pp.1334-1339, 2005.

[7] Evgeniya Balmachnova, Luc Florack and Bart ter Haar Romeny, Feature Vector Similarity Based on Local Structure, *SSVM 2007, LNCS 4485*, pp. 386–393, 2007.

[8] Y. Ishikawa, R. Subramanya and C. Faloutsos, MindReader: Querying Database Through Multiple Examples, *Proc. 24th Int. Conf.Very Large Database*, pp.218-227, 1998.

[9] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Pektovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content using color, texture, and shape. *Proc. of SPIE Storage and Retrieval for Image and Video Databases*, pp.173–181, 1993.

[10] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of large image data. IEEE Transactions on PAMI, 18(8): 837–842, 1996.

Integrated Test Framework Model for E-business Systems

Pasha Vejdan Tamar[1,a] , Dr. Abbas Asosheh[2,b] , Hourieh Khodkari[3,c]

[1]Department of Information Technology Faculty of Engineering Tarbiat Modares University, Tehran, Iran

[2] Faculty of Engineering Tarbiat Modares University, Tehran, Iran

[3] Faculty of Engineering University of Tehran, Tehran, Iran

[a]pvejdan@modares.ac.ir , [b]asosheh@modares.ac.ir , [c]khodkari@ut.ac.ir

Keywords: E-business Test Framework, Test Case, Test Bed, Interoperability Test, Conformance Test

Abstract. There is an increasing need to an integrated test framework to do conformance and interoperability testing in all layers of e-business standards without dependency on a specific standard. In this paper is shown that, abstracting of test scenarios in a modular manner makes them easy understandable and more reusable. Also embedding a test case tool in the test framework provides capability of automatic generation of executable test cases, and simultaneously makes them more manageable. A modular test bed design with considering some interfaces to pluggable adaptors and applying event driven execution model, make it configurable and applicable to the various test types. With embedding management components into the test bed, it provides more controlling and monitoring over the test and its steps, which they are functional requirements of test beds now.

Introduction

Also, there are test frameworks such as TestBATN[1][3,10,11,12] that could be used in domains further than their original domains. Also OASIS TaMIE[2] Technical Committee, respectively by representing eTSL[3][4], eTSM[4] , and recently Xtemp[5][13] with an event-driven execution model, and a XML based scripting markup for test cases, has provided appropriate background to fulfillment integrity in the e-business test frameworks. Execution model of Agile Test Framework (ATF)[8] and NIST[6] Athena TestBed are based on eTSL v0.78.

Also, from 2007 an international activity started in CEN[7], in order to creation of a Global Interoperability Test Bed for e-business systems. This activity is supported by most of e-business test correlated organizations. The purpose is to establish a base for a test framework (i.e. GITB[8]) to effective development of global distributed e-business test beds. From 2008 October the activity has continued in form of a project with three phases: Feasibility Study, Conceptualization of the target architecture, and Realization. The project first phase has been ended in December of 2009 and its final document is published in February of 2010 entitled 16093 CWA[9], and from January 2011 the

[1] Testing Business, Application, Transport and Network Layers
[2] Testing and Monitoring Internet Exchange
[3] Event-driven Test Scripting Language
[4] event-driven Test Scripting Model
[5] XML Testing and Event-driven Monitoring of Processes
[6] National Institute of Standards and Technology
[7] European Committee for Standardization (CEN)
[8] Global Interoperability Test Bed
[9] CEN Workshop Agreement

project has entered its second phase. In GITB feasability study requirements for an integrated test bed has been estimated and classified in 1)Business-level, 2)Engineering-level, and 3)Operating environment requirements. The engineering-level requirements directly affects test framewok design, and we have leaf-levels of these requirements in Appendix [1,2,5].

In this article, design factors of an integrated test framework are obtained from features of the latest test frameworks which are used in more domains testing, and by referring to the first phase of GITB project requirements and commendations. Then a conceptual high level model for an e-business test framework is developed based on those factors. This model is not hard-code to a specific standard at any layer, and is capable of handling testing activities at all layers of the interoperability stack.

Test Case Model

To fulfill engineering level requirements of e-businesse Testing, in the CWA 16093 under the suggestion [GITB Requirements 1] has said: "… a new test framework should be developed to satisfy the requirements extracted from the three use cases and to enhance the reuse of existing capabilities."[1]

In the Testing literature, typically parts of a test framework are consisting of two general categories of test case and test bed. A test case is a set of steps, variables, and conditions under which the test user verifies the function of SUT. A test bed is a set of facilities that test user by using it tests SUT(s). Therefore, the model of test framework is combination of two test case and test bed models. Also test cases design is prior to test execution model because “most functional requirements for test execution depend on the functional requirements of the test case design. For example, the definition of “capability of test preparation and setup” for test execution depends on how “test configuration information” is represented. Therefore, the structure and grammar necessary to represent a test case should be considered first."[1]

There are some problems that make the process of test cases production to be complex, lengthy, costly, difficult to maintain, Low reusability, not allowing customization and applying the test runtime data. These problems could be resolved by applying the design factors such as *abstracting*, *modularity*, *event-driven*, and *automating* of test case generation and management process.

Test Case Structure (Layers and modules): The layering in the test case producing is separation it into two *abstract test case* and *execution test case* layers. This is an issue that currently has been used by TestBATN and ATF test frameworks. Abstract layer in the TestBATN framework is called *test scenario*. Abstract test suite is made from base of standard and application specification in interaction with domain expert and test user. So an *abstract test suite* could be read and understand by domain expert and test user. Therefore, it's based on specification and independent of any test bed. An abstract test suite could be converted by a test case generation tool or test case developer, to scripts which are executable by test bed. These scripts are said *execution test suite*. In the [GITB Requirements 2-3] suggestion, advantage of this type layering is expressed as: 1) an abstract test case can be developed in a more generic manner without considering test bed system details, and 2) many test cases used for existing test frameworks may be executed by a GITB test bed as an executable test case [1]. Also abstracting makes automation and management of test case production process easier by a tool.

A test case is including information about procedures, assertions, and environment of test. Breaking down each of this information into separate modules, make them reusable and easy to maintenance. Procedure module in the abstract test case includes the partners’ life cycles and actions during testing. Actions are abstract descriptions and contain no message instances. On the other hand, the procedural script in the executable test case represents a business transaction that will be executed and contains specific instances and references to the actors in the business process [7]. Before verification of a test item such as document or message against an assertion, there must be provided some conditions. Consequently, verification scripts may be reused readily within a new testing procedure because the verification script is independently executed by the events during the

test procedure [7]. Test environment information introduce and identifies test participants, services that each participants provides or uses, and specifies type and format of messages used by each services. The abstract test case doesn't have information of specific participant or message instance. But runtime information include test harness based on specific participant instance information, specific messages instance created based on message templates and test users information.

Designing test cases in two abstract and executable layers, indeed is applying [GITB Requirements 2-3] suggestion and making them modular in the above manner is applying [GITB Requirements 2-4] suggestion of CWA 16093. Thereby [Fun-TCM/R01], [Fun-TCM/R02], [Fun-TCM/R03], and [Fun-TCM/R05] requirements of that documents are also addressed and from that also some of test execution requirements will be affected.

Executable Test case structure and scripting grammar: The execution test cases are conversion of abstract test cases, so that can be executed by the test bed. It is necessary therefore, their structure and syntax to be standard, because only the standardized test cases can be identified, interpreted and executed by means of a test bed. Also, it provides the possibility of test repeating, and the portability from a test bed to the others. Hence a test suite is generally written in XML scripts. It is according to [GITB Requirements 2-2]**.**

CEN 16093 CWA Report, suggests the event-driven scripting model (eTSM) developed by OASIS TAMIE TC as a candidate approach to standardization. Recent work of this committee is provided in title of Xtemp [13]. Xtemp has simple structure, and limited but sufficient number of instructions, so its implementation is easy for extension and evaluation. Its execution model is event -driven and independent of time that makes test scripts equally applicable to the both real-time and deferred events validation. Coordination of test case execution into the test suite also test workflow status has shown as events. This makes easier management of the large test suite. It is independent of any platform, and protocol.

Using Xtemp, enhances the satisfying of engineering requirements stated in the abstract layer, specially [Fun-TCM/R02 -1,2,3,4,5] and [Fun-TCM/R02 -1].

Test Case Generation and Management tool: Using Test Case management tool is earlier carried in TestBATN test framework as Test Design GUI, which is used to dynamic defining of Test scenarios For creation corresponding TDL[10][10], [11], [12]. Also NIST has implemented a test case generation method by using a tool in healthcare domain that the essential goal for it is to facilitate specification, generation, and traceability of test cases [7].

A test cases generation and management tool, leads to automation of test case generation process; and makes possible development, implementation, and maintaining of test cases; auxiliary materials of different standards; and grouping them in test suites. It also makes capability of test case customization and applying user actual data, creating message instance from message template.

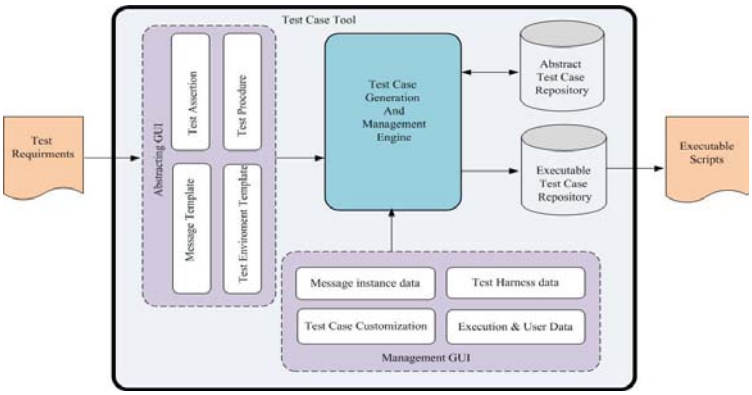


Fig.1. Test Case Tool

[10] Test Description Language

As you see in Fig.1, in test case tool, by using designed abstracting forms instances, such as abstract test rule form in represent of business process, message template form, environmental information template form, and test abstract assertion form in the abstract test case specific GUI, test cases requirements are defined and saved in the abstract modules corresponding tables in the abstract test cases repository. These abstract test cases which created for a standard or a type of test could be reused in generation of executable test cases for different test models and test conditions.

Test case tool makes possible control and changes in test steps, this helps to test control and running in more or less then its prospected steps. Using test case tool helps to fulfillment of all GITB engineering functional requirements, specially [Fun-TCM/R04] and [Fun-TCM/R05], and it also helps to fulfillment of non-functional requirements modularity and extensibility in test cases.

Test Execution Model

Execution model is architecture of testing tools to execute a test. This model shows the test bed components and their orchestration together in interact with the test environment to fulfillment of test process. The integrated test bed execution model wants to show that, it is able to testing of any system under any standard and in any interoperability layer, and with providing test management capabilities can rich testing process. It requires existence of *pluggable* and *Plug-n-Play* test components. Pluggable components provide capability of testing different standards, also Plug-n-Play feature makes it easy to setup test bed and automation of test process. *Modular* design provides better resource management and reduces maintenance cost and increases reusability. Also an *event driven* structure causes a test bed with agility manage different resources, because components can interact with each other without a direct connection between themselves. "All of the activities of the pluggable components and the test infrastructure are coordinated via events [6,9]".

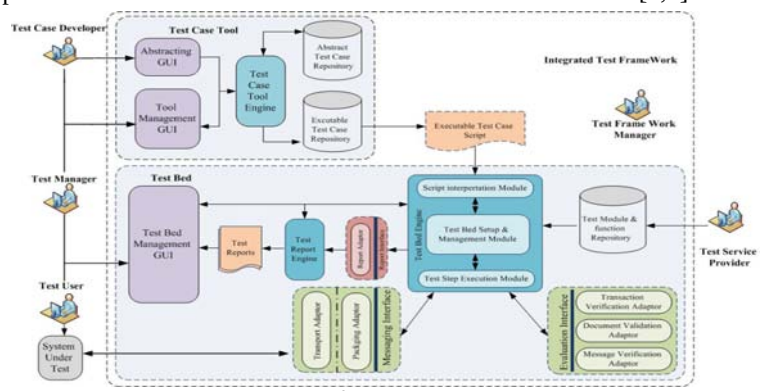


Fig.2. integrated test framework conceptual model

An integrated test framework conceptual architecture model could be as Fig.2. Its components descriptions are as following:

Test bed infrastructure: Test engine is the test infrastructure. Its structure has considered similar to ATF [8,9] , with some extended managerial duties. Test engine, is the test bed brain that controls the entire test process. It reads and interprets a test case, coordinates interfaces and other test components, conducts the execution of test with the SUTs which should be tested, and it also interacts with management components, provides the test management. The test engine has the following three major internal modules:

- **Test Script interpertation module**: This module reads presented executable test case and then separate it to test procedures for defining test execution steps; test assertions to evaluate the messages, documents, and transactions; and configuration information to set up a test-bed and discovering and plugging appropriate test interface adaptors; and finally And puts them to the Test Bed Setup and Management Module.
- **Test Bed Setup and Management Module**: This module allows executable test case to the

interpreter module, and uses the configuration profiles provided by it, to discover and connect appropriate test adaptors. It defines a process model to test, from test procedure, discovered instances of pluggable adaptors, management components, test engine, their messages and calling sequence; and then puts it to Test Steps Execution Module. Also this module in interaction with test users and manager through Test Bed Management GUI provides controlling of test steps and test effective management.

- **Test Steps Execution Module:** this module interprets the Test Bed Setup and Management Module defined process and execute each activity of process by calling a function

Test Adaptors Interfaces: Test adaptors interface provides test bed interactions with the SUT and test pluggable module. Test engine places and calls them dynamically according to configuration specifications. A test bed could have multiple adaptor instances to each interface, and during test execution inside test service, at each step recall appropriate instance. By designing these interfaces, [GITB Requirements 3-1] from CWA16093 report is used, and it leads to fulfilment of [Fun-TCE/R03] to [Fun- TCE/R05] engineering-level requirements.

there are defined the following interfaces for an Integrated Test Framework:

- **Messaging adaptors interfaces**

TestBATN test framework [11] has broken messaging adaptors interface functionality in two interfaces that, it makes their corresponding adaptors more reusable and lightweight

  - **Transport adaptors interface:** this interface facilitates using adaptors for receive or sending messages by protocols such as TCP, HTTP, SMTP and etc...
  - **Pack/Unpacking adaptors interface:** this interface provides using adaptors which used in packing and unpacking messages according to higher layer communication protocols such as SOAP or ebMS.

- **Verification adaptors interfaces**

This interface is designed for plugging test verification and evaluating adaptors to test bed, and general classification these adaptors can be as follows. This type classification is to show verification coverage of an integrated test bed, and that many of them could be plugged and used simultaneously by a test bed.

  - **Content validation adaptor:** these type pluggable adaptors are used to validation contents against a schema, to generate a verdict and a structured test report, about has down validation. E.g. XML schema validation adaptors or Schemata valuator.
  - **Message verification adaptor:** these type pluggable adaptors are used to perform complex testing over any contents of messages. And its example is XPATH verification adaptors.
  - **Transaction verification adaptor:** this type adaptor is designed to verify transactions of a process. In this time there aren't any well-known this type adaptors, and it is one of main weakness of existing test beds, that is emphasised in CWA 16093 report.

TestBed Management Components: Management components have been embedded in test bed to make rich controlling and managing test execution by test manager, and helping test users in preparing various reports to identification the faults in their systems.

- **Test bed management GUI**

The idea of this component is driven from TestBATN framework [11], but defined more functionality to it to cover testing requirements. This component is interface between Test Bed Setup and Management Module and Test Report Engine with test manager and test users for supervision on test execution,controlling test steps, providing graphical facilities, managing test bed resources and required test reports.

- **Test Report Engine and its interface**

Report engine is to providing test steps verifications and final verdict, showing accurately error location, providing Test Log report and analytical reports for test users . Report engine uses test transactions event-board to produce necessary test reports. This engine also have interface to several report supporting adaptors for different report purposes according to test use cases. Test Report engine allows this output to the test users through Test Bed Management GUI.

Conclusion

In this paper, an integrated test framework was interduced to e-business systems interoperability testing. First step was to model test cases, using abstracting and modularity factors. Test cases were considered in two abstract and executable layers. Abstract layer is to making a common understanding among test agents, and also to providing suitable context for automatic generation and management of test cases by a tool. To enhance reusability test cases are separated into three modules, test procedure, test assertion, and test environment. Using test case tool speeds up and facilitates test cases generation process, and provides capabilities such as: test cases customization, and applying test users data to the test cases.

Then modularity and event-driven factors were applied to the test execution model. Test Engine component as an infrastructure interprets test cases, configures test bed, and manages test execution. Test interfaces define high level functionalities and requirements about corresponding adaptor instances. This makes it possible to implement an interface in different ways, and also imply test bed to testing different standards or specifications and test requirements in e-business different layers. Also a Test Bed Management GUI in interaction with Test Engine makes capable more controlling and monitoring over test steps and process. Test Report Engine and Test Report Interface provide detailed and analytical reporting capability from different test usage.

References

[1] CEN: Feasibility Study for a Global eBusiness Interoperability Test Bed(GITB), ftp://ftp.cen.eu/CEN/Sectors/TCandWorkshops /Workshops/CWA16093TestBed.pdf (2010)

[2] CEN: Terms of Reference for the GITB–Phase 2 Project Team, http://www.ebusiness- testbed. eu/dynamics/modules/SFIL0100/view.php?fil_Id=1010 (2011)

[3] Dogac A., et al: Electronic Health Record Interoperability as Realized in Turkey's National Health Information System, Methods of Information in Medicine Volume: 50, Issue: 2, Pages: 140-149 (2011)

[4] Durand, J. OASIS ebXML IIC TC.: Event-driven Test Scripting Language. http://kavi.oasis-open.org/committees/download.php/22445/eTSL-draft-085.pdf.(2007)

[5] Ivezich N., et al, : Towards a Global Interoperability Test Bed for eBusiness Systems, Proceedings of the 2009 eChallenges Conference Istanbul (2009)

[6] Ivezich N., Woo J., Cho H.: Towards Test Framework for Efficient and Reusable Global e-Business Test Beds, In Proceedings of I-ESA 2010 Conference, Coventry, UK, (2010)

[7] Ivezich N., Woo J.: Testing Interoperability Standards – A Test Case Generation Methodology, proceedings of the international conference on interoperability for enterprise software and applications A-ESA (2010)

[8] Woo J.: Agile Test Methodology for B2C/B2B Interoperability ,Department of Industrial and Management Engineering Pohang University of Science & Technology (2007)

[9] Woo J., Ivezic N., Cho H.: Agile test framework for business-to-business interoperability, Springer Science+Business Media, Inf Syst Front, DOI 10.1007/s10796-011-9303-3 (2011)

[10] Namli T., el at: Testing the Conformance and Interoperability of NHIS to Turkey's HL7 Profile, 9th International HL7 Interoperability Conference (IHIC), pp. 63-68. (2008)

[11] Namli T., Aluc G., Dogac A.: An Interoperability Test Framework for HL7-Based Systems, IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 13, NO. 3 (2009)

[12] Namli T., Dogac A., Sinaci A., Aluc G.: Testing the Interoperability and Conformance of UBL/NES based Applications, Middle East Technical University (2009)

[13] OASIS TAMIE: XTemp- XML Testing and Event- driven Monitoring of Processes, http://docs.oasis-open.org/tamie/v1.0/200906/xtemp-1.0-csd01.pdf (2011)

Appendix: GITB Engineering Level Functional Requirements [1]

Key Capability Index for Engineering Level Functional Requirements	
Test Execution Model	**[Fun-TCE/R01] Capability of test preparation and setup**
	1) Capability of providing the setup information to SUT(s)
	2)Capability of requesting SUT's parameters and information
	3)Capability of test case customization
	4)Capability of configuration of setup information
	[Fun-TCE/R02] Capability of controlling test steps
	1)Capability of display of test flow and test progress
	2)Capability of requesting/storing user's information
	3)Capability of binding user's information into test case
	4)Capability of manual execution of test steps
	[Fun-TCE/R03] Capability of message exchange
	1) Capability of sending/receiving message payloads?
	2) Capability of uploading/downloading message payloads
	3)Capability of capturing message
	[Fun-TCE/R04] Capability of message pre/post-processing
	1)Capability of decomposing message
	2)Capability of retrieving the value from message
	3) Capability of generation message template from schema
	4) Capability of generation test data for a specific message template
	5) Capability of message transformation
	[Fun-TCE/R05] Capability of message pre/post-processing
	1) Capability of detecting unknown problems
	2) Capability of employing the existing validation engines
	3) Capability of recovery from errors
	[Fun-TCE/R06] Capability of reporting
	1) Capability of display of error location
	2) Capability of display of test log information
	3) Capability of display of the detail test result
	[Fun-TCE/R07] Capability of B2B system emulation (optional)
	1) Capability of emulation of an arbitrary unit
Test Case Design	**[Fun-TCM/R01] Capability of representing test configuration information**
	1) Capability of representing declaration of messaging protocol to
	[Fun-TCM/R02] Capability of representing test procedural information
	1) Capability of representing message to be sent
	2) Capability of representing message choreography
	3) Capability of representing conditional expressions (test step) for test case
	4) Capability of representing iterative expression (test step) for test case
	5) Capability of representing manual steps
	[Fun-TCM/R03] Capability of representing test verification information
	1) Capability of using external documents for verification(e.g. XML Schema)
	[Fun-TCM/R04] Capability of representing test suite which contains a set of test cases
	1) Capability of representing precedence relationships between test
	[Fun-TCM/R05] Capability of representing test data
	1) Capability of representing of user's defined values
	2) Capability of representation of automatically generated values (i.e. using metadata)

Design and Enhancement of Mandarin Emotional Speech Database

Liqin Fu [a], Hongli Jin [b], Xinjie Wu [c]

Department of Mechanical & Electrical Engineering, Beijing Institute of Economic Management, Beijing, 100102

[a]liqin_fu@126.com, [b]Jinhongli@biem.edu.cn, [c]Wuxinjie@biem.edu.cn

Keywords: Mandarin Emotional speech database; naturalness; Assembling; Sharing; Description of database

Abstract. Progress in both speech emotion recognition and emotional speech synthesis relies heavily on the development of appropriate databases. This paper addresses the main issues that need to be considered in developing Mandarin emotional speech database. It shows what the challenge is for current databases and indicates the future directions for the development of Mandarin emotional speech databases. Firstly we should mainly collect speeches that are genuinely emotional rather than acted or simulated. Secondly data assembling from diverse sources is an effective solution to improve the scale and quality of speech database and sharing strategy should be established. Thirdly the material collected should be audio–visual rather than audio alone. Finally perfect database description is essential. And it discusses the ways of database description in detail.

Introduction

Research on speech emotion plays an important role in human–computer interaction. Accurate detection of the affective state from speech provides clear benefits for the design of natural human–machine speech interfaces. For the growing emphasis on applications in the area—the synthesis of emotional speech and recognition of emotional speech，the developing of the scale and quality of emotion speech databases is an urgent need. Several efforts about speech emotion are relevant to data collection. The easiest way to collect emotional speech is to have actors simulate it which has been adopted by most of the native speech communities. In addition, these datasets are comparatively small-scale collections of material, typically created to examine a single issue, and not widely available.

This paper analyzes the character of existing projects and discusses the strategies which can be adopted by new database projects and the way assembling databases.

Key issues

There are many problems surrounding database development and four main problems need to be considered in developing a database.

The first issue covers several kinds of variation that a database may incorporate, Including language spoken; type of dialect (e.g. standard or vernacular); number of different speakers; gender of speakers; types of emotional state considered; tokens of a given state; social setting. These kinds of variation are potentially important for any attempt to generalize.

To understand or to match human performance, databases that contain evidence on the way vocal signs relate to their context are needed. So the second issue is about context. For acted speech which is the main part of databases adopted to verify speech emotion technique, it is doubtful whether it

captures subtler aspects of contextualization in naturally emotional speech. Four types of context can be distinguished--semantic context; structural context; inter-modal context and temporal context [1].

The third issue is about the naturalness of the speech material. It is certainly true that good actors can generate speech which can be classified reliably by listeners. But there are many differences between acted and natural emotional speech. It leads to that an algorithm verified by acted speeches can not achieve good result when handling real speeches. For instance, acted speech is often 'read' but spoken, and read speech is well known to have distinctive characteristics [2].

The fourth issue is description of database. Constructing a database requires techniques for describing the linguistic and emotional content of speech.

Present of Mandarin Emotional Speech

With the development of mandarin emotional speech research, there appear some databases for mandarin emotion recognition. Most of them consist of acted speeches, only several ones include natural speeches. Moreover, their scale is small relatively, as shown in Figure 1.

Table 1: Mandarin emotional corpus

Database/Community	collection mode	Speakers	Emotion states
CASIA/Chinese Academy of Sciences	acted speech	4 (2 males,2 females)	happiness, sadness, anger, fear and neutral
Southeast University	acted speech Induced by computer game	10 (5 males, 5 females)	3 (boring, nervous and happiness)
BUMEC /Beihang University	acted speech Induced by text	15 (7 males, 8 females)	7 (anger, disgust, happiness, surprise, sadness, fear and neutral)
CSED /Tsinghua University	acted speech conversation		5 (anger, impatience, neutral, joy, and happiness)
MASC /Zhejiang University	acted speech	68(23 female and 45 male)	5(neutral, anger, elation, panic and sadness.)
CASS_ESC/Chinese Academy of Social Sciences	acted speech	4	7 (sneer, happiness, fear, anger, sadness, disgust and neutral)
CASS_EXP/Chinese Academy of Social Sciences	broadcast and television		
CASS-IBM-ESC/Chinese Academy of Social Sciences	Natural speech (real estate field)		

CASIA Mandarin emotional corpus has been carefully established by Institute of Automation of Chinese Academy of Sciences [3]. It is recorded by two males and two females under studio conditions. The corpus contains 9600 speeches in five emotion states (happiness, sadness, anger, fear and neutral).

The speech emotion database established by Southeast University contains speeches recorded by 10 students (5 males and 5 females) whose ages are between 20 and 30. Emotion states (boring, nervous and happiness) in data collection are induced by computer games.

Beihang University mandarin emotion corpus (BUMEC) covers mandarin utterances of six emotions, including anger, disgust, happiness, surprise, sadness, fear and neutral, twenty texts and fifteen actors, seven males and eight females. Each speaker repeats each text three times in each

emotion, meaning that sixty utterances per emotion. For classifier evaluation, all samples have been assessed and speeches which are regarded expressing emotion correctly are selected

Chinese Speech Emotion Database (CSED) is established by Tsinghua University contains 600 utterances in the form of dialogues with 20 emotional variation modes consisting of 5 different emotions including anger, impatience, neutral, joy, and happiness [4].

MASC (Mandarin Affective Speech Corpus), established by Zhejiang University, contains recordings of 68 native speakers (23 females and 45 males) and five kinds of emotional states: neutral, anger, elation, panic and sadness. Each speaker pronounces 5 phrases, 10 sentences for three times for each emotional states and 2 paragraphs only for neutral. These materials covers all the phonemes in Chinese [5].

In most projects, neither the words nor the phrasing are typically chosen to simulate emotional speech.

Strategies

Natural speech collection. Research aimed at recognizing emotion needs databases that encompass as many as possible of the signs by which a given emotion may be expressed. It is certainly true that good actors can generate speech that listeners classify reliably. But that kind of source does not mirrors spontaneous expression of emotion closely.

Naturalness is at the centre of the problem. The essential aim is to collect speeches that are genuinely emotional rather than acted or simulated. Speech material collection should be guided by two principles. First, the material should be derived from interactions and from people who at least appeared to be experiencing genuine emotion. Second, the type of emotional states labeled in material should be same as occur in everyday interactions rather than archetypal examples of emotion (such as full-blown fear or anger).

In spontaneous speech recording, utterances that express emotions in the real-world are recorded. Although this method can obtain speeches which have the best naturalness, there are many inconveniences in the process because we need to follow the speaker. When he/she is in some emotion state, his/her voice is recorded immediately.

Radio and television provide rich sources in chat shows, documentaries, etc. In some broadcast material, it is felt that the speaker was genuinely affected by emotion. And some television programme can present real interactions with a degree of emotional content reliably.

The collection of natural material may leads to lacking of control. For instance, there is unpredictability for emotion of natural speech, so it is a substantial issue to identify the emotion that is being expressed. When recording spontaneous speech, we must hide our recording device in order to avoid the speaker having any pressure to express his real emotion. We also can not guarantee the quietness of environment. In addition, from truly natural speech, it is difficult to achieve phonetically and prosodically balanced data sets needed in some applications (e.g. concatenative synthesis).

Assembling and sharing. For improving the scale and quality of speech database, another answer depends on assembling information from diverse sources. And the value of a database increases enormously if it is available to all the speech researchers.

It is clear that there are challenges in assembling and describing databases of the type that meet the needs we identify. Two main issues must be considered. First, the format of the data files, including raw material (e.g., wav) and the coding of descriptors, should be standard or transparent. When choosing a file format, it should be considered whether these files can provide full enough information about the signal or the details of its collection. The second problem is about ethics and copyright, particularly with natural data. Subjects of natural emotional may object to wide circulation. Moreover, accessing radio or television source may raise serious copyright problems.

Multi-modal collection. In general, vocal signs of emotion form part of a multi-modal signaling system. It is not true that audio and visual channels function independently. So it will be useful to consider speech as only one of several mutually supportive information sources. But the great

majority of the database is purely audio, or only used audio information. Therefore, the material collected should be audio–visual rather than audio alone.

It is known that high rates of emotion recognition are unlikely to be achieved from speech alone in applied settings. So in practical view, we also should consider speech as one of many inputs. Owing to collecting visual as well as audio material, we can get benefit from awareness of visual information when recognizing emotion from speech

Database description. There are two issues need to be solved in speech description. First, it needs to specify the full range of features involved in the vocal expression of emotion. Second, it needs to describe the attributes being relevant to emotion.

Acted material may well be adequately described in terms of category labels such as happy, angry, sad, etc. Some Mandarin Emotion Speech Corpus include the corresponding label files. For instance, the label files of CSED include silence or effective segments, emotional classes, emotional variation segment and emotional quality. 67 normal acoustical features are extracted based on the Praat tool and stored in the database. Two feature analysis methods are employed on the MASC. One of them shows the contours for prosodic features and the other gives a statistical acoustical analysis.

Natural databases, though, are likely to involve gradation in and out of emotional peaks, coincidence of different emotions, and relatively subtle states. A fundamental choice is between categorical descriptors and continuous variables. The relative merits of the two types remain to be resolved.

For large corpora, manual coding is very slow and expensive, and so it is critical that description of emotional speech is automatic—or rather semi-automatic. So a software system for speech description is needed. For each speech, the system constructs description specifying features including at least voice quality, prosody and non-linguistic features such as laughter, crying, etc. In our project, F0 contours, intensity, pause boundaries, high frequency bursts, and basic spectral properties are calculated. The summary of prosodic and paralinguistic features is shown in Table 2.

The description is implemented through three steps. First, it divides the passage into tunes (episodes of speech between substantial pauses) using pause boundaries. Then statistics are derived for each tune and for the passage as a whole. Finally, we obtain a battery of measures which cover properties related to F0 profile, intensity profile, and its spectrum for each unit.

Table 2：Summary of prosodic and paralinguistic features

Feature type	Specific coding
Pitch range	F0 contours (High/low wide/narrow)
Loudness	Loud/quiet, crescendo/diminuendo
Tempo	Fast/slow, ccelerating/decelerating, clipped/drawled
Voice quality	Falsetto, creak, Whisper, rough, breathy, ventricular, ingressive, glottal attack
Pause	Pause boundaries
Reflex behaviours	Clearing the throat, sniffing, gulping Audible breathing, yawning
Voice qualifications	Laugh, cry, tremulous voice
Spectral properties	Formant, LPCC, MFCC, etc.

Conclusion

For real-world applications of speech emotion techniques, it is essential to understand the ways people express emotion which are complex and variable. But the current native databases always use pure simulations which are too far from the reality for application of techniques. So we suggest natural emotional speech collection and database assembling to improve the scale and quality of

database on the one hand, and to consider the whole domain of emotion as well as to focus on a specific sub-region on the other hand.

Moreover, description of emotion plays an important role in database assembling and sharing. For the description of speech, there also appears to be consensus on the need to consider two levels. For linguistic descriptions, standard labeling systems exist and are increasingly becoming automated. For emotion description, we replace a small number of primaries with a new system which includes both a larger number of categories and dimensional description.

References

[1] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, Peter Roach: *Emotional speech: Towards a new generation of databases*. Speech communication (2003), p.33.

[2] Johns-Lewis, C: *Prosodic differentiation of discourse modes*. In: Johns-Lewis, C. (Ed.), Intonation in Discourse. College-Hill Press, San Diego (1986), p. 199.

[3] Jianhua Tao, Fangzhou Liu, Meng Zhang, Huibin Jia: *Design of Speech Corpus for Mandarin Text to Speech.* The Blizzard Challenge 2008 workshop (2008), p.1

[4] Lu Xu, Mingxing Xu: *Chinese emotional speech database for the detection of emotion variation.* Tsinghua Science and Technology, Volume S1 (2009), p.10

[5] Tian Wu, Yingchun Yang, Zhaohui Wu and Dongdong Li: *MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition*. IEEE Odyssey 2006, Speaker and Language Recognition Workshop, June (2006), p.1

Knowledge Management Platform Based on the Environmental Monitoring System with Energy Harvesting Sensor Motes for Tea Farming

Eiji Aoki[1,a], Ken Kudo[1,b], Akira Fukuda[2,c], Tsuneo Nakanishi[2,d],
Shigeaki Tagashira[2,e], Takashi Okayasu[2,f],
Naoyuki Tsuruda[3,g], Satoru Yamasaki[3,h], and Yasuhito Imura[3,i]

[1]Institute for Hyper Network Society, 51-6, Higashikasugamachi, Oita, Japan

[2]Kyushu University, 6-10-1, Hakozaki, Higashiku, Fukoka, Japan

[3]Oita Computer Engineering & Consulting, Ltd., 21-1, Kumano, Kitsuki, Oita, Japan

[a]blue@hyper.or.jp, [b]kudo@hyper.or.jp, [c]fukuda@ait.kyushu-u.ac.jp, [d]tun@f.ait.kyushu-u.ac.jp,
[e]shigeaki@f.ait.kyushu-u.ac.jp, [f]okayasu@bpes.kyushu-u.ac.jp,
[g]turuda@cec-ltd.co.jp, [h]S.Yamasaki@cec-ltd.co.jp, [i]y-imura@cec-ltd.co.jp

Keywords: Agricultural infomatization, Sensor network, Energy harvesting system, SNS, Web

Abstract. In regards to computerization for the field of agriculture, various initiatives are being started using production and distribution as the subjects. However, since it has not been long since this began in reality, there are still insufficient matters that have been resolved in regards to the technical validations and cost effectiveness. Unlike factories, designing and operation of computerization for farms, necessitates pouring in a tremendous amount of expertise. Therefore to making this a possibility is ICT and its related technologies of sensors and energy harvesting systems. In our research and developments, we configure hardware that implement monitoring system as well as the knowledge management platforms that utilize the very data that were collected by them.

Introduction[1]

So far, laborsaving and automation in agriculture are mainly achieved by mechanical devices. However, advancement in ICT can bring them in knowledge works in agriculture. Sensing and sensor network technologies gathering data on crops and their environment are promising technologies. Many researchers have been tackling to introduce them into farms [1-10]. At the same time, there are some barriers to introduce of ICT into agriculture especially in rural areas of Japan and some Asian countries [11].

In Oita prefecture, a production region development contract has been agreed with major beverage manufacturers, and a plan to develop the tea plantations of 100 hectares within prefecture is moving forward. There is now a delivery standard set for the ingredients meant for drinking of pet bottles and cans, and in that standard, the determination of the harvesting time is heavily emphasized, and is evaluated based on the data of climate. This means that grasping the detailed climatic data of the tea plantations will become of utmost necessity.

Although there are services available for wide range fixed point observations including AMeDAS and the MESH climatic data or such, however with these, since the topography of the tea plantations

[1] This work was supported in part by Ministry of Internal Affairs and Communications (MIC) in Japan, Strategic Information and Communications R&D Promotion Programme (SCOPE) No.112310004.

is complex, it is not possible to grasp the minimum atmospheric temperature and such in details with solar radiation, or wind speed, or whenever frost damages occur. Accordingly, we will place weather observation on several points, measure the weather data (atmospheric temperature, wind speed, solar radiation, amount of rainfall, and such), and using a personal computer or cell phones and such, collect them in real time. Also, as data is being accumulated, it becomes easier to assess the harvesting time. Moreover, in addition to being able to clarify the generating mechanisms of a cold wind damage of the winter season, or the frost damage of the spring season, or the drought damage of the summer season, as well as their preventative measures, by using simulations from the climatic data, it will become possible to examine and implement the countermeasures.

Table 1, The proportion of hilly and mountainous areas in Oita prefecture.

District	Cultivated Acreage (hectares)			Proportion	National Rank	Kyushu Rank
	Whole	Hilly Areas	Mountainous Areas			
Fukuoka	70,169	12,910	2,069	21.3%	38	7
Saga	47,379	11,729	546	25.9%	32	6
Nagasaki	33,247	11,984	868	38.7%	25	4
Kumamoto	87,478	29,916	3,818	38.6%	26	5
Oita	**40,849**	**23,113**	**6,401**	**72.3%**	**3**	**1**
Miyazaki	51,234	23,657	6,549	59.0%	8	2
Kagoshima	80,642	38,847	3,111	52.0%	14	3
National	3,693,026	1,027,105	368,817	37.8%	-	-
Kyushu	411,000	152,155	23,361	42.7%	-	-

However, on the semi mountainous areas in Oita prefecture, environment for connecting a broadband or a wide range of the tea plantations network, as well as the power source facilities are not sufficiently established, and the reality is that collecting the climatic data in real time is a difficult situation. With “energy harvest type environmental monitoring system” and “knowledge management platform” of our research and developments, our objective is to supply the relevancy of environmental information and the farm work processes to the farmers, even to these farms that do not have the power sources or the networks established. In other words, by using the current energy harvest model, it will become possible to supply information of the farm work processes that are affected by environmental factors, to every farm (regardless of presence or absence of power sources as well as networks).

Also, SNS feature which is packaged with the platforms, allows speeding up of information sharing between farmers as well as between the farmers and the related parties. For example, the production technological improvement for working groups of “Japan Agriculture” or smooth technology transfer and such from farmer selections to the selected farmers regarding proving test developments for new variety developed by every “Prefectural Agriculture Research Centers”, and will contribute to betterment of the agricultural production system that has become regionally unified into one. As the collected data amount increases, the different analysis with the platform will be possible, and it becomes possible to form countermeasures that are appropriate towards matters that are regionally native amongst the linked facilities. With this, a new public awareness of new technology for the particular region is formed, turning into training of the people, and will enhance motivation from an individual level or as a society. Moreover, with the technology and technical expertise acquired from our research and developments, we can expect the effects of being able to performing the acquisition of technology within a short period of time for agriculture, through various ways such as in the selecting of the suitable spots as the tea plantations within prefectural expands, or ICT utilization for the agricultural production, or for training manuals for the new participants.

Research Subjects

With all expertise and subjects related to the operations and management of the environmental monitoring system and developing and supplying of tools that utilize collected data, that have been acquired from past investigative researches and proving tests, we will definitely actualize and by adding energy harvesting related technology, also enhance functionality. Moreover, we will apply the technological seeds that exist in universities and "Prefectural Agriculture Research Centers", and solve the problematic points of the current agricultural helping systems that have actualized, using ICT and the networks. In our research and developments, we design and develop the platform that makes the data acquisitions and expertise to "becoming visible" and we improve the system that materializes these types of environments, and we perform those functional verifications empirically on the farms.

Experiments for Evaluation of Adequacy. In regards to our research and developments, at "Kyushu University", a member of our research project, in addition to developing environmental monitoring system based on sensor networks that applies ICT and sensing technology as well as developing a system that displays environmental information on the web, and using these, with the cooperation of agricultural corporations, individual farmers, and agriculturally related research and development institutes and such, they have also been accumulating proving tests that are used for evaluating validities of these and have been producing various results and subject matters. Fig. 1 will display the system configurations, and web display screens of the proving tests.

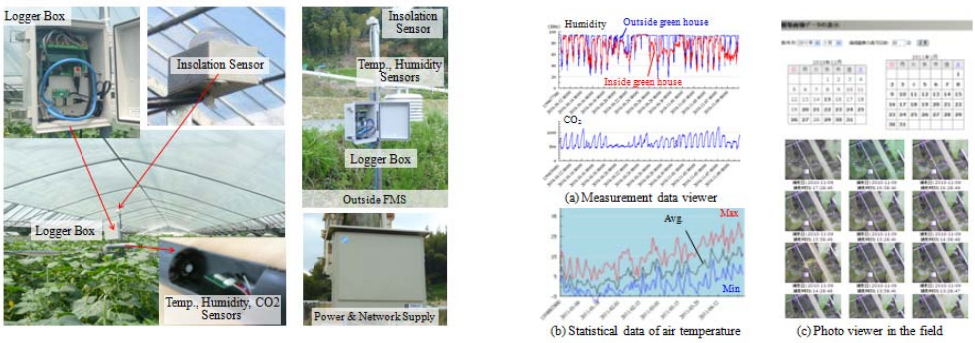


Fig. 1, These photos are environmental monitoring system in-greenhouse, the sensor stores in a logger box. The sensing data will be viewing as environmental information on the PC screen. The source of this figure is from Mr. Okayasu, T. Kyushu University.

Case of High Quality Strawberry Production. "Act Group Strawberry Farms Inc" prides itself as having the largest class production area (2.1 hectare, according to a proving test 0.21 hectare in operation) of facility cultivating strawberries in west of Japan. By introducing with the system mentioned a collection of in-greenhouse environmental information, they have been overseeing proving tests for actualizing the cultivation of high quality strawberries and for controlling production periods beginning from 2010, with "Oita Prefecture Research Center". In terms of accomplishments, based on the climatic environmental information inside and outside the greenhouse, we are now able to control the skylight window (timing of open/close and operation time), ventilation (timing of ventilation and operation time), heating appliances (timing of heating and operation time), carbon dioxide generator (release period, fertilization method, and amount of applied fertilizer). Since it is now possible to display environmental information of inside and outside of the greenhouse with a graph, table, camera images, as well as work history (3rd party comments included this time from a research associate "Oita Prefecture Agriculture Research Center") all unto a list, it now possible for data to be shared between farmers and researchers.

Amongst the subject matters, improvements in constructing, relocation, operation, maintenance of the environmental monitoring system, fulfilling the creation of low priced sensors, and cost lowering of the devices, are being requested. Also, since current system assumed the displaying and sharing on desktop computers only, it is extremely difficult to immediately display agricultural information for performing comparisons, verifications, analysis, or sharing, that are all required steps for cultivated fields. It is necessary to develop the information sharing platform that uses a multifunctional portable terminal for smart phones, tablet computers.

Case of Private Own Farms. The installation rate of ICT in regards to privately owned farms are extremely low when compared to industrial and corporate farms. By performing trial installations of the previously mentioned systems to privately owned farms, from 2010 with cooperation of "Japan Agriculture Itoshima", "Itoshima City Council", "Itoshima Agricultural Council for Promoting Industry Academic Government Collaboration", we have been implementing a foundation research that investigates subject matters and solution strategies regarding ICT installations for each management entities. In terms of the achievements, we showed that weather information can be checked at all time and that these can be applied for catching issues during operation and for making decisions during work implementation.

Moreover in terms of subject matters, the 3 points that were brought up include simplification of construction so that even farmers can perform relocation and repairs for the devices, fulfilling the creation of low priced sensors, and cost lowering of the devices.

Case of Open Cultivation. Many of the produce those are cultivated on bare earth in terms of growth and quality are affected by abnormal weather due to global warming and such factors. Particularly in recent years in Fukuoka prefecture, due to high temperature increases during summer season period, a decrease in the ratio of 1st class rice has been continuing. Therefore, from 2005 together with "Japan Agriculture Kasuya", we have been implementing a review of fertilizer management and proper time of harvesting, by using precise measurements of the local climate collected by the environmental monitoring system and the information it provides. In terms of the achievements, while the grades and such of the harvested rice in other regions are experiencing a sluggish business, we were able to actualize an improvement in the ratio of the 1st class rice for year 2009 rice productions and year 2010 rice productions.

Moreover in terms of subject matters, simplification of construction so that even "Japan Agriculture" staff members can perform relocation and repairs for the devices, fulfilling the creation of low priced sensors, and cost lowering of the devices were requested. Also, for the cultivated lands, since securing a power supply facility or a network environment is difficult, a securing of power source through solar power or such, as well as developing a system that can indirectly retrieve climatic environmental data that was measured and to be registered into a database, are becoming necessary.

Challenge of Agricultural Helping System. Through proving tests, after research the results from installation, operation, and utilization of ICT based decision helping system, the following subject matters have been actualized.

Improvement. Develop in aspects of operation and maintenance of the environmental monitoring system. Help for easy installation and expansion of the sensor network system. Help for installation at locations where a power source and networks cannot be secured. Development of low priced environmental monitoring devices and sensor types (agricultural specific sensors considered).

Storage. Development and supplying of the tools related to the utilization of gathered agricultural information as well as storage of the expertise. Discover methods of using information and using technology (ICT or statistical figures). Develop, supply, and prepare a utilization manual involving the display, comparative verification, sharing, analysis and such of information. Disclose measurement data and such information that are intended for sharing on websites. Advice the helping comments by the specialist on the development and utilization of SNS.

Purpose. In order to solve the problematic points of current agricultural ICT based decision helping system that we actualized, while proactively applying technological seeds, we will promote

setup and development of following 3 research themes. For sake of advancing the regions, our goal is to develop and substantiate ICT based decision helping system, hardware, and software in regards to the agricultural sector. Enhancement of the environmental monitoring system that is compatible to various environments and user needs. Construct the platform for displaying, sharing, analyzing of agricultural information, and their validity evaluation.

Research Details

For overall picture of our research and developments, towards the tea plantations located at the semi mountainous area, consists of constructing energy harvesting sensors in wide and sparse areas, mobile relay stations for vehicles and such, and ultimately, a data management center that collects measurement data. In regards to the energy harvest type environmental monitoring system, the summary of natural energy and sensor is displayed on Fig. 2. It allows a combination of a storage battery that stores natural energy from solar panels, wind turbine generators, hydroelectric generators from the rain water, a device that is equipped with a communication module, and every sensors and memory for the measurement data are installed.

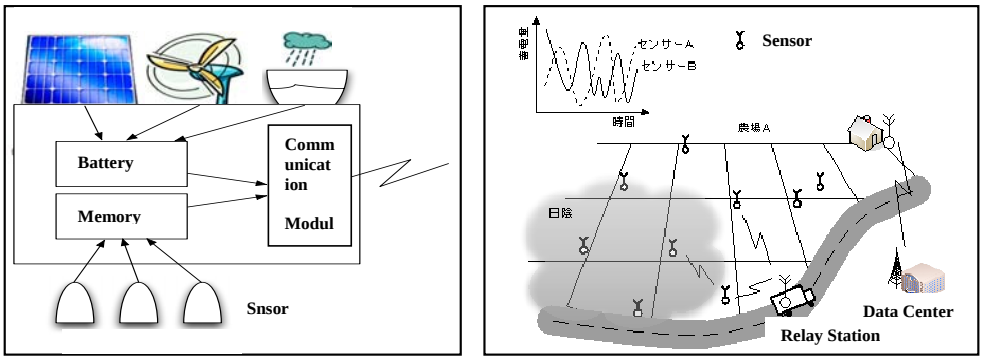


Fig. 2, This is energy harvest type sensor using solar, wind, rain water and the network of 3rd party relay system in the field. The source is from Kyushu University.

On the actual farms we will use the solar panels or drive the wind turbine generators and such, and perform an investigative analysis on how much electric power can be gained. Also for communication device, electric energy in the storage battery will be monitored, and when required amount of electric energy is charged, data will be sent. The matters that involve this data transmission including effects of communication range are all important points for the environmental monitoring system and will be part of research for upgrading.

Concretely speaking, this includes method of determination for data size that is dependent on remaining electric power, the method of locating mobile relay stations, and transmission timing control. At the same time, from each sensor as well, we will perform detailed investigative analysis of consumed electric energy, clarify relationship between electric power generated through natural energy and data transmission electric energy, and we will develop a communication protocol that enables an autonomous collection of data.

We will perform investigation of a 3rd party relay system (Fig. 2) in which acquired measurement data is passed through vehicles for transmission. The mobile relay station will tour through all the sensors, and ultimately transmit gathered data to “Data Center”. With this approach, even with sensors that are placed unto wide and sparse areas, it will be possible to efficiently collect while at the same time, limit consumption of electric energy.

By operating the trial prototype energy harvesting sensor on the actual farms, while evaluating analysis and simulation results up to this day at same time, in regards to any insufficient points a

feedback can be placed at any moment, and by improving the energy harvest type environmental monitoring system including communication protocol, we plan to devise upgrades.

The ability to manage data of the environmental monitoring system in a unified way and to speed up the secondary use (optimization of agricultural production, improvement of production technology, administrative improvement, and feedback to research and development) of accumulated agricultural information database is one of the most important points amongst platform construction. On the system being developed, every type of environmental information of the farm is displayed, and not only can the worker share and analyze, but also by equipping SNS that has experienced a boom in the number of registered users in recent years, their communication on the day to day level becomes much easier. Moreover, through management of all shared and collective expertise of up to today (a place of wisdom knowledge) between farmers, between the farmers and the related parties ("Japan Agriculture" staff, staff of "Improving and Popularizing Agriculture", researchers and such are assumed) whom are all users, it becomes possible for experienced people and field experts to give their advises. This is what we call knowledge management platform.

Summary

Research accomplishments of research subject were unfolded at the tea plantations in Oita prefecture, and with cooperation of the farmers and agriculturally related parties, we will implement a comprehensive operations test for the environmental monitoring system and the platform. We will evaluate and verify overall stability, user-friendliness, availability. By planning to successively improve and modify problems that are discovered, we will build it up unto a more versatile specification. In order to perform experiment that assumes application and utilization of achievements up to now in a concrete way, we will perform regional and coordinated environmental constructions of the field, as well as perform verification of functions, performance, and such of developed technology.

Although it is surmised that turning agricultural information into "knowledge" will be a difficult task, it is also strongly identified from the past, that data accumulation and its application is a critical factor and we are expecting that methods will be established for information sharing of agricultural knowledge through ICT. Also through improvements and generalization of the sensor network, a motivating factor to the farmers as well as farm management system will follow, and through establishment of a foundation emphasizing environment, optimization and social participation of agricultural industry, and safety information transmission regarding food for the consumer will ripple as a huge effect.

From here on, we are scheduled to ultimately compile the verification experiment of this year implementations together with research and examination results related to the scope of application of technology, its validity.

References

[1] J. Burrell, T. Brooke, and R. Beckwith, "Vineyard Computing: Sensor Networks in Agricultural Production," IEEE Pervasive Computing, Vol.3, No.1, pp.38-45, Jan. 2004.

[2] W. Zhang, G. Kantor, and S. Singh, "Integrated Wireless Sensor/Actuator Networks in an Agricultural Application," Proc. 2nd. Int. Conf. on Embedded Networked Sensor Systems (SenSys), p.317, Nov. 2004.

[3] A. Baggio, "Wireless Sensor Networks in Precision Agriculture," Proc. Workshop on Real-World Wireless Sensor Networks, 2005.

[4] T. Fukatsu and M. Hirafuji, "Field Monitoring Using Sensor-Nodes with a Web Server," J. Robotics Mech., Vol.17, No.2, pp.164-172, 2005.

[5] N. Wang, N.-Q. Zhang, and M.-H. Wang, "Wireless Sensors in Agriculture and Food Industry: Recent Development and Future Perspective," Int. J. on Computers and Electronics in Agriculture, Vol.50, No.1, pp.1-14, Jan. 2006.

[6] T. Wark, P. Corke, P. Sikka, L. Klingbeil, Y. Guo, C. Crossman, P. Valencia, D. Swain, and G. Bishop-Hurley, "Transforming Agriculture through Pervasive Wireless Sensor Networks," IEEE Pervasive Computing, Vol.6, No.2, pp.50-57, Apr. 2007.

[7] F. J. Pierce and T. V. Eliott, "Regional and On-Farm Wireless Sensor Networks for Agricultural Systems in Eastern Washington," Int. J. on Computers and Electronics in Agriculture, Vol.61, No.1, pp.32-43, 2008.

[8] L. Ruiz-Garcia, L. Lunadei, P. Barreiro, and J. I. Robla, "A Review of Wireless Sensor Technologies and Applications in Agriculture and Food Industry," Sensors, Vol.2009, No.9, pp.4278-4750, 2009.

[9] T. Okayasu, H. Yoshida, T. Miyazaki, T. Nanseki, M. Mitsuoka, and E. Inoue, " Feasibility Study on Field Monitoring and Work Recording System in Agriculture," Proc. ASABE Annual Meeting 2011, No.110909, pp.1-9, Aug. 2011.

[10] C. S. Ryu, M. Suguri and M. Umeda, "Estimating Quality and Quantity of New Shoots for Green Tea in Field Using Ground-Based Hyper Spectral Imagery," Proc. 8th European Conf. on Precision Agriculture, pp.143-154, July 2011.

[11] S. Ninomiya, "Successful Information Technology(IT) for Agriculture and Rural Development," Extension Bulletins, Food and Fertilizer Technology Center, Vol.549, pp.1-19, Sep. 2005.

Mentally Framing a Three-dimensional Object from Plane Figures Increases Theta-Band EEG Activity

Koji Kashihara[1, a]

[1]Institute of Technology and Science, The University of Tokushima,

2-1 Minamijyousanjima, Tokushima, Japan 770-8506

[a]kojikasi@is.tokushima-u.ac.jp

Keywords: mental imagery, EEG, theta waves, wavelet transform, 3D objects

Abstract. Although mental imagery is a crucial psychological theme, the complicated brain function in mental images still remains unknown. The dynamical brain activity during the mental image of two- or three-dimensional (2D or 3D) shapes was assessed in this study. A healthy male volunteer for a pilot feasibility study participated in the repeated EEG measurements. The mental image task to frame a 3D object from actual 2D figures (2D-3D task) had the longest reaction time; the theta power at the electrode site of Fz was greater in the 2D-3D task than in the simple tasks of 2D or 3D, suggesting the existence of an important process in the central frontal region during such a mental task.

Introduction

Mental rotation generally refers to the ability to imagine the rotation of a two- (2D) or three-(3D) dimensional object [1,2]. The observer's reaction time is linearly increased as the angles of the object increase [1]. Whereas such mental image is a crucial psychological theme, there are few applied studies of it. Actually, there exist some cases to mentally image 3D shapes from plane figures [3]. For example, computer aided design (CAD) or clinical assessment [4] in magnetic resonance imaging (MRI) and/or computed tomography (CT) data [5] needs the accurate and quick judgment of the 3D image, operating plane figures. However, it is hard to estimate how the brain function works in such complicated situations; the applied cases have been inadequately assessed.

The real world can be perceived as 3D, despite the fact that images projected on the retina are reduced to two dimensions. The perception indicates that the 3D shape is framed from plane images on the retina by the brain [6]. Because traditional working memory tasks activate the frontal lobe during temporal memory [7], the mental operation of 2D and 3D shapes according to the positional relationship among elements may need further high cognitive processes [2]. Recently, the image task to mentally frame a 3D object from plane figures showed the longer reaction time; the accuracy was lower than that in other tasks [8]. However, the brain dynamics is still unknown during such a mental image task. Accordingly, this study aimed to investigate the brain activity during mentally framing a 3D object from plane figures.

Methods

To estimate brain activity during mental imaging, a healthy male volunteer with normal vision (age = 35 yr.) was participated in this feasibility study of electroencephalogram (EEG) recording. Written informed consent was obtained from the participant after a complete description of the experiment. The experiment for the mental image tasks was repeated three times, with a time interval of two days.

Image Tasks. The four mental image tasks were performed to estimate the actual reaction times. (A) Simple 2D task. After a 2D figure appeared, a participant mentally imaged the 2D figure. (B) Selective 2D task. After three 2D figures were presented, a participant mentally imaged the 2D figures. (C) 2D-3D task. Three 2D figures appeared in order of top, front, and side views. A participant imaged the 2D figures and mentally framed a 3D figure. (D) Simple 3D task. After a 3D figure was shown, a participant mentally imaged the presented 3D figure.

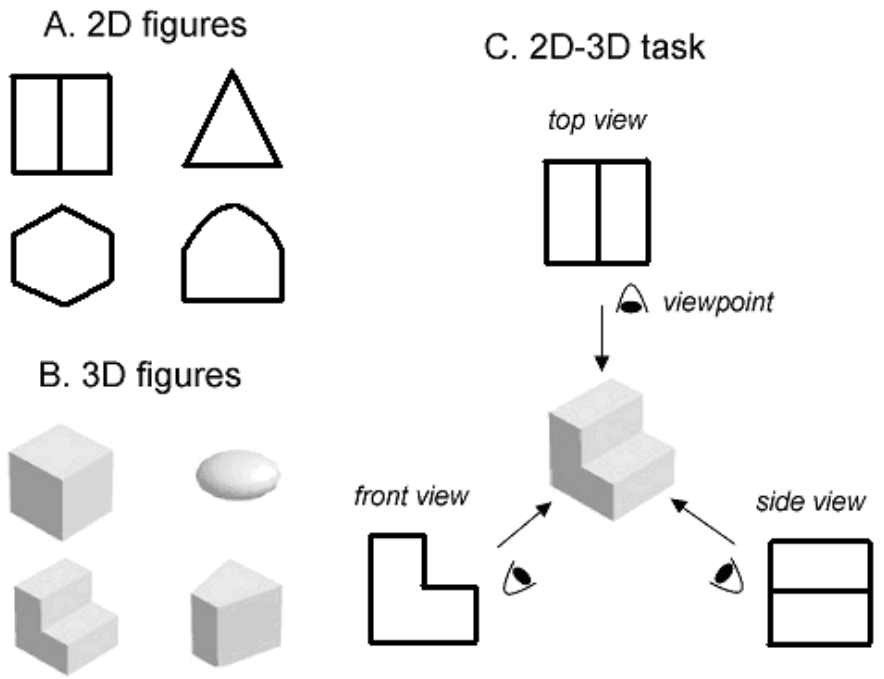


Figure 1. Examples of 2D (A) and 3D (B) figures. C. Scheme of the 2D-3D task. The 3D object was indicated by three 2D figures: top, front, and side views.

Procedures. A participant was sufficiently instructed to press the button quickly and accurately and to concentrate on the next trial, irrespective of their mistakes. The participant confirmed all figures shown in this experiment before all tasks. In all tasks, the sufficient practice was performed to familiarize a participant with the task procedure. The shown figures were randomly selected from the dataset.

A "Start" stimulus in each trial was initially shown for 500 msec. After a period for 600-1200 msec. with a cross bar (+) on the center of a screen, a figure or figures were pseudo-randomly presented from the prepared stimuli: (A) a 2D figure (300 msec.) in the simple 2D task, (B) three 2D figures (300 msec. in each figure; time interval of 150 msec.) in the selective 2D task, (C) three significant 2D figures (300 msec. in each figure; time interval of 150 msec.) in the 2D-3D task, and (D) a 3D figure (300 msec.) in the simple 3D task.

After a black background was shown for 150 msec. to remove an afterimage, the "Go" stimulus (300 msec.) appeared. With the stimulus, the participant was required to mentally image the figure(s) previously presented as correctly and quickly as possible. When the figure(s) was clearly imaged, the mouse button was quickly pressed. The response time was limited within 7 sec.; if the time was surpassed without response, the next trial was automatically started.

EEG Recording. First, the above mental image tasks (20 trials in each task) were performed to estimate the reaction times. The order of the tasks was randomized. Then, EEG signals with the image tasks were recorded. Each task consisted of four blocks (20 trials per block). Four blocks in reference to all tasks (i.e., one block in each task) were randomly allocated; this process was repeated four times (a total of 80 trials every task). To remove noise contamination from muscle activity, the participant was needed to keep the mental image for five sec. without pressing a button when the figure(s) was clearly imaged or framed in the brain. The participant had time to rest for a few minutes between the tasks.

EEG signals were recorded for the experimental period from the sites of Fz, Cz, and Oz in reference to the left earlobe. Electrooculogram recording was performed using electrodes placed on the upper and lower sites of the right eye. All signals from electric amplifiers (EEG100C: gain = 10,000 with a high-pass filter of 0.1 Hz; EOG100C: gain = 1,000 with a high-pass filter of 0.05 Hz) were sampled at 250 Hz with data acquisition system (MP150; BIOPAC Systems, Inc., USA).

Data Analysis. The average reaction times for the image tasks were computed from the first 20 trials. All trails with artifacts of body movements, eye blinks, and background noise ($> \pm 30$ μVolts) were excluded from EEG analysis. EEG activity was estimated for theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), and early gamma (30-40 Hz) frequency bands.

Power spectral analysis.—To express the signal's oscillation amplitude across the entire frequency spectrum, the fast Fourier transform was applied to the EEG data. The power spectral densities during a 4-s period immediately after the "Go" stimulus (i.e., the mental imagery period) were calculated from the EEG traces, applying a linear detrend and the Hanning window (half-overlapping segments of 256 data points). After the calculation of the power spectra for a single trial, those for all trials at Fz, Cz, and Oz were averaged every task.

Time-frequency analysis.—The recorded EEG signals can be classified into evoked and induced rhythms. The evoked activity is strictly phase-locked to a stimulus in an experimental condition. On the other hand, the induced activity is not strictly phase-locked to the cue stimulus, and it can extract the frequency property appearing at a different time every trial [9,10]. To evaluate the induced EEG activities during the mental image of 2D and/or 3D, a time-frequency analysis was performed in this study. The wavelet transform was applied to each single trial. Especially in theta and alpha bands, those values at Fz were averaged across all trials.

The signals were convoluted by complex Morlet wavelet [9]:

$$w(t, f_0) = \exp\left(\frac{-t^2}{2\sigma_t^2}\right) \cdot \exp(2\pi f_0 it) / \sqrt{\sigma_t \sqrt{\pi}} \,. \quad (1)$$

The standard deviation (σ_t) of the time domain is inversely proportional to the standard deviation (σ_f) of the frequency domain [$\sigma_f = (2\pi\sigma_t)^{-1}$]. The f_0/σ_f determining the effective number of oscillation cycles comprised in the wavelet was set at 7 with f_0 ranging from 4 to 12 Hz in increments of 0.1 Hz.

After a linear trend was subtracted, the continuous wavelet transform of a time series u(t) was calculated as the convolution of a complex wavelet with the u(t):

$$\tilde{u}(t, f_0) = w(t, f_0) * u(t) \,. \quad (2)$$

The squared norm of the wavelet transform was calculated in a frequency band at around f_0.

Results

The average reaction times in the simple 2D, selective 2D, 2D-3D, and simple 3D tasks were 405.9 ± 174.2, 1059.1 ± 257.0, 2137.2 ± 645.9, 437.0 ± 125.6 msec., respectively. The longest reaction time of all tasks was shown in the 2D-3D task. Figure 2A shows the power spectrum density at Fz, Cz, and Oz. At around 6 Hz in the theta band, the peak value at Fz of the power spectrum was larger in the selective 2D and 2D-3D tasks than in the simple tasks of 2D or 3D; however, those in the alpha band at Fz and Cz were inversed. The values of the power spectrum in the early gamma band (30-40 Hz) at Oz were not differed among those in the four image tasks. Figure 2B shows the results of

time-frequency analysis during the mental image period (4 sec. of total) in the theta and alpha bands at Fz. The peak values in the selective 2D or 2D-3D tasks were greater than those in the simple tasks of 2D or 3D; the values in the selective 2D task were higher than those in 2D-3D task.

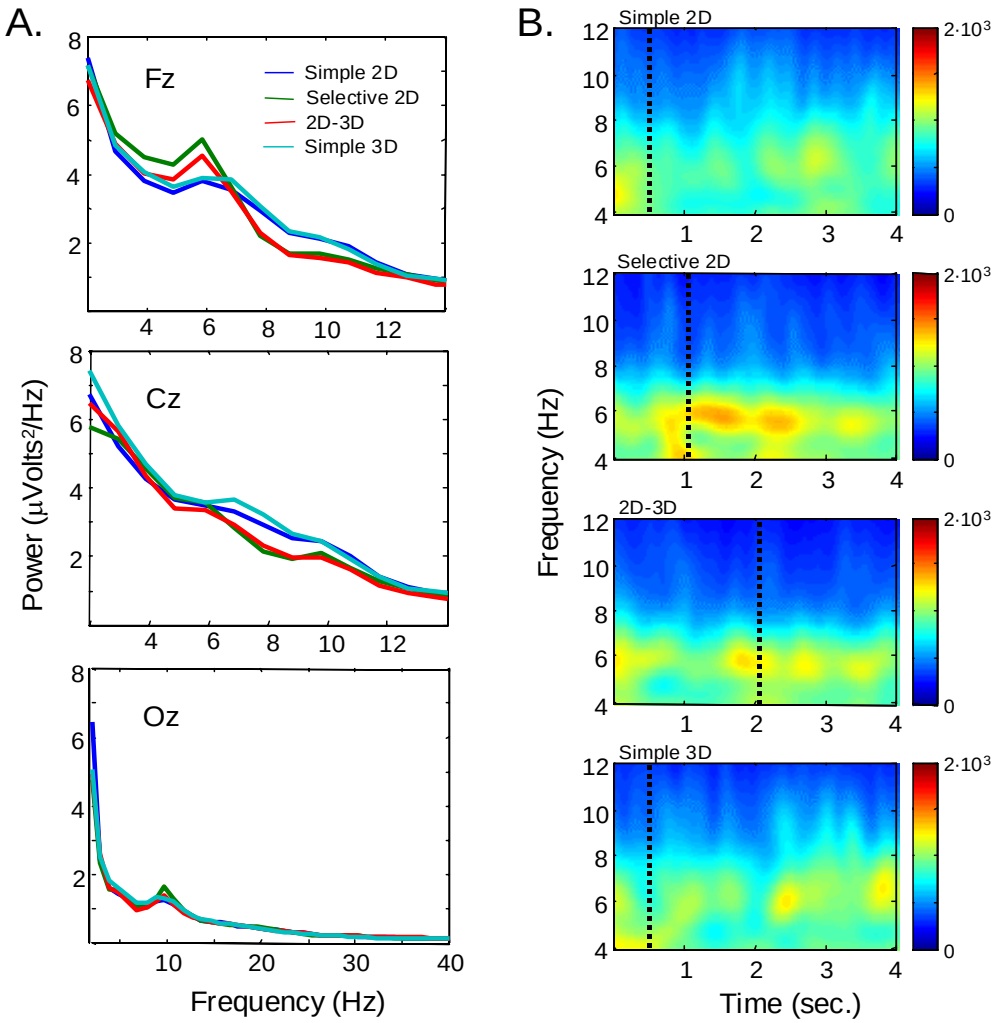


Figure 2. The results of frequency analysis during mental imaging in a subject. A. The values of the power spectrum in the four tasks at Fz, Cz, and Oz. B. Time-frequency analysis in the theta and alpha bands at Fz. Dotted lines show the estimated reaction times of mental images.

Discussion

The theta power at the site of Fz, which usually depends on working memory load [10], was greater in the selective 2D or 2D-3D task than in the simple tasks of 2D or 3D (Fig. 2). Because the 2D-3D task with the longest reaction time requires a more complicated operation to mentally frame a 3D object, it might have activated the wide brain areas (e.g., the primary visual cortex, the parietal lobe related to a spatial operation, and the temporal lobe related to shape recognition) as well as the frontal lobe for memory retention. However, the power of the theta band at Fz was slightly lower in the 2D-3D task than in the selective 2D task. The three plane figures in the 2D-3D task have a meaningful

relationship with each other, compared to those in the selective 2D task requiring the memory of three unmeaning figures. Accordingly, the theta power at Fz during the short-term memory retention in the selective 2D task would have been stronger than that in the 2D-3D task. The top-down process based on experiences [11] might have been also activated in the 2D-3D task.

Conclusion

The mental image task for the construction of a 3D object from plane figures was the longest reaction time of all tasks. In such a task, the theta power at Fz was greater than in the simple tasks of 2D or 3D, suggesting the existence of a crucial role in the frontal lobe. Further brain imaging studies will facilitate the understanding of the mechanism in the complicated mental images. The brain activities at other sites and individual differences should be also considered in future studies.

Acknowledgment. This study was partially funded by a Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (KAKENHI, 22700466).

References

[1] R.N. Shepard and J. Metzler: Mental rotation of three-dimensional objects. Science, Vol. 171 (1971), p. 701-703

[2] M.S. Cohen, S.M. Kosslyn, H.C. Breiter, G.J. DiGirolamo, W.L. Thompson, A.K. Anderson, S.Y. Brookheimer, B.R. Rosen and J.W. Belliveau: Changes in cortical activity during mental rotation. A mapping study using functional MRI. Brain, Vol. 119 (1996), p. 89-100

[3] S. Grossberg: Cortical dynamics of three-dimensional figure-ground perception of two-dimensional pictures. Psychol Rev, Vol. 104 (1997), p. 618-658

[4] R.S. Sidhu, D. Tompa, R. Jang, E.D. Grober, K.W. Johnston, R.K. Reznick and S.J. Hamstra: Interpretation of three-dimensional structure from two-dimensional endovascular images: implications for educators in vascular surgery. J Vasc Surg, Vol. 39 (2004), p. 1305-1311

[5] S.K. Warfield, F. Talos, A. Tei, A. Bharatha, A. Nabavi, M. Ferrant, P.McL. Black, F.A. Jolesz and R. Kikinis: Real-time registration of volumetric brain MRI by biomechanical simulation of deformation during image guided neurosurgery. Computing and Visualization in Science, Vol. 5 (2002), p. 3-11

[6] K. Tsutsui, H. Sakata, T. Naganuma and M. Taira: Neural correlates for perception of 3D surface orientation from texture gradient. Science, Vol. 298 (2002), p. 409-412

[7] E.E. Smith and J. Jonides: Storage and executive processes in the frontal lobes. Science, Vol. 283 (1999), p. 1657-1661

[8] K. Kashihara and Y. Nakahara: Evaluation of task performance during mentally imaging three-dimensional shapes from plane figures. Percept Mot Skills. Vol. 113 (2011), p. 188-200

[9] C. Tallon-Baudry, O. Bertrand, C. Delpuech and J. Pernier: Stimulus specificity of phase-locked and non-phase-locked 40 Hz visual responses in human. J Neurosci, Vol. 16 (1996), p. 4240-4249

[10] O. Jensen and C.D. Tesche: Frontal theta activity in humans increases with memory load in a working memory task. European Journal of Neuroscience, Vol. 15 (2002), p. 1395-1399

[11] I. Bülthoff, H. Bülthoff and P. Sinha: Top-down influences on stereoscopic depth-perception. Nat Neurosci, Vol. 1 (1998), p. 254-257

Semantic Categorization of Emotional Pictures

Koji Kashihara[1, a]

[1]Institute of Technology and Science, The University of Tokushima,

2-1 Minamijyousanjima, Tokushima, Japan 770-8506

[a]kojikasi@is.tokushima-u.ac.jp

Keywords: emotional pictures, support vector machines, bag of features, generic object recognition

Abstract. It is critical to determine the correct semantic categories for classification of generic or social scenes in photographs; however, it is still unfinished, especially in the categorization of emotional pictures. Accordingly, the method to search for emotional information from a picture database was investigated using the bag of features scheme. The SVM classifier for emotional pictures was more accurate compared with the categorization by image similarity. It was more remarkable in the neutral picture category than in other emotional categories. There were some cases in which the bag of features scheme based on local features mistakenly selected similar images that were not in the same semantic category of emotion. Further efficient methods for emotion categorization should be considered in future studies.

Introduction

Computer vision technology has been rapidly developed with the increase of Web contents and materials. Generic object recognition and image categorization are crucial themes in computer vision [1]. Although specific object recognition identifies strictly defined targets, generic object recognition categorizes visual object classes and labels them. Thus, generic object recognition requires grouping across semantic categories in general scenes or objects; it needs a wide range of image categorization, suggesting the difficulty of the extraction and definition of the image features in all objects. However, a late development of the method for the expression of image features and for the categorization of visual images has made it possible to recognize generic objects with high accuracy. For example, the speeded-up robust features (SURF) and scale-invariant feature transform (SIFT) can effectively search for local information of the object boundary [2], and they have been applied for the abstraction of features in target objects. The bag of features (BoF) scheme has been demonstrated as the efficient style of image feature distribution [1]. In addition, machine learning methods with discriminative models such as support vector machines (SVM) and a boosting algorithm can construct model-based recognition with high precision [3].

One of the current tasks for generic object recognition is the extraction of regions for objects and labeling of the image regions. It is also crucial to consider how to determine the correct semantic categories for classification of generic scenes or states in photographs. However, the method for the judgment of social scenes or situations is still not established, and this problem comes under the categorization of emotional pictures. Even if the categorization and labeling of generic objects are successful, it is hard for computer vision to understand the actual situation correctly. For instance, the scene of a fire picture must be recognized from the construction and meaning of whole objects after the generic object recognition of a flame and objects (e.g., a house, a forest, etc.). In this case, computer vision has a possibility to misread the target image and categorize it incorrectly (e.g., the neutral image categorized as a campfire, not an unpleasant one). Even when a fire is directly and

correctly recognized by the image scene analysis—which may require a great number of databases to label—it will be impossible to correspond to all emotions because of individual differences to visual scenes.

Thus, the categorization of emotional images is a difficult task in computer vision. Accordingly, the purpose of this study was to categorize various emotional pictures into correct groups, using the image similarity based on the BoF scheme. First of all, it is important to know how similar pictures in the same emotion category are. Searching for similar pictures from an emotional database and their classification is effective for the construction of a human-machine or computer interface for reflecting individual preference [4]. This study also investigated the method of semantic image categorization in emotion; the correct classification rate was evaluated in emotional images categorized by the SVM.

Feature Extraction for Emotional Pictures

Dataset. Emotional pictures were selected from the International Affective Picture System (IAPS) [5]. All pictures were converted to 8-bit grayscale bitmaps. The mean and distribution of the luminance of grayscale pictures were adjusted such that they did not differ among the pictures sets. The size of a picture was 360 × 270 pixels. The emotional pictures for the visual stimuli were rated by volunteers (N = 9; age 27.2 ± 3.9 years.), using a 1–7 scale (1: extremely negative and 7: extremely positive). In the rating scores, the paired t-test revealed that there was a significant difference among the picture conditions ($p < 0.01$ in each): 4.4 ± 0.8, 4.9 ± 0.6, and 2.5 ± 0.5 for the neutral, pleasant, and unpleasant pictures.

Methods. A picture dataset with three emotional categories (neutral, pleasant, and unpleasant pictures, 60 images in each category) was set for the test of this system. The BoF, which is based on the idea of the bag of words scheme for categorization of textural data, was selected to search for similar pictures. The following steps were performed to classify emotional images [1].

(1) The SURF technique was selected to extract the descriptor as image features, and it is a robust feature descriptor showing the distribution of the pixel intensities within a scale dependent neighborhood of each interest point. The Haar wavelet to increase robustness and decrease the computation time is adopted for a simple filter [2].

(2) The k-means clustering was adopted for this study because of the simplest square-error partitioning method. Visual words were created by the k-means algorithm to cluster the feature vectors of SURF and to create a visual vocabulary. This algorithm proceeds by iterated assignments of points to their closest cluster centers and recomputation of the cluster centers [6].

(3) The BoF was constructed by counting the number of patches assigned to each cluster. Emotional pictures were represented by histograms of visual words, which make it possible to compute the similarity among images and to be utilized for features of an SVM classifier to judge the categories of emotional photographs.

Evaluation by Similarity Calculation

Similarity Calculation. After the construction of histograms as the feature vector, the kd-tree method was applied for the similarity calculation between the target and database pictures. The kd-tree method for a fast approximate nearest neighbor search algorithm [7] was applied for the selection of similar images. The top five similar images were searched from the picture database (60 pictures in each emotional category).

Results. Figure 1A shows the average accuracy to correctly classify the emotional categories between a query image and the top 1 or top 1–5 similar images. The accuracies between the target and selected images ranged from chance level (33.3 %) to 50 % in all categories; those values were higher in the emotional categories than in the neutral pictures. Figure 1B represents the mean image

similarity selected as the top 1 or top 1–5 in each emotional category. The value of 1 in the similarity indicates the same pictures. The values of the similarity in all categories ranged from 0.7 to 0.8; those values were higher in the pleasant and unpleasant categories than in the neutral category.

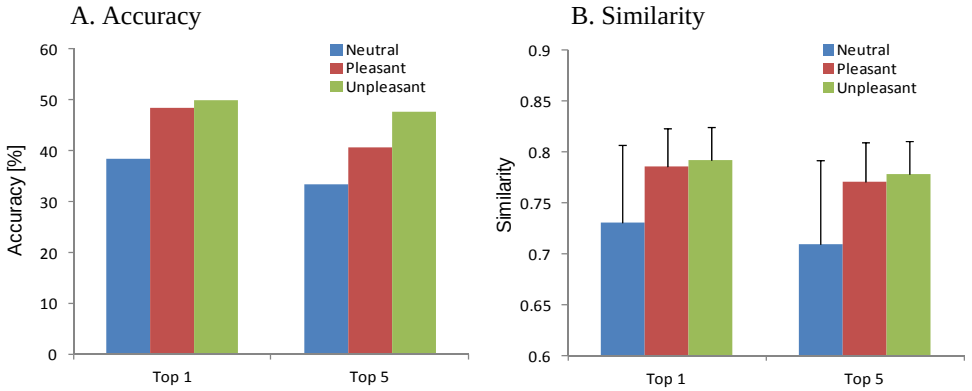


Figure 1. A. Classification of emotional categories based on image similarity. B. Means ± standard deviation of the similarity in selected pictures.

Figure 2A–C provides typical examples of detected similar pictures at high accuracy in all picture categories. With good results, most of the similar pictures selected in each category seems to contain the resemble objects and situations (e.g., pretty animals, accidents, etc.). Figure 2D indicates a poor example of the similar picture selected from another category (a query image in the unpleasant category and a selected one in the pleasant category). In this case, the BoF scheme was able to search for similar images of children. Although the query image had a child's face in pain during dental treatment in the unpleasant category, the selected similar image was a smiling face of a child in the pleasant category, indicating that the social scene of a query image is not consistent with that of the selected image.

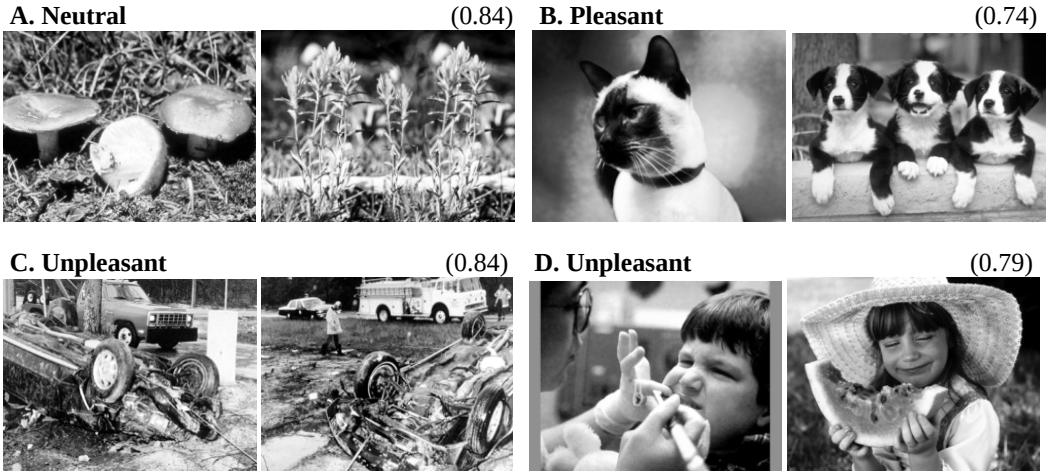


Figure 2. Typical examples of the detected similar images in each emotional category (A: Neutral, B: Pleasant, and C: Unpleasant). A query image (*left*) and the top similar one (*right*), which were selected from the same emotion category in the cases of A–C. D. An example of the picture selected from another category at the highest similarity (an unpleasant query image and a pleasant one with the top similarity). The values in parentheses show the similarity between a query image and a target one.

Classification by Support Vector Machines (SVM)

Method for classification. The SVM classifier was applied to determine a hyperplane that optimally separates samples from two classes with the largest margin [3]. An optimal separating hyperplane is calculated by solving the constrained optimization:

$$\min_{z,b,\xi}\left(\frac{1}{2}\|z\|^2 + C\sum_{i=1}^{l}\xi_i\right) \quad (1)$$

subject to $y_i(z\cdot\Phi(x_i)+b)+\xi_i \geq 1$ and $\xi_i \geq 0$ (i = 1, · · · , l), where l is the number of training vectors, y_i ∈ {-1, +1} is the class label of the output, and $\|z\|^2 = z^T z$ is the squared Euclidean norm. The weight parameter $\boldsymbol{z}$ determines the orientation of the separating hyperplane, b is a bias, ξ_i is the i^{th} positive slack parameter, and Φ shows a nonlinear mapping function. A parameter C is the penalty term and controls the size of $\boldsymbol{w}$ and the sum of ξ_i. To control the fractions of support vectors and margin errors, we chose the v-SVM method, which is a formulation of SVM using a parameter $0 < v \leq 1$.

The vector $\boldsymbol{\Phi(x_i)}$ corresponding to a nonzero value is a support vector of the optimal hyperplane. It is desirable to have a small number of support vectors to achieve a compact classifier. The optimal separating hyperplane is computed as a decision surface of the form sgn:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{L}\alpha_i y_i K(x_i, x) + b\right), \quad (2)$$

where sgn(.) ∈ {-1, +1}. The nonlinear kernel function K projects the samples to a feature space of higher dimension via a nonlinear mapping function, and L is the number of support vectors. The radial basis function as the nonlinear kernel was defined as

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2), \quad (3)$$

where the value of the kernel parameter γ determines the variance of the function.

Validation. The 180-fold cross-validation was performed between the two types of categorized data (the neutral, pleasant, or unpleasant category and others) from all pictures. The feature for SVM was the value of histogram acquired from the BoF scheme in emotional images. The nonlinear kernel parameter γ of Eq. (3) was set at four types: 0.0001, 0.01, 1, and 1 / (the number of features). The linear kernel was also evaluated. The tolerance of termination criterion for the SVM was set at ε^{-12}.

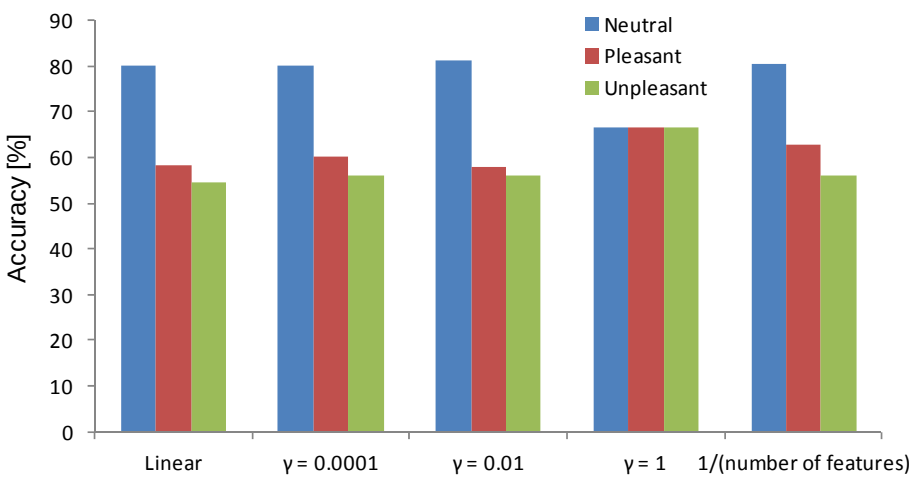


Figure 3. Results of cross-validation in an SVM classifier for emotional images. A nonlinear kernel parameter is shown as γ.

Figure 3 represents the results of 180-fold cross-validation for the SVM classifier to judge the emotional categories from histogram elements of visual words. All picture categories had higher

accuracy compared to the chance level (i.e., 33.3 % in each category). The accuracy was higher in the neutral picture category than in other categories. The accuracy in the feature of neutral images was constantly kept at a high level of accuracy (80 %), compared with those of pleasant or unpleasant data (66.6 % at the maximum).

Discussion

The effective method of emotion categorization may be different between the pleasant and unpleasant pictures. The unpleasant images will induce common feelings among most people (e.g., traffic accidents, injury, snakes, etc.), but the emotion of generally pleasant images presumably includes many individual differences (e.g., foods, vehicles, sports, etc.). In addition, whereas the unpleasant situation and meaning are not clearly consistent with the pleasant ones, there are some cases showing similar image features. For example, a child's face in a pain during dental treatment categorized in the unpleasant category is definitely different from a smiling face of a child in the pleasant category (Fig. 2D). In such a situation, we can recognize the correct emotional situation easily, but the result of the classification in this study was incorrect due to the very high similarity between the image features based on the BoF scheme. This result suggests that the understanding of emotional scenes or situations requires judging the correct semantic category of images.

A possible method to improve the poor rate of emotion categorization is the analysis of the text-based information on emotional images; however, it is hard to apply it to all images in databases. The application of an SVM classifier to emotional images made it possible to increase the accuracy of classification, compared to that categorized by the image similarity alone (Fig. 1 vs. Fig. 3). The category of neutral images showed a high identification rate (80 %). However, even in the SVM classifier, there existed a difficult case to categorize the pleasant and unpleasant images. The correct adjustment of the nonlinear kernel parameter for the SVM may improve this problem. Considering the increment of a classifier's rate for generic object recognition, further efficient methods or technology for the correct categorization of emotional images would be required as the next step for computer vision.

Conclusion

The BoF scheme was applied as a method to search for similar emotional images. In contrast to the categorization by image similarity, the SVM classifier was more accurate for emotion classification; however, it was more remarkable in the neutral picture category than in other emotional categories. In some cases, the BoF scheme based on local image features detected similar objects or humans in pictures, regardless of the different emotion categories. More effective methods of semantic categorization must be considered in future studies.

Acknowledgment. This study was partially funded by a Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (KAKENHI, 22700466).

References

[1] G. Csurka, C.R. Dance, L. Fan, J. Williamowski and C. Bray: Visual categorization with bags of keypoints. Proceedings of the IEEE Workshop on Statistical Learning in Computer Vision (SLCV'04) (2004), p. 1-16

[2] H. Bay, A. Ess, T. Tuytelaars and L.V. Gool: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, Vol. 110 (2008), p. 346-359

[3] N. Cristianini and J. Shawe-Taylor: Introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)

[4] K. Kashihara, M. Ito and M. Fukumi: Development of automatic filtering system for individually unpleasant data detected by pupil-size change., Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (2011), p. 3311-3316

[5] P.J. Lang, M.M. Bradley and B.N. Cuthbert: International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual (Technical Report A-6). Gainesville: University of Florida (2005)

[6] R.O. Duda, P.E. Hart and D.G. Stork: Pattern classification, second ed. John Wiley & Sons, Singapore (2001)

[7] M. Muja and D.G. Lowe: Fast approximate nearest neighbors with automatic algorithm configuration. In VISSAPP (2009), p. 331-340

Embodied Conversational Agent Model

Hima Bindu Maringanti, Aditya goil, Indraneel Srivastav, Sonali Satsangi
Jaypee Institute of Information Technology,

hima.bindu@jiit.ac.in, adigoil@gmail.com, indraneel.max@gmail.com,

sonalisatsangi89@gmail.com

Keywords: Embodied Conversational Agent, Emotions, Human Computer Interaction, Computer Vision, Intelligent Agent, 3D modeling Avatar, Questioning System, Natural Language Processing.

Abstract. As the advancement of technology, state of the art I/O devices are being created to ease the mode of interaction with the computer and to make it more fun. But they still lack the human factor. All these technologies and gadgets are pulling humans closer to the machine. Rather than bringing the digital world to the physical world the exact opposite is happening. For that matter, the concept of embodied agents have been touched, where the user need not have any prior knowledge of the computer or how to work with an interface so as to communicate with the computer, instead he can simply talk to the computer as if having a normal human-human interaction.

Although there are many implementations of embodied agents that have been developed and used for day to day use and fast interaction with the computer to complete simple tasks. For instance, online juke boxes are controlled by simple talking to an agent and bank transactions are handled by agents by simply talking to them. But still these agents lack the power to understand and analyze the human perception and the emotional state of the human being.

Introduction

Rapid increase in technology has led to the need for more interactive ways for communicating with the digital world. Since a few years, the traditional mouse and keyboard have been replaced by touch screens, hand gestures, body gestures and even expressions portrayed by a human being. Technology is trying to use the natural human input so as to minimize the way we interact with the computer. Recent advances in technology and the invasion of computing into more and more social domains has created more opportunities for interfaces which are closer to human. From interfaces for controlling smart homes to personal news-reader to question answering systems; interaction designers are being asked to create more generic interfaces.

An extremely new domain is being explored in which computer is given vision to understand and analyze its environment just like a human would analyze. As users, humans would simply have to communicate with the computer as they would interact with another human being. Such a free form interface for easy and interactive use is called an Embodied Conversational Agent. As the name specifies, the computer is personified with a virtual human being who interacts with us and analyzes us.The fields of Human-Computer Interaction and Computer vision are thoroughly being integrated and researched to make ECA happen. Recently Apple's visionary Knowledge Navigator video with the embodied agent Phil, who assists a college professor in planning and managing the interaction between his work and family life through social dialog, provides a well-known projection of what embodied agents might do. However, Microsoft's much-maligned "Clippy"—an animated paperclip that often interrupts users when it recognizes an activity with an offer of "help", and who never

remembers how users respond—provides a counterpoint, demonstrating how difficult it is to make embodied agents work in the real world.

The major focus of this paper is on:

* An intelligent agent which independently moves in the natural environment
* A human like personification of the computer, which interacts with us.
* A questioning system which uses world knowledge to ask questions.

Background

Embodied Conversational Agent

Justine Cassell [1] discusses how Rea (the conversational agent) makes conversations with breathing human users with a wink, a nod and a sidelong glance. He mentions how we make complex representational gestures with our prehensile hands, gaze away and toward one another out of the corners of our centrally set eyes, and use the pitch and melody of our flexible voices to emphasize and clarify what we are saying. Human capabilities of face to face conversation are portrayed on Rea to make her respond in the same way as one human would respond to another in a normal human conversation. Four abilities understood are: Recognizing and responding to verbal and nonverbal input, generating verbal and nonverbal output, dealing with conversational functions and giving signals that indicate the state of the conversation, as well as contributing new propositions to the discourse.

Maurizio Mancini [2] studies how our mental and emotional states can be communicated with a large variety of nonverbal behaviors, while influencing the interaction with people. In a first study, many characteristics of behavior were evaluated: tendency to use body, face, head, gestures; qualities of movement, like fast-slow, small-large, smooth-jerky, etc. The person's behavior tendency was shown to be an innate individual characteristic that the author claimed to be a personality trait.In the second study the author investigated the consistency of a person's behavior across time and situations. Results demonstrated this consistency: people that are quick when writing have a tendency to be quick at eating, if a person produces wide gestures, then he/she walks in large steps. A group of people judged the actors' behavior and annotated them. In the study, he authors found that the way actors' portrayed emotional state also seemed to be actor dependant and it depended on the actors' personal way of expressing those emotions.

John Zimmerman [3] and his colleagues, in their paper discussed how people's expectations for the visual form of the agent were clearly influenced by the task the agent performed and by the social and cultural experiences of the user. The paper talks about a conceptual model which was kept in mind to see the gap in the research by exploring the relationships between the visual form of an agent, the task the agent performs, and the gender of the user.

IpkeWachsmuth [4] paper talks about natural communication and human language developed in intimate connection with body. When a person speaks, not only symbols are transmitted, but the whole body is in continuous motion. While speaking we can indicate the size and shape of an object by a few hand strokes, direct attention to a referenced object by pointing or gaze, and modify what we communicate by emotional facial expression. The meanings we transmit this way are multimodality encoded and strongly situated in the present context.

Face Detection

Zhang Bao-jian [5] in their paper, suggested a fast face location method based on gradient distributions. This method can locate face with certain rotation, or light variation or with glasses or beard in the image. It begins with vertical location by use of vertical integral projection on the two-valued image of the original image, and then proceeds with horizontal location according to the distributions of gradient direction. The proposed algorithm is rule-based approach which is locating face by the structural characteristics of face gradient direction.

Jianbo Shi [6] uses a window based technique to track the moving features in the image sequence. The need for a careful choice of the windows to proposed. The proposed criterion, based on the size of the smaller eigenvalue of the tracking matrix G, is well justified by the nature of the tracking method. Furthermore, it subsumes previous feature selection criteria; in that it detects corners equally well as regions with high spatial frequency content, or with high second-order derivatives, or high values of intensity variance.

Carlo Tomasi [7] theoretically explains how feature points are detected and tracked in an image sequence. Two basic questions: how to select the features, and how to track them from frame to frame have been discussed in this paper. Their approach is to minimize the sum of squared intensity differences between a past and a current window. Because of the small inter-frame motion, the current window can be approximated by a translation of the old one. Furthermore, for the same reason, the image intensities in the translated window can be written as those in the original window plus a residue term that depends almost linearly on the translation vector.

Various techniques like Boosted cascade of simple features [8], the well-known Adaboost algorithm[9], Haar-like features for Rapid Object Detection [10], Joint-Haar like features for Face Detection [11], Minimum Facial Feature algorithm [12] were used by people working in the area of Face detection and feature extraction. G.J. Edwards[13] demonstrated a fast, robust method of interpreting face images using an Active Appearance Model (AAM).

Natural Language Processing

Dunwei Wen et.al, discusses a question answering system based on VerbNet frames. The group of rules in a VerbNet frame can be regarded as instances of the case in case grammar theory. In our system, verb frames in question and candidate sentences are detected, so that the thematic information can be therefore obtained. When one of the frames of the question is matched to one of the frames of a sentence in the retrieval corpus, this sentence is regarded as an answer sentence and a more specific answer chunk is extracted from the chunk whose thematic role is identical to the thematic role of the interrogative word in the question. In this way, the sentences with key words which do not have the matching thematic roles are filtered out.

A question semantically matches an answer candidate sentence and to extract the answer chunk from the sentence. Only if at least one of the frames of the question matches at least one of the frames of the answer candidate sentence, the answer candidate sentence is regarded as one answer sentence. And the answer chunk is extracted from the answer sentence.

3D Modeling and Rendering

3D Modeling [17][18] of the human like animation character development is possible using ADOBE software suites and other commercially available software tools, especially used in education and Game development domains, like OpenGL, Autodesk 3DS Max, Autodesk Maya, Google Sketchup Pro and Lightwave 3D and certain Open Source[19] also for Rendering the expression on the character's face to look like a human.

Design and Implementation

This work of design of a conversational agent (Fig. 1) is divided into various functional units as shown below in (Fig.1). The face detection and feature extraction process (Fig. 2) followed the steps given below:

* Input Video is extracted into frames which are images in 'bitmap' format

* Images which are to be used for further processing have to be in Grayscale

* Each image has to be preprocessed by lightning correction and histogram equalization

* Face Detection is implemented using the HAAR like feature detector of OpenCV.

* This HAAR like feature has already been trained with thousands of face images in different lighting and poses.

* Images with this face ROI have to be processed to find the features of the face.

* Another database, with each face image annotated with 58 features points is used to create a statistical shape model.

* Each shape model is normalized using the Procrustes Analysis to a frame of reference.

* Next we do a PCA analysis on each face model to reduce the dimensionality and to calculate the eigenvectors and eigenvalues.

* These eigen vectors are then used to find the shape parameters for each face model which describe the shape of the model.

* Now the new face image which was extracted from the video above, is analyzed to detect the features.

* The mean shape model is projected onto the new image and iteratively fitted onto the new face.

* The fitting is done by an adjustment along the normal to the model boundary towards the strongest image edge, with magnitude proportional to the strength of the edge.

A complete 3D rigged model of a human being was created from scratch in Autodesk Maya, a 3D modeling software. The model was given texture, and skeleton. The skeleton is used to control and move the model. The movement of various joints in the skeleton of the model has been constrained by using IK-Handles. These handles control the smooth movement of the joints and make the animation look realistic. For the rendering of the above rigged model, Unity Game Engine was used [19]. It is an interactive development environment which helps the user seamlessly create 3D environments, render and move characters. It allows C# and JavaScript scripting to help add dynamism to the 3D virtual world. The terrain [20] assets were used and the online documentation to understand the working of the IDE and used it to make the interface of the work.

To give a sense of understanding (Fig. 4) to the avatar, the 3D character, a basic conversational algorithm was implemented, based on an ontology developed by us. Python and NLTK Toolkit [21] was used to develop the application for conversation. Few algorithms mentioned in [14][15][16] were implemented for achieving POS tagging, Chunking and Theme role detection of the sentences.

Results

The object detector of OpenCV has been initially proposed by Paul Viola [8] and improved by Rainer Lienhart [10]. First, a classifier (namely a cascade of boosted classifiers working with HAAR-like features) is trained with a few hundreds of sample views of a particular object (i.e., a face), called positive examples, that are scaled to the same size (say, 20x20), and negative examples - arbitrary images of the same size. After the classifier is trained, it can be applied to a region of interest (of the same size as used during the training) in an input image. The classifier outputs a "1" if the region is likely to show the face, and "0"otherwise. To search for the object in the whole image one can move the search window across the image and check every location using the classifier. The classifier is designed so that it can be easily "resized" in order to be able to find the objects of interest at different sizes, which is more efficient than resizing the image itself. So, to find an object of an unknown size in the image the scan procedure should be done several times at different scales. Feature Extraction

using Active Shape Model technique was done. Active Shape Model is a concept where a statistical model is created for the face using a pre annotated database, the IIM face dataset , consisting of 240 images of different people in various illuminations and poses. Each image contains 58 annotated points (Fig. 3) which best describes the contour of the face. In ASM, this database information is used to extract statistical face from the face images. After extracting the face models of each face, we normalize them using the procrustes analysis. The procrustes analysis is basically done to make the model Translation, Rotation and Scale invariant. On the normalized face(Fig. 5) PCA is applied for dimensionality reduction and then Point Distribution Model (PDM) to search for the best fit face shape from a given database. A face-shape PDM can be used to locate faces in new images by using Active Shape Model (ASM) search. The mean shape is projected into the image and iteratively modified to better fit the image evidence, subject to the shape constraints represented by the model. At each step, the region around each model point is searched for the best match to a local grey-level model learnt during training. This gives a new proposed shape. An ontology was created, which is referred by the embodied agent to generate questions according to the responses given by the user. When the user introduces himself, a random topic is chosen from the ontology and a question is asked from it.

When the user answers this question, the attribute from the response is matched to the answer and is stored in the database to keep a track/record of the user. Once a sub tree is complete, the level is changed, that means the topic of conversation is changed and so on. This process keeps on happening till the ontology is exhausted. The amount of conversation this agent can make depends upon the size of the ontology created. Hence the conversational capacity is limited by the size of the ontology. A complete 3D rigged model of a human being was created from scratch in Autodesk 3DS Max, a 3D modeling software. The model (Fig. 6) was give texture, skin and skeleton. The skeleton is used to control and move the model.The movement of various joints in the skeleton of the model has been achieved by importing into Unity Game Engine and constrained by using IK-Handles. These handles control the smooth movement of the joints, manipulate and move,make the animation look realistic.

References

[1] Justine Cassell: Embodied Conversational Agents, April 2000/Vol. 43, No. 4(2000), Communications of The ACM.

[2] Maurizio Mancini: Multimodal Distinctive Behavior for Expressive Embodied Conversational Agents, ISBN-13:978-1-59942-699-0(2008).

[3] John Zimmerman, Ellen Ayoob, Jodi Forlizzi, Mick McQuaid: Putting a Face on Embodied Interface Agents", CMU, as part of DARPA project, Contract no. NBCHD030010.

[4] Ipke Wachsmuth: Embodied Communication, Artificial Intelligence Group, Faculty of Technology, University of Bielefeld, Germany.

[5] Zhang Bao-jian, GaoGuohong, Lv Jinna, Zhu Yanli: Human Face Location Based on Gradient Distributions, Third International Conference on Knowledge Discovery and Data Mining, 978-0-7695-3923-2/10, IEEE Computer Society.(2010).

[6] Jianbo Shi and Carlo Tomasi: Good features to track, 1063-6919/94, IEEE (1994).

[7] Carlo Tomasi,Takeo Kanade: Detection and Tracking of Point Features, Technical Report, CMU-CS-91-132 (1991).

[8] Paul Voila and Micheal Jones: Rapid Objects Detection using Boosted Cascade of Simple Features, Proceedings of ACCEPTED Conference on Computer Vision and Pattern Recognition (2001).

[9] Hossein Falaki: AdaBoost Algorithm, University of California, LA.

[10] Rainer Lienhart and Jochen Maydt: An Extended Set of Haar-like Features for Rapid Object Detection, Intel Labs, Intel Corporation(2002).

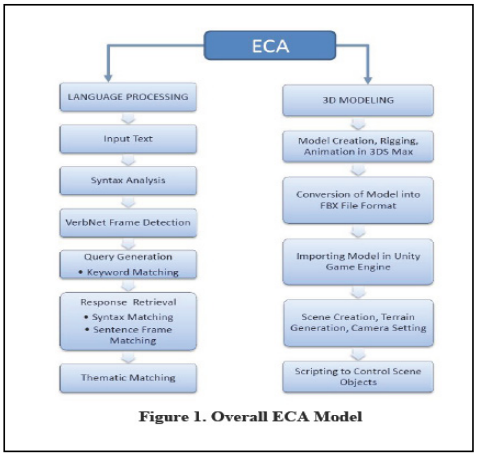


Figure 1. Overall ECA Model

Figure 2. Facial Feature Extraction

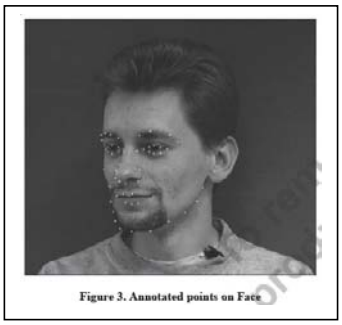


Figure 3. Annotated points on Face

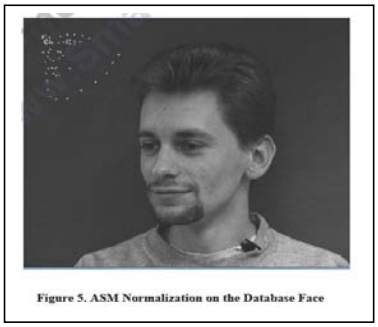


Figure 5. ASM Normalization on the Database Face

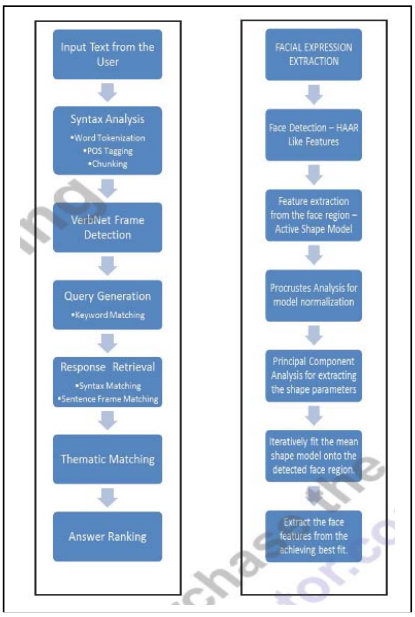


Figure 4. Face Detection & Expression Extraction and Response Generation

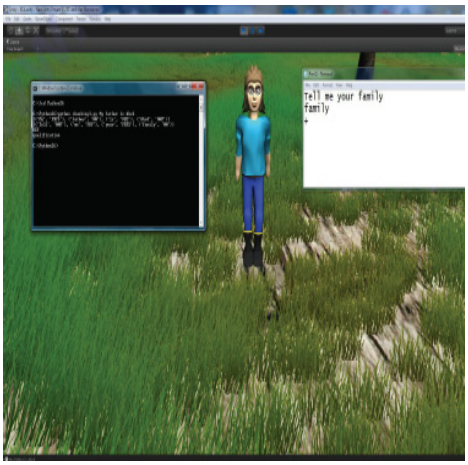


Figure 6. Avatar Interaction

The Novel Application of Bioelectrical Impedance Analysis with Back Propagation Artificial Neural Network to Assess the Body Compositions of Lower Limbs in Elite Male Wrestler

Tsong-Rong Jang [1,a], Hsueh-Kuan Lu [2,b], Ruey-Tyng Kuo [3,c], Yu-Yawn Chen [4,d] and Kuen-Chang Hsieh [5,e*]

[1] Department of Combat Sports, National Taiwan College of Physical Education, Taiwan

[2] Sport Science Research Center, National Taiwan College of Physical Education, Taiwan

[3] Computer Center, National Taiwan College of Physical Education, Taiwan

[4] Department of Physical Education, National Taiwan College of Physical Education, Taiwan

[5] Research Center, Charder Electronic Co., LTD, Taiwan

[a]tong510315@yahoo.com.tw, [b]sk.lu2002@gmail.com, [c]rtkuo@ntcpe.edu.tw, [d]yu11.tw@yahoo.com.tw, [e]abaqus0927@yahoo.com.tw

*Corresponding author.

Keywords: artificial neural network, dual-energy X-ray absorptiometry, fat free mass, tissue mass.

Abstract. The aim of this study was to develop a novel predictive model of fat free mass of lower limbs (FFM_{LL}) by using a bioelectrical impedance analysis (BIA) with a Back Propagation Artificial Neural Network (BP-ANN). The variation in prediction of FFM_{LL} between by the traditional linear regression (LR) equation and the optimal BP-ANN model was elucidated with the identically anthropometric data of 24 elite male wrestlers in Taiwan. The input, hidden and output layers of the BP-ANN consisted of four inputs (age, height, weight and bioelectrical impedance values), five neuron units and two outputs. The evaluation of body composition with 2-C components was with the criterion of the dual-energy X-ray absorptiometry (DXA). Our results showed that the correlation coefficient (r) of the FFM_{LL} estimated by LR ($FFM_{LL\text{-}LR}$), and the FFM_{LL} estimated by BP-ANN ($FFM_{LL\text{-}ANN}$) between the FFM_{LL} by DXA ($FFM_{LL\text{-}DXA}$) were 0.865 and 0.965, respectively. In conclusion, the BP-ANN model had better performance than the traditional LR model in the prediction of the FFM_{LL} in elite male wrestlers in Taiwan.

Introduction

Body composition of athlete is different depending on the type of sport and the level of competition. Nearly every athlete would benefit from an increase in muscle mass for more muscle meaning more strength or better athletic performance [1, 2]. Greater speed performance and strength advantage through field were relative to lower body fat in soccer [3] and football player [4]. There was correlation between maximum aerobic capacity and body composition in Sumo wrestlers [5]. The body composition of athlete is thought to be quite different from that of general people [6, 7].

Lower limbs include the pelvic girdle, buttocks, hip, and thigh, as well as the components distal to the knee. There was evidence that muscles of lower limbs are critical to moving the body mass over the base of support [8]. However, the role of lower limbs is not only to provide a larger base of support but also play an active role in balance [9]. Specifically, the aims of training were to increase the distance reached and the active contribution of the lower limbs to support and balance [10]. Wrestling consists of Greco-Roman and Freestyle. In Greco-Roman, wrestlers can only hold the opponent's body above the waist, while use of lower limbs is forbidden. In Freestyle, wrestlers can

978-1-61275-004-0/10/$25.00 ©2012 IERI

use the technique of Greco-Roman wrestling and that of using lower limbs to trip and grasp. Research showed that there is highly correlation between muscle anaerobic energy metabolism and fat free mass of lower limbs in freestyle wrestlers of the Polish Olympic team [11].

Non-invasive methods of evaluating human body composition were developed, such as air-displacement plethysmography, body circumference, computed tomography (CT), dilution method, dual-energy X-ray absorptiometry (DXA), magnetic resonance imaging (MRI), neutron activation analysis, skinfold thickness and underwater weighing [12]. However, limitations on ether the accuracy or the cost appears in above instructions or methods for evaluating body composition [12]. The easy operation, non-invasive, portable and fast characteristics in BIA estimation of body composition have render it become a feasible application for wildly usage [13-15]. Furthermore, evaluating athletic body composition with its specific predictive equation by BIA was constructively developed [16-18].

The FFM in limbs was high correlation with the bioimpedance index (BI) of limbs; therefore, the BIA measurement could be applied in predicting the body composition of limbs [19, 20]. The prediction equation for evaluating FFM in the lower limbs of young adult male Rugby Union players [21] and skeletal muscle mass (SMM) in the lower limbs of general people [22] by anthropometry referenced with DXA was developed. In addition, the limb muscle volume (MV) could be evaluated by BIA for as the same principle as above mentioned with high correlation between target tissue and BI [23]. However, using BIA measurements to assess FFM, SMM, LBM and MV in the lower limbs of athlete are still necessary.

There were mathematic methods such as ANN [24, 25], Cox regression [26], logistic regression [27], discriminate analysis and recursive partitioning [28] in clinical medicine fields for outcome prediction. The ANN model exhibited well performance with greater validity in the prediction of intercellular fluid (ICF) and total body water (TBW) in chronic hemodialysis patients than that of the linear regression model [29, 30]. The diagnosis of risk in dengue 481 patients by using bioelectrical impedance analysis and artificial neural network was studied [31]. There were very rare application of ANN model to predict the body compositions such as FFM and segmental tissue masses. It was an interesting issue to evaluate whether the greater precision and accuracy in prediction of lower limbs mass in wrestler by the application of BIA with ANN mathematical model.

The present study will focus on the development of a novel mathematical BP-ANN model in BIA measurement to accurately predict the tissue compositions in lower limbs in wrestler. Simultaneously, they were measured by high resolution DXA [32, 33], instrument and with the two components (2-C, including fat free mass and fat mass) model. The precision and accuracy in prediction of tissue compositions by BIA measurements with a novel BP-ANN model and with a linear regression model were compared.

Materials and methods

Subject

Twenty four elite male wrestlers, who were trained for more than 12 hours per weeks and have been disciplined about 10 years, have been recruited under the formal permission of Institutional Review Board (IRB) of Advisory Committee at Jen-Ai Hospital of Taiwan. The consumption of alcohol for 2 days and administration of diuretics for 7 days were restricted before experiments. The basic subject's characteristics were shown as Table 1.

Table 1. Characteristics of the subjects in the present study (n = 24).

	Mean (SD)	Range
Age (years)	19.3 (0.95)	18.0-25.6
Height (cm)	169.9 (5.1)	156.5-178.2
Weight (kg)	70.8 (9.7)	56.4-86.0
BMI (kg/m^2)*	24.5 (2.6)	21.3-28.7

*BMI, body mass index.

The measurement of BIA instrument

The BIA instrument, with independent detection electrodes and current source electrodes in platform embedded with tetra-polar electrodes and griped handle embedded with bi-polar electrodes, can create different circuits to measure the BIA values in corresponding segments by switching measuring models [34]. The QuadScan4000 (Bodystat Corp., U.K.), connected with computer was operated by the current at 400μA with frequency at 50 KHz during measurement. As shown in Fig. 1., the E1, E2, E4 and E6 were current electrodes and E3 and E5 were measuring electrodes. All the electrodes were made of stainless with high conductivity. The E1 and E2 placed on the handle, and the E3, E4, E5 and E6 placed on a platform. The bioelectric impedance value (Z) yielded in left lower limbs and right lower limbs were termed as $Z_{L\text{-}LL}$ and $Z_{R\text{-}LL}$, respectively. The summation of $Z_{L\text{-}LL}$ and $Z_{R\text{-}LL}$ create total lower limbs impedance as Z_{LL}.

The measurement of dual-energy X-ray absorptiometry instrument

Subjects in standard cotton dress without any metal attachment were scanned over whole body by DXA (Lunar Prodigy, GE Corp, USA.) and then analyzed by the software "enCore 2003 Version 7.0". Body weights and tissue weights were determined within the error at 0.1 kg and body heights at 0.5cm. The FFM and tissue weights were subsequently computerized quantified. The scanning protocol was operated at 20μGy in twenty minutes by legally registered and well trained medical technologists in Department of Radiology, Dah Li County Jen-Ai Hospital in Taiwan.

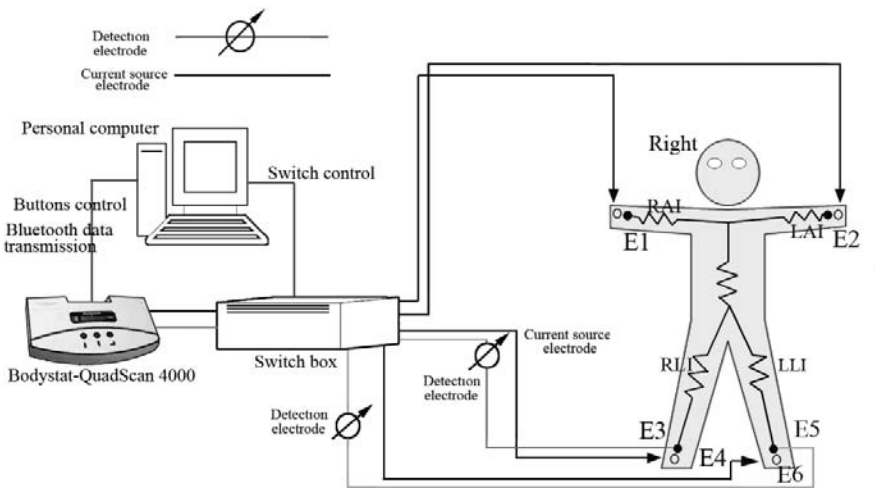


Fig.1 Modified bioelectric impedance measurement system with measuring platform. E1, E2, E4, and E6, measuring electrodes; E3 and E5, current electrodes; LAI, left arm impedance; LLI, left leg impedance; RAI, right arm impedance; RLI, right leg impedance.

Back Propagation-Artificial Neural Network (BP-ANN)

We constructed the FFM_{LL} estimating system by BIA measurement with the Back Propagation - Artificial Neural Network (BP-ANN) mathematic model (Fig. 2), which are composed input layer, hidden layer and output layer.. The input layer input four values, including age (y), height (h), weight (m), and Z_{LL}. The $f^1$and f^2, a type of Log-Sigmoid function, act as transfer functions in hidden layer and output layer, respectively. The hidden layer (S1) contained five neuron units and f^1 transfer functions. The output layer contained two neuron units and f^2 transfer functions to outcome the amount of tissue weight of lower limbs ($Tissue_{LL\text{-}ANN}$) and the amount of the FFM of lower limbs ($FFM_{LL\text{-}ANN}$). As the training rule in present work, we set the maximum iteration as 200 times with the minimum gradient value as 10^{-6}. All of the BP-ANN programs of algorithms above were coded by Matlab Ver.7.0 (MathWorks, Inc. MA). In our work, we adopted the 2-C model to obtain the percentage of fat mass (FM %) by the amount of total body weight and FFM in lower limbs.

Statistical analysis

All the collected data were analyzed by SPSS.12.0 software (SPSS Inc., Chicago, IL, USA). Data were expressed as mean ± SD (standard deviation). The prediction equations for $FFM_{LL\text{-}LR}$ and $Tissue_{LL\text{-}LR}$ were separately processed by multiple regression analysis. Input (R1), hidden (S1) and output (S2) layers of the BP-ANN consisted of 4, 5 and 2 neuron units (estimated $FFM_{LL\text{-}ANN}$, $Tissue_{LL\text{-}ANN}$) as described in above paragraph. R values obtained from linear regression analysis and Pearson were presented to describe the correlation between any variability. We followed the program suggested by Bland and Altman Plot [35] to survey the distribution and variability between segmental FFM values estimated by our developed equations vs. that of measured by DXA. A confidence level of 5% ($p < .05$) was considered as significant.

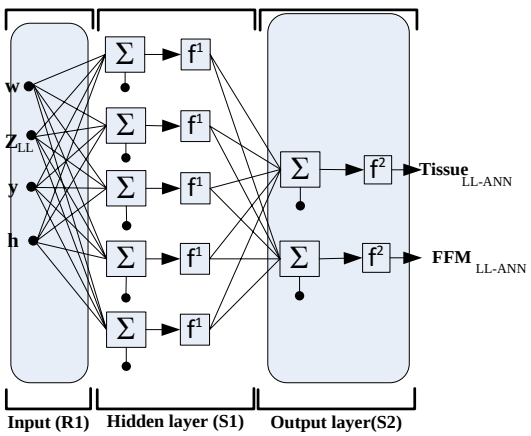


Fig. 2 The BP-ANN model in present study. The input layer (R1) contained four variabless as weight (w), age (y), height (h), bioelectrical impedances values in lower limbs (Z_{LL}). The hidden layer (S1) contained five neuron units, f^1 transfer functions for hidden layer and f^2 transfer functions for output layer. The output layer contained two neuron units to outcome the amount of tissue weight of lower limbs ($Tissue_{LL\text{-}ANN}$) and the amount of the FFM of lower limbs ($FFM_{LL\text{-}ANN}$).Σ, was the summation function.

Results

Subject characteristics

The subjects' basic physical characteristics, including age, weight, height and body mass index (BMI) were shown as Table 1. The range of subjects' age was from 18.0 to 25.6 years old. The range of body weight was from 56.4 to 86.0 kg and the range of BMI was from 21.3 to 28.7 kg/m^2 in twenty four elite male wrestlers.

Predictive equations of body composition in lower limbs by linear regression analysis

The linear regression equations Eq. 1 and Eq. 2 were obtained by the independent variables including the height (h, as cm), weight (w, as kg), age (y, as year) and bioelectrical impedances (Z, as ohm) and by the dependent variables as $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ values measured by DXA (Table 2), respectively.

Table 2. The body composition and bioelectrical impedance in lower limbs (n = 48)

Region & Item	Mean(SD)	Range
$FFM_{LL\text{-}DXA}$(kg)	11.56(1.46)	9.22-14.43
$Tissue_{LL\text{-}DXA}$(kg)	12.58(1.93)	9.69-16.62
$FM\%_{LL\text{-}DXA}$(%)	12.74(5.40)	6.41-24.52
Z_{LL} (ohm)	207.93(14.18)	181.30-240.10

$FFM_{LL\text{-}DXA}$, fat free mass in lower limbs by DXA; $Tissue_{LL\text{-}DXA}$, tissue mass in lower limbs by DXA; FM%, fat mass percentage in lower limbs by DXA; Z_{LL}, impedance value of lower limbs by BIA.

$$FFM_{LL\text{-}LR} = 3.052 + 0.031\ w - 0.169\ y + 0.069\ h^2/Z_{LL} \quad (r = 0.865,\ SEE = 0.756\ kg,\ P < 0.001) \tag{1}$$

$$Tissue_{LL\text{-}LR} = -0.884 + 0.145\ w - 0.149\ y + 0.044\ h^2/Z_{LL} \quad (r = 0.941,\ SEE = 0.674\ kg,\ P < 0.001) \tag{2}$$

h: height (cm), w: weight (kg), y: age (years), Z_{LL}: bioelectrical impedances (ohm), $FFM_{LL\text{-}LR}$: fat free mass (kg) of lower limbs by linear regression analysis, $Tissue_{LL\text{-}LR}$: tissue mass (kg) of lower limbs by linear regression analysis

Evaluating body composition in lower limbs by BP-ANN

The same variable data as above linear regression analysis were input into the input layer in a BP-ANN with training procedure to obtain the predictive value of $FFM_{LL\text{-}ANN}$ and $Tissue_{LL\text{-}ANN}$. The best weight matrix, bias vector, hidden layer weight matrix and layer bias vector after training procedure were presented as input weight matrix (Eq. 3), bias vector (Eq. 4), layer weight matrix (Eq. 5) and bias vector (Eq. 6), respectively.

$$\mathbf{W_1} = \begin{bmatrix} 0.0035 & -0.1263 & 0.1693 & 0.0538 \\ 1.2931 & 0.1385 & -0.7733 & 0.0883 \\ -1.8432 & 0.2779 & -0.2637 & 0.0389 \\ 1.0159 & -0.1346 & 0.0451 & -0.0162 \\ 0.1379 & 0.0855 & -0.0584 & -0.0408 \end{bmatrix}, \tag{3}$$

$$\mathbf{b}_1^{\mathrm{T}} = [-0.8764 \quad -0.0158 \quad -0.0365 \quad 0.2101 \quad -4.3942], \tag{4}$$

$$\mathbf{W}_2 = \begin{bmatrix} 3.0715 & 4.1573 & 0.5584 & -1.1294 & 4.2336 \\ 3.4663 & 2.9498 & -0.3831 & -4.1617 & 4.2052 \end{bmatrix}, \tag{5}$$

$$\mathbf{b}_2 = \begin{bmatrix} 3.7803 \\ 3.3562 \end{bmatrix}, \tag{6}$$

M: Matrices, as capital **BOLD** nonitalic letters, **b**: Vectors, as small **bold** nonitalic letters, **X**i: The"i"suffix, the series number of neuron, **T**: Transpose

The measurements of body composition in lower limbs

The predictive value of FFM_{LL} and $Tissue_{LL}$ by LR or ANN were showed as $FFM_{LL\text{-}LR}$ and $Tissue_{LL\text{-}ANN}$ in Table 3. The predictive data of FFM_{LL} by the regression equation of $FFM_{LL\text{-}LR}$ and $FFM_{LL\text{-}DXA}$ or by the $FFM_{LL\text{-}ANN}$ and $FFM_{LL\text{-}DXA}$ were presented in Fig. 3(a). The correlation coefficient (r) of $FFM_{LL\text{-}LR}$ and $FFM_{LL\text{-}ANN}$ by $FFM_{LL\text{-}DXA}$ were 0.865 and 0.965, respectively. The difference distribution (± 2 SD) between $FFM_{LL\text{-}LR}$ and $FFM_{LL\text{-}DXA}$ was from -1.464 to 1.464 kg, and the difference distribution between $FFM_{LL\text{-}ANN}$ and $FFM_{LL\text{-}DXA}$ was from -0.754 to 0.754 kg by Bland-Altman analysis in Fig. 4(a), respectively.

The predictive data of $Tissue_{LL}$ by the regression equation of $Tissue_{LL\text{-}LR}$ and $Tissue_{LL\text{-}DXA}$ or by the $Tissue_{LL\text{-}ANN}$ and $Tissue_{LL\text{-}DXA}$ were presented in Fig. 3(b). The correlation coefficients (r) of $Tissue_{LL\text{-}LR}$ and $Tissue_{LL\text{-}ANN}$ by $Tissue_{LL\text{-}DXA}$ were 0.941 and 0.981, respectively. The difference distribution between $Tissue_{LL\text{-}LR}$ and $Tissue_{LL\text{-}DXA}$ was from -1.304 to 1.304 kg, and the difference distribution between $Tissue_{LL\text{-}ANN}$ and $Tissue_{LL\text{-}DXA}$ was from -0.745 to 0.745 kg by Bland-Altman analysis in Fig. 4(b), respectively.

The correlations between independent and dependent variable

The correlations of independent variable between dependent $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ variable were presented as a matrix and most of them showed with highly correlation (Table 4). The correlation coefficient of Z_{LL} between $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ was -0.69 and -0.75; the correlation coefficient of "h" $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ was 0.77 and 0.76; and the correlation coefficient of "w" between $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ was 0.78 and 0.92, respectively. Only the correlation coefficient of "y" between $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$ showed 0.09 and 0.23 and was not with highly correlation. There was highly correlation between $FFM_{LL\text{-}DXA}$ and $Tissue_{LL\text{-}DXA}$.

Table 3. The predictive data of FFM and Tissue in lower limbs by linear regression and by BP-ANN model (n = 48).

Item	Mean(SD)	Range
$Tissue_{LL\text{-}LR}$ (kg)	12.51(1.81)	9.32-16.02
$FFM_{LL\text{-}LR}$ (kg)	11.59(1.26)	9.23-14.35
$Tissure_{LL\text{-}ANN}$ (kg)	12.57(1.88)	9.74-16.40
$FFM_{LL\text{-}ANN}$ (kg)	11.56(1.40)	9.09-14.28

$Tissue_{LL\text{-}LR}$, tissue mass in lower limbs by linear regression model; $FFM_{LL\text{-}LR}$, fat free mass in lower limbs by linear regression model; $Tissue_{LL\text{-}ANN}$, tissue mass in lower limbs by BP-ANN model; $FFM_{LL\text{-}LR}$, fat free mass in lower limbs by BP-ANN model.

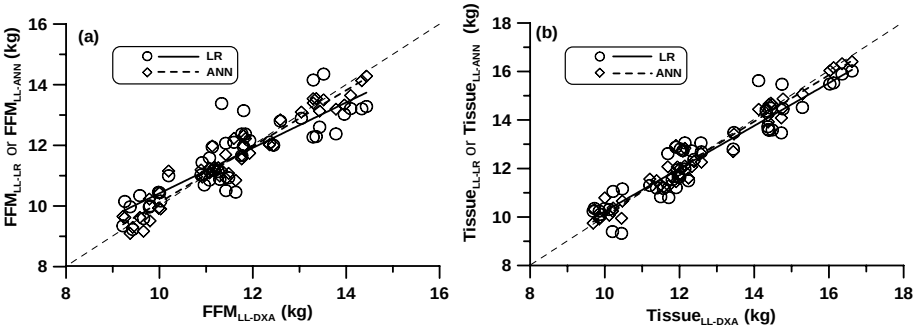


Fig. 3 Linear regressions of predictive data in FFM_{LL} and $Tissue_{LL}$. (a) The linear regressions of $FFM_{LL\text{-}LR}$ and $FFM_{LL\text{-}DXA}$ as well as $FFM_{LL\text{-}ANN}$ and $FFM_{LL\text{-}DXA}$. $FFM_{LL\text{-}LR}$ = 0.749 $FFM_{LL\text{-}DXA}$ + 2.924 (r = 0.865, P < 0.001), $FFM_{LL\text{-}ANN}$ = 0.925 $FFM_{LL\text{-}DXA}$ + 0.864 (r = 0.965, P < 0.001). (b) The linear regressions of $Tissue_{LL\text{-}LR}$ and $Tissue_{LL\text{-}DXA}$ as well as $Tissue_{LL\text{-}ANN}$ and $Tissue_{LL\text{-}DXA}$. $Tissue_{LL\text{-}LR}$ = 0.881 $Tissue_{LL\text{-}DXA}$ + 1.417 (r = 0.941, P < 0.001), $Tissue_{LL\text{-}ANN}$ = 0.958 $Tissue_{LL\text{-}DXA}$ + 0.532 (r = 0.981, P < 0.001).

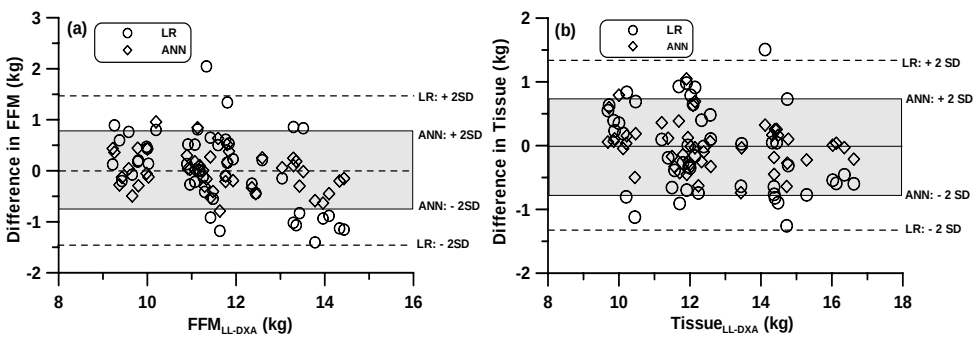


Fig. 4 The Bland-Altman plots in FFM_{LL} and $Tissue_{LL}$ by the linear regression and ANN model. (a) The Bland-Altman plots of $FFM_{LL\text{-}LR}$ and $FFM_{LL\text{-}DXA}$, SD = 0.732 kg, -2 SD = -1.464 kg, +2 SD =

1.464 kg. The Bland-Altman plots of $FFM_{LL\text{-}ANN}$ and $FFM_{LL\text{-}DXA}$, SD = 0.377 kg, -2 SD = -0.754 kg, +2 SD = 0.754 kg. (b) The Bland-Altman plots of $Tissue_{LL\text{-}LR}$ and $Tissue_{LL\text{-}DXA}$, SD = 0.652 kg, -2 SD = -1.304 kg, +2 SD = 1.304 kg. The Bland-Altman plots of $Tissue_{LL\text{-}ANN}$ and $Tissue_{LL\text{-}DXA}$, SD = 0.373 kg, -2 SD = -0.745 kg, +2 SD = 0.745 kg.

Table 4. The predictive data of FFM and Tissue in lower limbs by linear regression and by BP-ANN model (n = 48).

	Z_{LL}	h	m	y	$FFM_{LL\text{-}DXA}$	$Tissue_{LL\text{-}DXA}$
Z_{LL}	1.00					
h	-0.42**	1.00				
m	-0.78**	0.70**	1.00			
y	-0.10	0.19	0.37**	1.00		
$FFM_{LL\text{-}DXA}$	-0.69**	0.77**	0.78**	0.09	1.00	
$Tissue_{LL\text{-}DXA}$	-0.75**	0.76**	0.92**	0.23	0.90**	1.00

**, highly significant, $p < 0.01$

h, height (cm); w, weight (kg); y, age (years); Z, bioelectrical impedances (ohm); $FFM_{LL\text{-}DXA}$, fat free mass (kg) of lower limbs by DXA; $Tissue_{LL\text{-}DXA}$: tissue mass (kg) of lower limbs by DXA.

Discussion

This study applied a novel model for assessing the fat free mass in lower limbs (FFM_{LL}) in elite wrestling player by using a bioelectrical impedance analysis (BIA) with a Back Propagation Artificial Neural Network (BP-ANN). The results of by the traditional linear regression with the same data as the BP-ANN used were compared with that of by the BP-ANN. It was showed that the correlation coefficient of the BP-ANN referenced with DXA had better performance than that of traditional linear regression. In addition, the difference distribution of predictive values by the BP-ANN showed less range than that of by traditional linear regression.

Current application of BIA in estimation of body composition bases on parameters such as height, weight, age, gender, races and some specific cohorts to create prediction equations by linear regression analysis [36]. The created linear prediction equations with parameters tried to describe the relationship between independent variables and dependent variables [37]**.** The application of linear regression analysis on a single independent and a single dependant variable is a suitable way. Nevertheless, the multiple parameters, especially interaction exist between each other, will raise the standard deviation in prediction equation for their relationship is not ordinary linear. Artificial neural network (ANN) is a non-linear statistical data modeling or decision making tool, and can be used to model complex relationships between inputs and outputs or to find patterns in data. In this study, the results of the BP-ANN showed better performance than that of traditional linear regression could probably explain processing accuracy of non-linear data was better by using non-linear predictive models.

The predictive equations for measuring FFM_{LL} and $Tissue_{LL}$ must be separately established on 2-C model, and then the fat mass can be processed by the difference relationship between FFM_{LL} and $Tissue_{LL}$. However, one bias vector followed with one weight matrix can process both FFM_{LL} and $Tissue_{LL}$ by a BP-ANN. The superiority of the BP-ANN is not only on the accuracy but on its convenience.

Summary

Multiple parameters, especially interaction exist between each other, will raise the standard deviation in the predicted results for their relationship is not ordinary linear. Artificial neural network (ANN) is a non-linear statistical data modeling tool, and can be used to model complex relationships between

data. In this study, the BP-ANN model shown better performance than the traditional LR model in the prediction of the FFM_{LL} in elite wrestlers in Taiwan could be verified.

References

[1] T.J. Housha, W.G. Thorlanda, G.O. Johnsona, G.D. Tharpa and C.J. Cisara: J. Sports. Sci. Vol. 2 (1984), p.3

[2] N. Ebine, J.Y. Feng, M. Homma, S. Saitoh and P.J. Jones: Eur. J. Appl. Physiol. Vol. 83 (2000), p. 1

[3] J.A. Davis, J. Brewer and D. Atkin: J. Sports. Sci. Vol. 10 (1992), p. 514

[4] A.M. Luis, F.L. Juan, S. Ruth, I.M. Maria and F. Jesus: Nutr. Res. Vol. 24 (2004), p. 235

[5] M.D. Beekly, T. Abe, M. Kondo, T. Midorikawa and T. Yamauchi: J. Sports. Sci. Med. Vol. (2006), p.13

[6] C.M. Moldesky, H.K. Cureton, R.D. Lewis, B.M. Prior, M.A. Sloniger and D.A. Rowe: J. Appl. Physiol. Vol. 80 (1996), p. 2085

[7] B.M. Prior, C.M. Modlesky, E.M. Evans, M.A. Sloniger, M.J. Saunders, R.D. Lewis and K.H. Cureton: J. Appl. Physiol. Vol. 90 (2001), p. 1523

[8] H.G.J.M. Kuypers: *Anatomy of the descending pathways. In: Brooks VB, Brookhart JM, eds. Handbook of Physiology: The Nervous System II.* New York, NY: Am. Physiol. Soc. (1981), p.597

[9] J. Crosbie, R.Shepherd and T. Squire: J. Hum. Movement. Stud. Vol. 28 (1995), p.103

[10] C.M. Dean and R.B. Shepherd: *Stroke.* Vol. 28 (1997), p. 722

[11] E. Hübner-Woźniak, A. Kosmol, G. Lutoslawska and E.Z. Bem: J. Sci. Med. Sport. Vol. 7 (2004), p. 473

[12] K.J. Ellis: Physiol. Rev. Vol. 80 (2001), p. 649

[13] G. Sun, C.R. French, G.R. Martin, B. Younghusband, R.C. Green, Y.G. Xie, M. Mathews, J.R. Barron, D.G. Fitzpatrick, W. Gulliver and H. Zhang: Am. J. Clin. Nutr. Vol. 81 (2005), p.74

[14] M. Bussolotto, A. Ceccon and G. Sergi: Gerontology. Vol. 45 (1999), p. 39

[15] U.G. Kyle, I. Bosaeus, A.D. De Lorenzo, P. Deurenberg, M. Elia, J.M. Gómez, B.L. Heitmann, L. Kent-Smith, J.C. Melchior, M. Pirlich, H. Scharfetter, A.M. Schols and C. Pichard : Clin. Nutr. Vol. 23 (2004), p. 1430

[16] A.D. Stewart and W.J. Hannanl: J. Sports. Sci. Vol. 18 (2000), p. 263

[17] U. Svantesson, M. Zander, S. Klingberg and F. Slinde: J. Negat. Result. Biomed. Vol. 22 (2008), p.7

[18] M.R. Esco, M.S. Olson, H.N. Willford, S.N. Lizana and A.R. Russel: J. Strength. Cond. Res. Vol. 25 (2001), p.1040

[19] A. Tagliabue, A. Andreoli, S. Bertoli, E. Pagiato, M. Comelli, G. Testolin and D.E. Lorenzo: Ann. NY. Acad. Sci. Vol. 904 (2000), p. 218

[20] A. Pietrobelli, F. Rubiano, M-P. St-Onge and S.B. Hyemsfield: Eur. J. Appl. Physiol. Vol. 58 (2004), p. 1479

[21] W. Bell, D.M. Cobner and W.D. Evans: Ergonomics. Vol. 43 (2000), p. 1708

[22] R. Shih, Z. Wang, M. Heo, W. Wang and S.B. Heymsfield: J. Appl. Physiol. Vol. 89 (2000),

1380

[23] M. Miyatani, K. Kanehisa, I.M. Masuo and T. Fukunaga: J Appl. Phyisol. Vol. 91 (2001), p.386

[24] S.M. DiRusso, H.C. Suillivan, C. Holly, S.N. Cuff and J. Savino: J. Trauma. Vol. 49 (200), p. 212

[25] S.D. Izenberg, M.D. Williams and A. Luterman: Am. Surg. Vol. 63 (1997), p. 275

[26] D.R. Cox: J. R. Stat. Soc. Ser B Vol. 34 (1972), p. 187

[27] P. McCullagh and J.A. Nelder: *Generalized linear models.* 2nd ed. London: Champan and Hall, (1989)

[28] J.H. Watson, H.C. Sox, R.K. Neff and L. Goldman: N. Engl. J. Med. Vol. 313 (1985), p. 793

[29] J.S. Chiu, C.F. Chong, Y.F. Lin, C.H. Wu, Y.F. Wang and Y.C. Li: Am. J. Nephrol. Vol. 25 (2005), p. 507

[30] E.I. Mohamed, C. Maiolo, R. Linder, S.J. Pöppl and A. Lorenzo: Acta. Diabetol. Vol. 40 (2003), p. S15

[31] F. Ibrahim, T. Faisal, M.I. Salim and M. Taib: Med. Bio. Eng. Comput. Vol. 48 (2010), p. 1141

[32] R.B. Mazess, H.S. Barden, J.P. Bisek and J. Hanson: Am. J. Clin. Nutr. Vol. 51 (1990), p. 1106

[33] A.D. Lorenzo, S.P. Sorge, L. Iacopino, A. Andreoli, P.P. de Luca and G.F. Sasso: Appl. Radial. Iost. Vol. 49 (1998), p. 739

[34] L.W. Organ, G.B. Bradham, D.T. Gore and S.L. Lozier: J. Appl. Physiol. Vol. 77 (1994), p. 98

[35] J.M. Bland and D.G. Altman: Lancent. Vol. 8476 (1998), p. 307

[36] U.G. Kyle, I. Bosaeus, A.D. De Lorenzo, P. Deurenberg, M. Elia, J.M. Gómez, B.L. Heitmann, L. Kent-Smith, J.C. Melchior, M. Pirlich, H. Scharfetter, A.M. Schols and C. Pichard : Clin. Nutr. Vol. 23 (2004), p. 1226

[37] S.S. Guo, W.C. Chumlea and D.B. Cockram: Am. J. Clin. Nutr. Vol. 64 (suppl) (1996), p. 428S

The Establishment of Bioelectrical Impedance Analysis System with Neural Network Model to Estimate Segmental Body Compositions in Collegiate Wrestlers

Tsong-Rong Jang[1,a], Yu-Yawn Chen[2,b], Hsueh-Kuan Lu[3,c], Cai-Zhen Mai[4,d] and Kuen-Chang Hsieh[5,e*]

[1] Department of Combat Sports, National Taiwan University of Physical Education and Sport, Taiwan

[2]Department of Physical Education, National Taiwan University of Physical Education and Sport, Taiwan

[3] Sport Science Research Center, National Taiwan University of Physical Education and Sport, Taiwan

[4] Department of Ball Sports, National Taiwan University of Physical Education and Sport, LTD, Taiwan

[5] Research Center, Charder Electronic Co., LTD, Taiwan

[a]trong510315@yahoo.com.tw, [b]yu11.tw@yahoo.com.tw, [c]sk.lu2002@gmail.com, [d]tcmai@ntcpe.edu.tw, [e]abaqus0927@yahoo.com.tw

*Corresponding author.

Keywords: artificial neural network, dual-energy X-ray absorptiometry, fat free mass.

Abstract. To establish the precise estimation of FFM (fat free mass) of whole body, upper limbs, lower limbs and trunk in collegiate wrestlers, we created BIA (Bioelectrical impedance analysis) system by BP-ANN (Back Propagation Artificial Neural Network). The parameters of 24 elite wrestlers of their age, height, weight and bioelectrical impedance value of whole body and limbs were acted as input layer. The measured FFM of whole body, upper limbs, lower limbs and trunk as training data base as well as the estimation FFM of whole body and limbs as BP-ANN output layer. The obtained estimation data by above were compared to by BIA8 (8 contact electrode BIA of body composition analyzer). The correlation between the measured FFM in whole body, upper limbs, lower limbs and trunk by DXA and by BP-ANN are R^2 = 0.996, 0.853, 0.954, 0.945 and that by BIA8 are R^2 = 0.794, 0.374, 0.570, 0.628, respectively. In summary, the greater determination coefficients and smaller difference SD indicates the BIA system with BP-ANN model can estimate FFM in segments in wrestlers.

Introduction

The body composition and exercise capacities such as the maximal oxygen uptake, power out in running were depended on various types of sport items [1]. The relationships between the body composition and exercise capacities were well reported [2, 3]. The body composition can be a good index to predict not only the performance but also training outcomes. The data about the changes of body composition during training program can also be the index as training achievement [4, 5]. The current applications of BIA system on estimation of body composition are in not only the general population [5] but also the athletes [6-8]. However, the identical same system can't fit every specific physiologic status, especially, the totally different body compositions in various types of sport items. In other words, the specific system for estimation of body composition in athletes of specific sport items was needed [1]. To meet the requirement for current clinical application to estimated upper limbs, lower limbs and trunk [9, 10], BIA system with the multiple electrodes was developed instead of single circuit of hand-to-foot BIA estimation system in the past [11, 12]. Although the BIA with

eight electrodes at standing posture were well established in commercial devices, the specific model for the competitive athletes is not yet established and is needed.

The current mathematic outcome prediction adopted Cox regression [13], logistic regression [14], discriminate analysis, recursive partitioning [15] and artificial neural network-ANN [16]. Owing to the excellent performance in prediction for the outcomes with non-linear relationship of dependent variables, the ANN was wildly adopted in many fields [17, 18]. There are rarely adopted in the studies about the estimation of body compositions by BIA, except some about the estimation of intracellular fluid and body water content [19, 20].

We tried to develop the BIA system adopted with BP-ANN model to accurately estimate segmental body compositions in collegiate wrestlers. To validate the accuracy and precision of our developed system, the comparison of commercial BIA instrument with 8 electrodes for multiple segmental body composition was also completed.

Materials and methods

Subject

The 24 Taiwan elite male wrestlers with average disciplined training over 10.12 (±1.2), age at 19.3 (±0.9) within 18.0 to 25.6 years old, weight within 56.4 to 86.0 kg and BMI at 24.5 (±2.6) kg/m^2 within 21.3 to 28.7 kg/m^2. The subjects underwent strength and specific training over 12.1 hour per week. No strength training were loaded before one week, no diuretic agent was administered for seven days, no alcoholic beverages were consumed for 48 hours, and no urination for 30 minutes before the examination of BIA and DXA measurements was allowed. All of the volunteered subjects with their informed consent examined by the Institutional Review Board (IRB) of Advisory Committee at Jen-Ai Hospital in Taiwan were recruited under clearly informed about the details and possible risks during experiment.

Octa-contact electrode BIA of body composition analyzer

The BC-418 measurement (Tanita Corp, Tokyo, Japan, BIA8) selected with athletes mode by standing up at platform embedded with tetra-polar electrodes and by griped a handle embedded with stainless bi-polar electrodes. The bioelectric impedance value (Z) of whole body, head skull, upper limbs, lower limbs and trunk of subjects were measured by BC-418 [21, 22]. The Z values of each body segment of elite female runners from by BIA8 were combined with age, body height and each body segments weight parameters to develop our predictive equation for athletes' body composition.

Impedance measurement device

The BIA instrument (QuadScan 4000; Bodystat, Ltd., Isle of Man, UK), which contains independent detect electrodes and current source electrodes in the platform and handle grip, was modified to exhibit a changeable switch for different circuits and to connect to computer for data transferring. To confirm the no changes of accuracy and precision compared to the original instrument after our modification, all of data from the prior test and post test were verified carefully by impedance measuring instrument with high resolution. As shown in Fig. 1. The E2, E4, E6 and E8 were current electrodes and E1, E3, E5 and E7 were measuring electrodes. For the electric impedance between measuring electrodes are much greater than human body, the effects of measuring electrodes can be neglected.

Besides, the electrodes are made of stainless slice with great conductivity. E1, E2, E5 and E6 placed on hand grip, and that E3, E4, E7 and E8 on bilateral sides in plate. The bioelectrical impedances value (BIV) of left lower limb (termed as Left leg impedance, Z_{lleg}), right lower limb (Right leg impedance, Z_{rleg}), right upper limb (Right arm impedance, Z_{rarm}), left upper limb (Left arm impedance, Z_{larm}) and whole body (Whole body impedance, Z_{whole}) were obtained by changing the switch to form different circuits at 50kHz, 400μA current. For example, the Z_{rleg} obtained while measure between E3 and E7 electrodes at circuit between E2 and E4 electrodes; the Z_{lleg} obtained while measure between E6 and E8 electrodes at circuit between E3 and E7 electrodes [23].

Experimental procedures

After being measured the weight at error within 0.1kg and the height within 0.5cm, all of subjects, in cotton robe without any metal attachments is lying supine, scanned for whole body by DXA (Lunar Prodigy, GE Corp, USA.), with the "enCore 2003 Version 7.0" software, at 20μGy. The bone mineral density, fat mass (FM), FFM and tissue mass were measured. While been measured BIV by our system, subjects stand on the stainless electrodes with holding two electrodes.

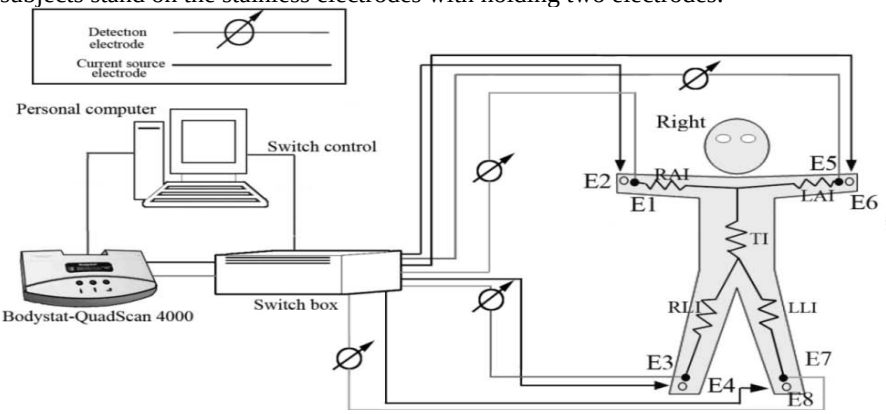


Figure 1. Measuring platform and bioelectric impedance measurement of improved system. E1, E3, E5, and E7, current electrode; E2, E4, E6, and E8, measuring electrode; LAI , left arm impedance (Z_{larm}); LLI, left leg impedance (Z_{lleg}); RAI, right arm impedance (Z_{rarm})); RLI, right leg impedance (Z_{rleg}); TI, trunk impedance.(Z_{whole}= RAI + TI + RLI).

Back Propagation – Artificial Neural Network (BP-ANN)

To establish the FFM estimation system of BP-ANN [24], three components of input layer, hidden layer and output layer were created (as shown in Fig. 2).

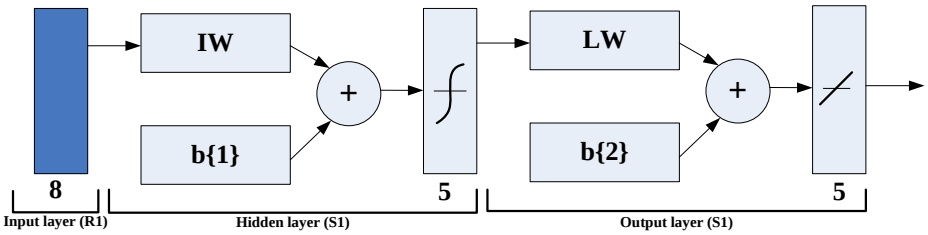


Figure 2. The BP-ANN in present study. The input layer (R1) contained four parameters as weight (m), age (y), height (h), bioelectrical impedances values in right lower limbs (Z_{rleg}),left lower limbs (Z_{lleg}), right upper limbs (Z_{rarm}), left upper limbs (Z_{larm}) and whole body (Z_{whole}). The hidden layer (S1) contained four neuron units. The output layer contained four neuron units to outcome the amount of appendicular and whole body of the FFM.

The 8 parameters of height (h), weight (w), age (y), Z_{whole}, Z_{larm}, Z_{rarm}, Z_{rleg} and Z_{lleg} construct input layer (R1). The relative training targets were the measured FFM by DXA in whole body ($FFM_{whole\text{-}DXA}$), right upper limbs ($FFM_{rarm\text{-}DXA}$), left upper limbs ($FFM_{larm\text{-}DXA}$), right lower limb FFM ($FFM_{rleg\text{-}DXA}$) and left lower limb ($FFM_{lleg\text{-}DXA}$). The output layer (S2) can yield $FFM_{whole\text{-}ANN}$, $FFM_{rarm\text{-}ANN}$, $FFM_{larm\text{-}ANN}$, $FFM_{rarm\text{-}ANN}$ and $FFM_{larm\text{-}ANN}$. The hidden layer, composed by 4 neuron units to connect the input layer and output layer, construct the transfer function f^1 as Hyperbolic Tangent log Sigmoid, and that, the output layer transfer function f^2 as linear. In order to optimize the best weight matrix and weight vector, we adopted the Levenberg-Marquardt algorithm for optimization of the network weights. The maximum Epochs were set at 200 times with the minimum gradient at 10^{-6}. All of programs mentioned above were composed by Matlab Ver.7.0 (MathWorks, Inc. MA).

Statistical analysis

The data all were analyzed by SPSS.12.0 software (SPSS Inc., Chicago, IL, USA). Data are shown as mean (±SD) (standard deviation). Values of $P<0.05$ were considered as significant. We used the Pearson Correlation Coefficient matrix to elucidate the relationship between the y, h, w, BIV of whole body and limbs vs. $FFM_{whole\text{-}DXA}$, $FFM_{rleg\text{-}DXA}$, $FFM_{lleg\text{-}DXA}$, ,$FFM_{rarm\text{-}DXA}$, $FFM_{larm\text{-}DXA}$ and $FFM_{trunk\text{-}DXA}$. The linear regression can be used to descript the degree of correlation between BIA8, ANN and DXA, and that, Bland-Altman Plots [25] point out the difference.

Results

Whole and segments of the impedance measurement

As shown in Table 1, the measured BIV of whole body, right upper limb (Z_{ram}) and left upper limb (Z_{larm}) were 483.36 (±32.99) 254.07 (±20.78) and 258.00 (±19.41) ohm, respectively. In pair t test, the α= 0.024 (<0.05) indicates the significant difference. The measured BIV of right lower limb (Z_{rleg}) and left lower limb (Z_{lleg}) were 207.14 (±13.86) and 208.74 (±14.75) ohm, respectively. In pair t test, α= 0.220 (>0.05) indicates no significant difference.

Table 1. Subject's whole body and limb segments of BIVs (n = 24)

Item	Mean(SD)	Range
Z_{whole} (ohm)	483.36(32.99)	434.33-542.67
Z^{*}_{rarm} (ohm)	254.07(20.78)	223.33-303.00
Z^{*}_{larm} (ohm)	258.00(19.41)	227.33-293.00
$Z^{\#}_{rleg}$ (ohm)	207.14(13.86)	184.33-234.67
$Z^{\#}_{lleg}$ (ohm)	208.74(14.75)	181.33-240.00

Z_{whole} is BIV of whole body, Z_{rarm} is BIV of right upper limb, Z_{larm} is BIV of left upper lim, Z_{rleg} is BIV of right lower limb, Z_{lleg} is BIV of left limb. (Z^{*}_{rarm} and Z_{larm} Paired T-Test, significant (two-tailed) of the value of 0.024 less than α = 0.05， Significant level of difference. $Z^{\#}_{rleg}$ and Z_{lleg} Paried T-Test, Significnat (two-tailed) of the value of 0.220 greater then α = 0.05， below the significant level difference.)

BP-ANN whole body and segments of the FFM estimation model

The h, w, y, Z_{whole}, $Z_{ram,}$ Z_{larm}, Z_{rleg} and Z_{lleg} input variables to BP-ANN modeling input hidden layer, the output layer of wrestling with the limbs of the FFM estimation model, the output layer were $FFM_{whole\text{-}ANN}$, $FFM_{lleg\text{-}ANN}$, $FFM_{rleg\text{-}ANN}$, $FFM_{larm\text{-}ANN}$ and $FFM_{rarm\text{-}ANN}$, BP-ANN after training, the best of the weight matrix and bias vector are:

Input weight matrix:

$$\mathbf{IW} = \begin{bmatrix} -0.06 & -0.06 & 0.07 & -0.09 & 0.01 & 0.10 & -0.03 & 0.12 \\ 1.45 & -0.00 & -0.02 & 0.47 & -0.07 & -0.53 & 0.01 & -0.48 \\ -3.62 & -0.74 & 1.28 & -1.13 & -0.07 & 1.60 & -0.39 & 1.71 \\ -2.58 & -0.57 & 0.97 & 2.92 & 0.60 & -3.37 & 0.88 & -3.85 \end{bmatrix},$$

Input bias vector:

$$\mathbf{b_1^T} = [2.31 \quad -5.51 \quad -0.42 \quad 3.06],$$

Layer weight matrix:

$$\mathbf{LW} = \begin{bmatrix} 12.31 & 17.17 & 14.11 & -16.83 \\ -1.08 & 3.84 & 4.41 & -3.75 \\ -1.32 & 3.65 & 4.51 & -3.89 \\ 3.49 & 1.03 & -0.05 & -1.16 \\ 3.27 & 0.97 & -0.50 & -1.06 \end{bmatrix},$$

layer bias vector:

$$\mathbf{b_2^T} = [16.43 \quad 3.32 \quad 3.29 \quad -0.47 \quad 0.21],$$

where: subscripts represent the layer (1 or 2), the superscript T is transpose.

Whole body and segments of the FFM measurements

The measured FFM of whole body, upper limb, lower limb and trunk by DXA, ANN and BIA8 were shown in Table 2. The measured FFM of trunk were defined to contain head part. The estimation of FFM of whole body, right upper limb, left upper limb, right lower limb, left lower limb and trunk by ANN were termed as $FFM_{whole\text{-}ANN}$, $FFM_{rleg\text{-}ANN}$, $FFM_{lleg\text{-}ANN}$, $FFM_{rarm\text{-}ANN}$, $FFMl_{arm\text{-}ANN}$ and $FFM_{trunk\text{-}ANN}$, respectively, similarly, that of by BIA8 as $FFM_{whole\text{-}BIA8}$, $FFM_{rleg\text{-}BIA8}$, $FFMl_{leg\text{-}BIA8}$, $FFM_{rarm\text{-}BIA8}$, $FFM_{larm\text{-}BIA8}$ and $FFM_{trunk\text{-}BIA8}$, respectively. The $FFM_{trunk\text{-}ANN}$ was yielded by following equation:

$$FFM_{trunk\text{-}ANN} = FFM_{whole\text{-}ANN} - FFM_{rleg\text{-}ANN} - FFM_{lleg\text{-}ANN} - FFM_{rarm\text{-}ANN} - FFM_{larm\text{-}ANN} \quad (1)$$

Table 2. DXA, BIA8 and ANN estimate body FFM results with the limbs segment

Item	Mean(SD)	Range
$FFM^{@}_{whole\text{-}DXA}$ (kg)	61.19(6.30)	48.99-76.16
$FFM^{*}_{arm\text{-}DXA}$(kg)	3.57(0.59)	5.11-2.30
$FFM^{*}_{leg\text{-}DXA}$(kg)	11.56(1.46)	9.22-14.43
$FFM^{@}_{trunk\text{-}DXA}$(kg)	26.50(2.85)	18.48-32.31
$FFM^{@}_{whole\text{-}BIA8}$ (kg)	61.55(6.13)	48.37-72.47
$FFM^{*}_{arm\text{-}BIA8}$(kg)	3.53(0.45)	2.40-4.47
$FFM^{*}_{leg\text{-}BIA8}$(kg)	12.32(1.49)	8.53-15.37
$FFM^{@}_{trunk\text{-}BIA8}$(kg)	29.86(2.62)	25.13-34.67
$FFM^{@}_{whole\text{-}ANN}$ (kg)	61.82(6.29)	49.04-76.13
$FFM^{*}_{arm\text{-}ANN}$(kg)	3.57(0.53)	4.94-2.72
$FFM^{*}_{leg\text{-}ANN}$(kg)	11.56(1.41)	9.18-14.37
$FFM^{@}_{trunk\text{-}ANN}$(kg)	31.59(2.85)	24.80-37.95

$^{@}n$ =24, $^{*}n$ = 48, subscript whole, whole, leg, trunk representing the whole body, upper and lower limbs and trunk (including head), DXA, BIA8, ANN represent the application DXA, BIA8 and ANN measurement and estimation results.

In Fig. 3, it presented the distribution and correlation between the measured FFM by DXA and estimated FFM by ANN or by BIA8. The determination coefficient (R^2) between $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}BIA8}$ is $R^2 = 0.794$ ($P< 0.001$) with the linear regression equation as $FFM_{whole\text{-}BIA8} = 0.867\ FFM_{whole\text{-}DXA} + 7.910$, and that of vs. $FFM_{whole\text{-}ANN}$ is $R^2 = 0.996$ ($P< 0.001$) with as $FFM_{whole\text{-}ANN} = 0.998\ FFM_{whole\text{-}DXA} + 0.072$. To clearly indicate the distribution of differences between $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}BIA8}$ and $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}ANN}$, we used Bland-Altman Analysis to obtain average difference between $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}BIA8}$ at -0.039 kg within 2SD from -0.985 to 0.908 kg as well as $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}ANN}$ at -0.001 kg, from -0.451 to 0.449kg (Fig. 4.(a)).

In Fig. 3.(b), it presented the distribution and correlation between the measured upper limb FFM by DXA and estimated FFM by ANN or by BIA8. The determination coefficient (R^2) between $FFM_{arm\text{-}DXA}$ vs. $FFM_{arm\text{-}BIA8}$ is $R^2 = 0.374$ ($P<0.001$) with the linear regression equation as $FFM_{arm\text{-}BIA8} = 0.469\ FFM_{arm\text{-}DXA} + 1.855$, and that of vs. $FFM_{arm\text{-}ANN}$ is $R^2= 0.853$ ($P<0.001$) with as $FFM_{arm\text{-}ANN} = 0.840\ FFM_{arm\text{-}DXA} + 0.567$. After being performed by Bland-Altman analysis, the average difference between $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}BIA8}$ at -0.039 kg within 2 SD from -0.985 to 0.908 kg as well as $FFM_{whole\text{-}DXA}$ vs. $FFM_{whole\text{-}ANN}$ at -0.001 kg, from -0.451 to 0.449kg (Fig. 4.(b)).

In Fig. 3.(c), it presented the distribution and correlation between the measured lower limb FFM by DXA and estimated FFM by ANN or by BIA8. The determination coefficient (R^2) between $FFM_{leg\text{-}DXA}$ vs. $FFM_{leg\text{-}BIA8}$ is $R^2 = 0.570$ ($P<0.001$) with the linear regression equation as $FFM_{leg\text{-}BIA8} = 0.772\ FFM_{leg\text{-}DXA} + 3.396$, and that of vs. $FFM_{leg\text{-}ANN}$ is $R^2 = 0.954$ ($P<0.001$) with as $FFM_{leg\text{-}ANN} = 0.949\ FFM_{leg\text{-}DXA} + 0.579$. After being performed by Bland-Altman Analysis, the average difference between $FFM_{leg\text{-}BIA8}$ vs. $FFM_{leg\text{-}DXA}$ at 0.761 kg within 2 SD from -1.305 to 2.827 kg as well as $FFM_{leg\text{-}ANN}$ vs. $FFM_{leg\text{-}DXA}$ at -0.001 kg, from -0.627 to 0.626kg (Fig. 4.(c)).

In Fig. 3.(d), it presented the distribution and correlation between the measured lower limb FFM by

DXA and estimated FFM by ANN or by BIA8. The determination coefficient (R^2) between $FFM_{trunk\text{-}DXA}$ vs. $FFM_{trunk\text{-}BIA8}$ is R^2 = 0.628 (P<0.001) with the linear regression equation as $FFM_{trunk\text{-}BIA8}$ = 0.705 $FFM_{trunk\text{-}DXA}$ + 7.595, and that of vs. $FFM_{trunk\text{-}ANN}$ is R^2 = 0.945 (*P*<0.001) with as $FFM_{trunk\text{-}ANN}$ = 0.941 $FFM_{trunk\text{-}DXA}$ + 1.853. After being performed by Bland-Altman Analysis, the average difference between $FFM_{trunk\text{-}BIA8}$ vs. $FFM_{trunk\text{-}DXA}$ at -1.697 kg within 2SD from -5.342 to1.948 kg as well as $FFM_{trunk\text{-}ANN}$ vs. $FFM_{trunk\text{-}DXA}$ at -0.001 kg, from -1.376 to 1.373kg (Fig. 4.(d)).

In Table 3, it has showed the correlation coefficient matrix between the input variables as age, height, weight and bioelectrical impedance values of whole body and limbs as well as the output variables as $FFM_{whole\text{-}DXA}$, $FFM_{rleg\text{-}DXA}$, $FFM_{lleg\text{-}DXA}$, $FFM_{rarm\text{-}DXA}$, $FFM_{larm\text{-}DXA}$, and $FFM_{trunk\text{-}DXA}$.
Five BIVs: Z_{whole}, Z_{rleg}, Z_{lleg}, Z_{rarm} and Z_{rarm}, between $FFM_{whole\text{-}DXA}$ exhibit great correlation coefficients as r= -0.854, -0.711, -0.711, -0.817 and -0.842, respectively, however, less correlation coefficients as r= -0.773, -0.705, -0.698, -0.621 and -0.723 exhibit between $FFM_{rleg\text{-}DXA}$. Five BIVs: Z_{whole}, Z_{rleg}, Z_{lleg}, Z_{rarm} and Z_{rarm}, between $FFM_{ram\text{-}DXA}$ exhibit correlation coefficients as r = -0.614, -0.414, -0.405, -0.712 and -0.683, respectively, and that, as r = -0.769, -0.631, -0.623, -0.851 and -0.779 between $FFM_{trunk\text{-}DXA}$. The other important factor, the weight, exhibits the correlation coefficients between $FFM_{whole\text{-}DXA}$, $FFM_{rleg\text{-}DXA}$, $FFMl_{leg\text{-}DXA}$, $FFM_{rarm\text{-}DXA}$, $FFM_{larm\text{-}DXA}$ and $FFM_{trunk\text{-}DXA}$ as r = 0.835, 0.794, 0.762, 0.626, 0.512 and 0.766.

Table 3. Anthropometric parameters, each BIV of limb segment, FFM of the correlation coefficient matrix (N = 24)

	y	h	w	Z_{whole}	Z_{rleg}	Z_{lleg}	Z_{rarm}	Z_{larm}	$FFM_{tk\text{-}DXA}$	$FFM_{la\text{-}DXA}$	$FFM_{ll\text{-}DXA}$	$FFM_{ra\text{-}DXA}$	$FFM_{rl\text{-}DXA}$	$FFM_{rl\text{-}DXA}$
y	1.00													
h	.194	1.00												
W	.372	.695**	1.00											
Z_{whole}	-.282	-.491*	-.802**	1.00										
Z_{rleg}	-.087	-.401	-.773**	.845**	1.00									
Z_{lleg}	-.118	-.436*	-.792**	.893**	.908**	1.00								
Z_{rarm}	-.308	-.511*	-.617**	.852**	.557**	.604**	1.00							
Z_{larm}	-.366	-.467*	-.717**	.948**	.685**	.707**	.924**	1.00						
$FFM_{tk\text{-}DXA}$	.253	.760**	.835**	-.854**	-.711**	-.711**	-.817**	-.842**	1.00					
$FFM_{la\text{-}DXA}$	.105	.768**	.794**	-.773**	-.705**	-.698**	-.621**	-.723**	.914**	1.00				
$FFM_{ll\text{-}DXA}$	.077	.770**	.762**	-.752**	-.658**	-.683**	-.612**	-.697**	.910**	.984**	1.00			
$FFM_{ra\text{-}DXA}$	.523**	.416*	.626**	-.614**	-.414*	-.405*	-.712**	-.683**	.747**	.492*	.507*	1.00		
$FFM_{la\text{-}DXA}$	.566**	.282*	.512**	-.559**	-.389	-.357	-.585**	-.626**	.614**	.375	.369	.917**	1.00	
$FFM_{rl\text{-}DXA}$	.229	.741**	.766**	-.769**	-.631*	-.623**	-.851**	-.779**	.927**	.763**	.738**	.695**	.538**	1.00

**. when the significance level of 0.01(two-tailed), the relevant significant.

*. when the significance level of 0.05(two-tailed), the relevant significant.

where ; $FFM_{whole\text{-}DXA}$ is $FFM_{tk\text{-}DXA}$, $FFM_{rleg\text{-}DXA}$ is $FFM_{ll\text{-}DXA}$, $FFM_{lleg\text{-}DXA}$ is $FFM_{ra\text{-}DXA}$, $FFM_{rarm\text{-}DXA}$ is $FFM_{la\text{-}DXA}$, $FFM_{larm\text{-}DXA}$ is $FFM_{la\text{-}DXA}$, $FFM_{trunk\text{-}DXA}$is $FFM_{rl\text{-}DXA}$。

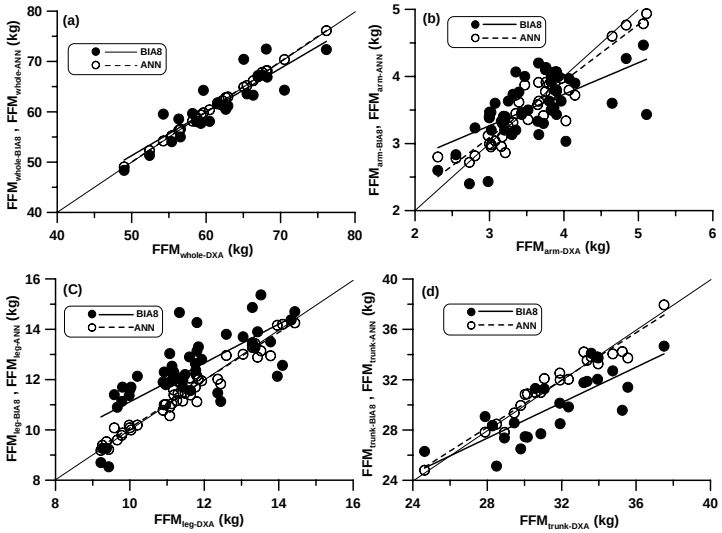


Fig 3. ANN and BIA8 estimates of whole body and segments of the FFM_{DXA} measurement results and correlation.

(a) whole body: $FFM_{whole-BIA8} = 0.867\ FFM_{whole-DXA} + 7.910$ ($R^2 = 0.794$, $P < 0.001$), $FFM_{whole-ANN} = 0.998\ FFM_{whole-DXA} + 0.072$ ($R^2 = 0.996$, $P < 0.001$)

(b)arm: $FFM_{arm-BIA8} = 0.469\ FFM_{arm-DXA} + 1.855$ ($R^2 = 0.374$, $P < 0.001$), $FFM_{arm-ANN} = 0.840\ FFM_{arm-DXA} + 0.567$ ($R^2 = 0.853$, $P < 0.001$)

(c)leg: $FFM_{leg-BIA8} = 0.772\ FFM_{leg-DXA} + 3.396$ ($R^2 = 0.570$, $P < 0.001$), $FFM_{arm-ANN} = 0.949\ FFM_{arm-DXA} + 0.579$ ($R^2 = 0.954$, $P < 0.001$)

(d)trunk : $FFM_{trunk-BIA8} = 0.705\ FFM_{trunk-DXA} + 7.595$ ($R^2 = 0.628$, $P < 0.001$), $FFM_{trunk-ANN} = 0.941\ FFM_{arm-DXA} + 1.853$ ($R^2 = 0.945$, $P < 0.001$)

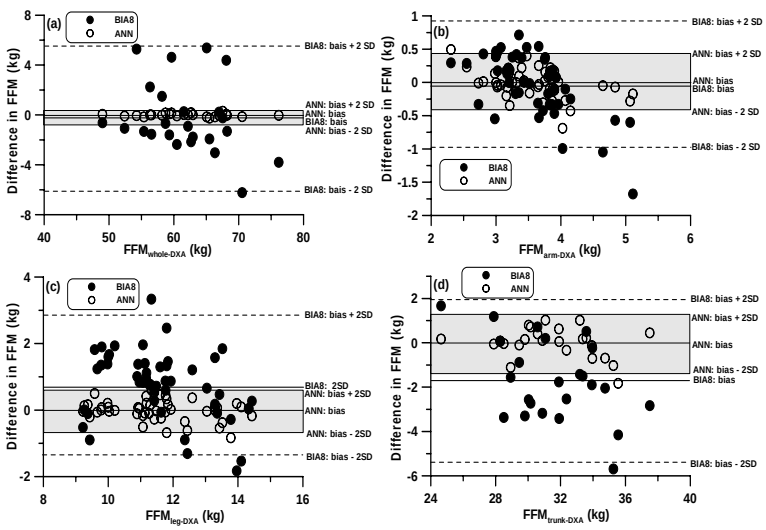


Fig 4. ANN and BIA8 estimates of whole body and segments of the FFM_{DXA} measurement results and distribution.

(a)Whole body: $FFM_{whole-BIA8} - FFM_{whole-DXA}$: bias = -0.278 kg, SD = 2.902 kg, bias -2 SD = -6.082 kg, bias +2 SD = 5.527 kg. $FFM_{whole-ANN} - FFM_{whole-DXA}$: bias = -0.004 kg, SD = 0.126 kg, bias -2 SD = -0.256 kg, bias +2 SD = 0.248 kg

(b) Arm: $FFM_{arm\text{-}BIA8} - FFM_{arm\text{-}DXA}$: bias = -0.039 kg, SD = 0.473 kg, bias -2 SD = -0.985 kg, bias +2 SD = 0.908 kg ◦ $FFM_{arm\text{-}ANN} - FFM_{arm\text{-}DXA}$: bias = -0.001 kg, SD= 0.225 kg, bias -2 SD = -0.451 kg, bias +2 SD = 0.449 kg

(c) Leg: $FFM_{leg\text{-}BIA8} - FFM_{leg\text{-}DXA}$: bias = 0.761 kg, SD = 1.033 kg, bias -2 SD = -1.305 kg, bias +2 SD = 2.827 kg. $FFM_{leg\text{-}ANN} - FFM_{leg\text{-}DXA}$: bias = -0.001 kg, SD= 0.313 kg, bias -2 SD = -0.627 kg, bias +2 SD = 0.626 kg

(d) Trunk: $FFM_{trunk\text{-}BIA8} - FFM_{trunk\text{-}DXA}$: bias = -1.697 kg, SD= 1.823 kg, bias -2 SD = -5.342 kg, bias +2 SD = 1.948 kg. $FFM_{trunk\text{-}ANN} - FFM_{trunk\text{-}DXA}$: bias = -0.001 kg, SD= 0.687 kg, bias -2 SD = -1.376 kg, bias +2 SD = 1.373 kg

Discussion

This study was to assess the segmental FFM by using a BIA with BP-ANN and referenced with DXA in elite male wrestling player. The accuracy of predictive results by the ANN models was compared with directly results of by the general BIA8. It was showed that the determination coefficient of the BP-ANN referenced with DXA had better performance than that of BIA8. In addition, the difference distribution of predictive values by the BP-ANN showed less range than that of by the BIA8. Artificial neural network is a non-linear statistical data modeling and can be used to model complex relationships between inputs and outputs in data. In this study, the results of the BP-ANN showed better performance than that of traditional linear regression could probably explain processing multiple interaction parameters was better by using the non-linear ANN model.

By judging the correlation coefficient matrix about the r between the $FFM_{whole\text{-}DXA}$ and Z_{whole}, Z_{rleg}, Z_{lleg}, Z_{rarm}, and Z_{rarm}, the range within -0.711 to -0.854 indicated that the BIV of above have high negative correlation to FFM_{whole}. In spite of both of BP-ANN and BIA8 contained height (h), weight (w), age (y) and BIV of whole body (Z_{whole}), other four important parameters as Z_{rleg}, Z_{lleg}, Z_{rarm} and Z_{rarm} were considered in BP-ANN. Besides, the interactions of all of parameters were well considered in hidden layer of BP-ANN rather than the independent relationship between parameters all each other in traditional linear regression, especially, the complex physiologic phenomenon. Artificial neural network is a non-linear statistical data modeling and can be used to model complex relationships between inputs and outputs in data. In other words, the calculations of interactions between physiologic parameters all in BP-ANN more agree to the reality of natural creatures than description of the simple linear regression

No matter what the whole body, upper limbs, lower limbs and trunk were, the greater R values between measured FFM by DXA vs. estimated FFM by BP-ANN than that of vs. by BIA8 were observed. The line of the best fit (y = x) of the measured FFM by DXA itself is the ideal line. From our data, the linear regression lines of estimated FFM by BP-ANN are almost overlapped to line of the best fit (y = x) of the measured FFM by DXA, however, that of by BIA8 are deviated much far from. In other words, the slope and regression coefficient between measured FFM by DXA vs. estimated FFM by BP-ANN, which are the index about the relationship between two variables, are more close to 1.0. From Bland-Altman analysis in the FFM of whole body, upper limbs, lower limbs and trunk, the differences of average and differences of SD between measured FFM by DXA vs. estimated FFM by BP-ANN are greater than that of by BIA8. It can also indicate the greater performance in prediction of body composition. In spite of the greater performance in prediction of body composition by BP-ANN than by BIA8, there are some different performances in prediction of body composition for whole body, lower limbs and trunks. While considered the R values between FFM_{whole} and other segmental FFMs, the lower R values between input variables and output variables were obtained, the lower correlation and greater differences between measured FFM by DXA vs. estimated FFM by BP-ANN were obtained. On the contrary, the greater R^2 values between input variables and output variables were obtained, the higher correlation and lower differences between measured FFM by DXA vs. estimated FFM by BP-ANN were obtained, especially, in $FFM_{whole\text{-}ANN}$.

Summary
In this study, we applied the simple Back Propagation Artificial Neural Network with single hidden layer to bioelectrical impedance analysis system for predicting the FFM of whole body and multiple segmental limbs in collegiate wrestlers, and that, with greater performance than that of BIA8. The greater R values, differences of average and differences of SD by BP-ANN also indicate the applicable properties.

References

[1] K.R. Segal : Am. J. Clin. Nutr. Vol. 57 (1996), p. 469S
[2] K.J. Ellis : Physiol. Rev. Vol. 82 (2000), p. 649
[3] D. Brodie , V. Moscrip and R. Hutcheon, : Nutrition. Vol. 14 (1998), p. 296
[4] U.G. Kyle, I. Bosaeus, A.D. De Lorenzo, P. Deurenberg, M. Elia, J.M. Gómez, B.L. Heitmann, L. Kent-Smith, J.C. Melchior, M. Pirlich, H. Scharfetter, A.M. Schols and C. Pichard : Clin. Nutr. Vol. 23 (2004), p. 1226
[5] G. Sun, C.R. French, G.R. Martin, B. Younghusband, R.C. Green, Y.G. Xie, M. Mathews, J.R. Barron, D.G. Fitzpatrick, W. Gulliver and H. Zhang: Am. J. Clin. Nutr. Vol. 81 (2005), p. 74
[6] A.D. Stewart and W.J. Hannanl: Vol. 18 (2000), p. 263
[7] A.C. Utter, D.C. Nieman, G.J. Mulford, R. Tobin, S. Schumm, T. McInnis and J.R. Monk: Med. Sci. Sports. Exerc. Vol. 37 (2005), p. 1395
[8] U. Svantesson, M. Zander, S. Klingberg and F. Slinde: J. Negat. Results. Biomed. 2 Vol. 2 (2008). 7:1
[9] D. Bracco, D. Thiébaud, R.L. Chioléro, M. Landry, P. Burckhardt and Y. Schutz: J. Appl. Physiol. Vol. 81 (1996), p. 2580
[10] S.P. Stewart, P.N. Bramley, R. Heighton, J.H. Green, A. Horsman, M.S. Losowsky and M.A. Smith: Br. J. Nutr. Vol. 69 (1993), p. 645
[11] G. Bedogni, M. Malavolti, S. Severi, M. Poli, C. Mussi, A.I. Fantuzzi and N. Battistin: Eur. J. Clin. Nutr. Vol. 56 (2002), p. 1143
[12] G. Medici, C. Mussi, A.L. Fantuzzi, M. Malavolti, A. Albertazzi and G. Bedgni: Eur. J. Clin. Nutr. Vol. 59 (2005), p. 932
[13] D.R. Cox: J. R. Stat. Soc. Ser B Vol. 34 (1972), p. 187
[14] P. McCullagh and J.A. Nelder: *Generalized linear models.* 2nd ed. London: Champan and Hall, (1989)
[15] J.H. Watson, H.C. Sox, R.K. Neff and L. Goldman: N. Engl. J. Med. Vol. 313 (1985), p. 793
[16] W.G. Baxt: Lancet. Vol. 346 (1995), p. 1135
[17] A. Vellido, P.J.G. Lisboa and J. Vaughan: Expert. Syst. Appl. Vol. 17 (1999), p. 51
[18] A.S. Chen, M.T. Leung and H. Daouk: Computers & Operations Research. Vol. 30 (2003), p. 901
[19] J.S. Chiu, C.F. Chong, Y.F. Lin, C.H. Wu, Y.F. Wang and Y.C. Li: Am. J. Nephrol. Vol. 25 (2005), p. 507
[20] E.I. Mohamed, C. Maiolo, R. Linder, S.J. Poppl and A.D. Lorenzo: Acta. Diabetol. Vol. 40 (2003), p. S15
[21] A. Piettoi, F. Rubiano, M.P. St-Onge and S.B. Heymsfield: Eur. J. Clin. Nutr. Vol. 58 (2004), p. 1479
[22] M.G. Shaikh, N.J. Crabtree, N.J. Shaw and J.M.W. Kirk: Horm. Res. Vol. 68 (2007), p. 8
[23] L.W. Organ, G.B. Bradham, D.T. Core and S.L. Lozier: J. Appl. Phyiol. Vol. 7, p. 98
[24] M.T. Hagan, H.B. Demuth and M. Beale: *Neural Network Design*, Thomson Learning, Inc. (1996)
[25] J.M. Bland and D.G. Altman: Lancent. Vol. 8476 (1998), p. 307

Automated Text Illustrator Based on Keyword Sense Tagging

Savindhi Samaraweera[1] and Ravindra Koggalage[2]

[1]University of Moratuwa, Katubedda, Sri Lanka

[2]General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

{[1]sssamaraweera@yahoo.com, [2]koggalage@yahoo.com}

Keywords: Text illustrator

Abstract. The concept of automated text illustrator has been around for a while. These systems take in an article, identify the keywords of the article and outputs images from an image database based on the keywords. Such systems help writers and journalists to identify images that best suits their texts. Most of these systems have given more emphasis on the image retrieval component and less emphasis on the text processing component. Our attempt is to concentrate on the text processing component, so that it can be used by existing text illustrators. We try to identify the correct meaning of the keywords based on the context of the verb it appears. We also identify the best possible word that represents the identified meaning.

Introduction

In journalism it is important that writers accompany their articles with images that depict the substance of the text. To find images normally they would use image search engines. It would be remarkable if an automated system exists, which can process writers' written stories and be able to identify the keywords of the stories and then search an image database for a set of appropriate images. The solution is an automatic text illustrator.

Figure 1 depicts the functional flow of an automated text illustrator. The basic flow includes mainly two components: processing of the article and image pool retrieval. The first component extracts meaningful words from the article, which represent the substance of the article. The second component is to extract images from an image database. Our system concentrates on the text processing capabilities of the system. SPE [1] and TTP [2] are two examples under this topic. Most of the text illustrator systems have fallen short of the text processing component. Therefore, our attempt is to improve the text processing component with the existing image retrieval techniques that would give significant results for the existing automated text illustrator systems.

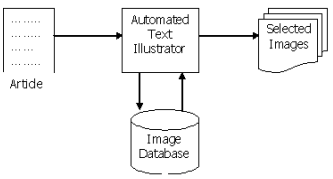


Figure 1: Functional flow of an automated text illustrator

Common mechanism of Information Retrieval systems, is to consider only the keywords that are been typed in. Some keywords have number of senses, and could be misleading. For example, 'case' has multiple senses – a container, an instance, etc. To differentiate and understand the meanings, humans normally consider the whole or part of the sentence. Another important concept in understanding the meaning of a word is to identify in which action context the word appears. For example consider 'play cricket' and 'jump cricket'. It is quite evident that in these two phrases the word 'cricket' means two completely different senses. In both these cases humans are able to differentiate the meanings based on the action performed by the object. Therefore, we propose to enhance the text processing component of automated text illustrator systems, by tagging correct senses to the keywords that are identified. For this we propose to consider the whole or part of the sentences and to identify the nouns and verbs of the sentences.

The hardest level in an NLP model would be the semantic analysis level, which relies on knowing the meaning of individual words, how the meaning of individual words combine to form the meaning of group of words and how it all fits in with the meaning of the sentence. This level includes word sense disambiguation (WSD). In order to solve issues in the semantic level, linguists consider the semantic relations between words. WordNet [3] is a research project, which attempts to model a lexical reference system. The system is a manually made database of lexical semantic relations and can be used as a dictionary, a lexical reference system, etc., with relationships, such as synonymy, hypernymy, etc. WSD is the process of determining the most relevant sense that applies to a word depending on the instance the word is been used. WordNet is extensively used in research related to WSD, because of its variety of detailed lexical relations.

Methodology

Figure 2 depicts the data flow of the overall system. The article is first fed to the system, which in turn forwards the article to the external keyword extractor, to identify the initial pool of keywords. These identified keywords are processed by the WSD component to identify the correct meanings of the keywords. For this we need to use the verb identifier to identify the verb of the concerned keyword. A semantic calculation is performed in the WSD component in order to identify the most probable meaning for the keyword. Then the lemma identifier identifies the probable synonymous word that should be used, which best represents the identified meaning. The lemma identifiers will constitute of the final set of keywords, which is forwarded to the image database in order to retrieve images.

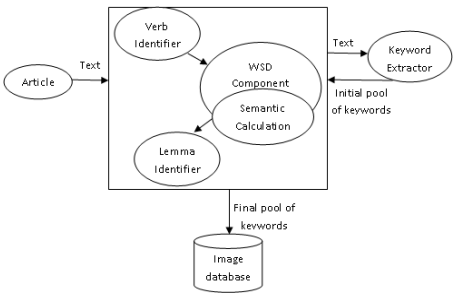


Figure 2: Data flow of the system

In this implementation we have considered words as elements of a sentence. We can say that a verb of a sentence plays a prominent role when trying to identify the correct sense of a word. Further, we could narrow down this idea to present that the meaning of a noun can be identified based on the verb associated with the noun. Therefore, to disambiguate the meaning of a given noun, we use the

main verb of its sentence. As the system would be processing (noun, verb) pairs we need to make sure that all keywords are paired up with appropriate verbs. Therefore, we have introduced the verb identifier component, which tries to identify the verb in several phases. The first phase is to identify verbs, from the identified keywords itself. Some of the keywords that are returned by the keyword extractor system contain phrases. For example the keyword 'shaved head' is used for the WSD component as (head, shave). Therefore, we check in the returned keyword set whether there are any verbs that accompany the noun. If so, then we use them as the (noun, verb) pair for the WSD component. For the keywords that such verbs do not exist, we use collocations to identify verbs. A collocation is a sequence of words that occur together unusually often. For example 'take care' is a collocation, where as 'do care' is not. To identify collocations we use Brown corpus. In these collocation entries we check whether the keyword exist, and if so, we check whether it exist with a verb, and check whether this verb exist in the sentence. For example the phrase 'unwed mother' can be found. If the word 'unwed' exist in the sentence with the word 'mother', then we could consider (mother, unwed) pair for the WSD component. The third phase of identifying verbs is to consider verbs from the sentence context. We pos tag all the words in the sentences that contain the keyword. Then based on basic linguistic rules we try to identify the main verb of the sentences. Still, if we cannot identify a suitable verb for the keyword then from the pre-compiled file of (noun, verb), we identify the verbs that could exist with the keyword, and then validate whether it exist in the sentence. If so then we use those (noun, verb) pairs for the WSD component.

The WSD component is based on [4] and is processed under four phases. The first phase is to identify if the noun in the (noun, verb) pair has only one single meaning. If so then that word will always hold one meaning, whatever the context of the rest of the sentence means. If the relevant noun contains multiple senses, then we need to evaluate which sense best suites the current context of the word. Therefore, in phase two we try to identify whether the noun contains a semantic similarity with another candidate noun in the noun's verb context. If a semantic similarity is identified between the noun and the candidate noun that exist with the verb, then we consider the meaning of that noun exist with the current verb context. To identify the candidate nouns that can exist with the verb we use a pre-compiled (noun, verb) file. For each candidate noun identified in the file, we check whether a semantic similarity exist between the noun and the candidate noun. If so, we interpret the considered noun's synset as a candidate synset. In phase three, we check whether any candidate verb exist within the noun context, and then check whether there is a semantic similarity with the verb and the candidate verb. To identify candidate verbs we use the pre-compiled file. After identifying that a similarity exists with the candidate verb, we process the (noun, candidate_verb) pair. If a resultant is found by processing (noun, candidate_verb) pair, the synset of the noun is considered as a candidate synset. Still, if the system is unable to find a correct sense for the noun, in phase four it will identify a separate (candidate_noun, candidate_verb) pair and check whether semantic similarity exist between the verb and the candidate verb as well as the noun and the candidate noun. In this method we first check whether semantic similarity exists with the noun and the candidate noun. If so, we then check whether a similarity exists between the verb and the candidate verb. If so, we identify the resultant synset between the noun and the candidate noun as to be the correct synset.

The system considers four semantic relations out of the large amount of semantic and lexical relations that exist in WordNet. Synonymy, hypernymy, hyponymy and coordinate relationships are considered. Synonyms were chosen as this would be the basic and obvious semantic similarity between two words. Hypernym and hyponym relationships were chosen as the parent and child synsets will denote a generalised and specialized concept of the concerned synset. Coordinate relationships were considered because, as peer synsets would have meaningful attributes of the concerned synset.

The WSD algorithm can retrieve multiple synsets within a phase. We need to identify only one synset, which would represent the most probable meaning of a word. To solve this issue, we extended the system into a similarity calculation component. We have experimented with different types of similarity calculation measures in order to choose the best calculation method that suite the system. In phases two to four we would be calculating the semantic relationship of hypernymy, hyponymy and coordinate relationship. We consider the noun synset and the candidate noun synset and perform the similarity measures. Based on the retrieved similarity measures we identify the maximum similarity measure that was retrieved and consider the synset with that measure to be the synset, which correctly represents the meaning of the noun. If we encounter multiple synsets with the same maximum value then the summation of the values relevant synsets are considered in order to identify the maximum value that is been returned. We have employed four similarity calculation measurements: path distance similarity, Leacock Chodorow similarity [5], Resnik similarity [6] and Jiang-Conrath similarity [7]. All these similarity calculation methods use synset pairs in order to perform the calculations. Path distance similarity measure returns a score denoting how similar two word synsets are based on the shortest path that connects the senses in the hierarchy. The path length is measured in nodes. If multiple paths exist between the two synsets then the shortest path is selected. Leacock Chodorow similarity measure returns a score denoting how similar two synsets are, based on the shortest path that connects the senses and the maximum depth of the hierarchy in which the synsets occur. Maximum depth of the hierarchy is the longest distance between the root and any leaf of the concerned hierarchy. Resnik presents a new measure of semantic similarity in a hierarchy, based on the notion of information content. This returns a score denoting how similar two word senses are, based on the information content of the least common subsumer that is the most specific ancestor node for the synsets that are considered in the similarity calculation. Jiang-Conrath similarity measure returns a score denoting how similar two word senses are, again based on the information content of the least common subsumer and the two input synsets.

In the lemma identifier module we identify the commonest or the most used word that describes a particular sense. This was considered as a needed component because some words are commonly not used in English language, and as we are using these words in order to search an image database it would be appropriate if we could identify the commonest word that is been used to denote the meaning. For example consider the ‘bass’, which contains a fish sense. There would hardly be images that are annotated by the ‘bass’ that depicts fish, but these images would definitely be annotated with the ‘fish’. Therefore, we believe it is important to have a component to identify the commonest used word for a meaning. To identify the commonest used word we use the frequency distribution of the words in Brown corpus. All the lemma names or simply synonyms, in the identified synset are extracted. Then using the constructed frequency distribution we identify the frequency of each lemma name and select the lemma name with the highest frequency.

Evaluation

We have disintegrated our evaluation into smaller components in order to identify strengths and weaknesses of each component. Therefore, our first experiment was based on the WSD component. The functionality of the WSD component is to identify the correct meaning of a noun for a given verb context in a (noun, verb) pair. The input of the component would be a (noun, verb) pair and the output would be a synset denoting the meaning of the noun for the given verb context. For each identified synset, we have calculated the outcome based on similarity calculation measures in order to identify the most probable synset, if multiple synset outputs exist from the WSD component. We have used four similarity calculations: path distance similarity, Leacock Chodorow similarity, Resnik similarity and Jiang-Conrath similarity. Out of these similarity measures Jiang-Conrath similarity proved to be the most successful, shown in table 1. The results show the number of

examples that were able to identify the correct meaning of the noun based on the verb, how many were identified incorrectly and how many examples were not identified.

If we consider the overall test results they show that all four phases have been utilised to identify the results. About 60% of the results had been resulted by phase 2, which is identifying the semantic relationship between the noun and the candidate noun. From the test results, 28% of the experiments had given incorrect senses. This is mainly due to identifying incorrect (noun, verb) pairs. According to the results most of the verbs were identified by the pre-compiled file with (noun, verb) pairs. Compilation of the file was not done manually, but done programmatically based on basic English grammar. During this compilation incorrect (noun, verb) pairs might have got entered and due to this reason the system will not function as expected. Therefore, to minimise the errors that occur due to this reason, it would be better to manually go through the set of identified (noun, verb) pairs and validate the entries. Another possible solution for this would be to have separate (noun, verb) pair based on domains, such as religion, sports, etc. Then each (noun, verb) file will only contain verbs that are relevant to that domain, which would reduce the probability of erroneous sense identification.

The next evaluation was performed for the lemma identifier component. The lemma identifier component was used in order to identify the best possible word that describes the meaning. The synonymous words of the synset are considered based on the frequency distribution of the words and collocations in Brown corpus. Table 2 shows results of the lemma identifier component. The whole idea of introducing the lemma identifier component was to find out the most commonly used word that best describes the meaning, based on the frequency distribution of the synonymous words, but still it gives incorrect results of 28%. Some of the results imply that after disambiguating the sense correctly, during the lemma identifier we again ambiguously name the keywords by incorrect wording. For example the Synset('book.n.02') contains synonymous values 'book' and 'volume'. The lemma identifier identifies 'book' as the correct wording, but still, if we had used a different corpus there is a possibility of identifying the word 'volume' as the lemma name for Synset('book.n.02'). We believe the solution for this would be to have different corpuses for different domains.

Table 1: Test results of Jiang-Conrath similarity measure

	No. of (noun, verb) pairs
Correct senses	72%
Incorrect senses	22%
Non identified words	6%

Table 2: Test results of lemma identifier

	No. of words
Correctly identified lemma names	72
Incorrectly identified lemma names	28

Next we performed an evaluation of the overall system. We have taken the output of hundred short articles and evaluated the sense tagging outcome of the system. We have considered the number of keywords that are given by the external system and the number of senses that were correctly tagged. The overall accuracy of the system is 66%. Further investigating we found out that most of the articles that had correctly tagged keywords contained about two to four keywords in the initial pool. This could mean that the initial keyword list might have contained unnecessary keywords, such as 'sustainable high quality', 'square miles', etc, were removed by the system as inappropriate keywords.

Another issue that we came across is that during the verb identifier for a noun, correct verbs of the sentences were not identified. The main reason for this is that some of the verbs are already contained in the stopwords list. The stopword list contains verbs, such as 'has', 'be', etc, and these

verbs do appear as main verbs of sentences. For example, consider the two sentences 'I had a guitar' and 'I am playing a guitar'. The system would correctly identify 'playing' as the main verb of the second sentence, but will be unable to identify 'had' as a main verb, as it is already removed from the stopword list.

Conclusion and Future Work

According to the evaluation results, the text illustrator system performs well even with some identified shortcomings. Therefore, we believe that the system could be further enhanced in order to give better results. One essential drawback of the system was the pre-compiled file of (noun, verb) pairs. This resulted mainly for the inaccuracies that were identified in the results. The file was generated based on the existing Brown corpus. Therefore, it would be useful if the (noun, verb) pairs were accurate and upto date. A manual validation of the entries of the file would be appropriate. In order to solve the above problem, we could extend our system to extract (noun, verb) pairs from articles according to domains, such as sports, politics, etc, and to save them in different files according to domains. So during the verb identification period, the system would only utilise the file that is relevant to the domain, and identify verbs that are relevant to the domain. We could also use these domain specific files for the lemma identifier to get the correct synonymous word. We believe that using domain specific entries will drastically increase the accuracy of the system. Currently, the system does not handle keyword extraction. It uses an external system to identify keywords of the system. Therefore, we could further extend this system to handle keyword extractions.

References

[1] Dhiraj Joshi, James Z. Wang and Jia Li "The Story Picturing Engine - A System for Automatic Text Illustration," ACM Transactions on Multimedia Computing, Communications and Applications (2006)

[2] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, "A Text-to-Picture Synthesis System for Augmenting Communication," Proceedings of the 22nd Conference on Artificial Intelligence: Integrated Intelligence Track (2007)

[3] Miller G.A. WordNet: An on-line lexical database. International Journal of Lexicography, 3(4): 235-312 (1990)

[4] Li, Xiaobin; Szpakowicz, Stan; and Matwin, Stan, "A WordNet-based algorithm for word sense disambiguation." Proceedings, 14th International Joint Conference on Artificial Intelligence, Montreal, (August 1995)

[5] A. Budanitsky and G. Hirst, Semantic Distance in WordNet: An Experimental, Application-Oriented Evaluation of Five Measures Proc. Workshop WordNet and Other Lexical Resources, Second Meeting North Am. Chapter Assoc. for Computational Linguistics, (June 2001)

[6] Resnik, P., Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of IJCAI-95, pages 448–453, Montreal, Canada

[7] Jiang, J., Conrath, D.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Proceedings of the 10th International Conference on Research on Computational Linguistics, Taiwan, (1997)

Using ASM-optical flow method and HMM in Facial Expression Recognition

Wencang Zhao [1, a], Junbo Zhang [1,b]

[1] College of Automation and Electronic Engineering, Qingdao University of Science & Technology, Qingdao 266042, China

[a]wencangzhao@gmail.com, [b]zhangjunboqust@163.com

Keywords: ASM, optical flow method, HMM, expression recognition.

Abstract. This paper presents a facial expression recognition method combining ASM-optical flow with hidden Markov model (HMM). First, we use active shape model (ASM) to obtain facial feature points, and track feature points by the means of optical flow method. And then count the variation of each feature point and extract the motion variations of feature sequences, such as eyebrows, eyes, nose, and mouth. Finally, the 6-dimensional expression feature sequences extracted above are used as the input of HMM to classify the facial expression. Experiments based on JAFFE expression database have obtained good results.

Introduction

FER (Facial expression recognition) is a very popular topic in the field of computer vision and pattern recognition. In recent years, scholars from all over the world have devoted themselves to solving with this problem, and a variety of effective methods have been proposed. The specific algorithms can be found in the reviewed preferences [1,2,3,4,5].

FER [6], which can separate the specific expressions from given sequences of images and videos, aims to make computer to understand human facial expression and then recognize. In general, FER can be divided into two primary parts: facial feature extraction and facial expression classification. According to the different nature of the image, facial expression extraction can be divided into static image feature extraction and image sequence feature extraction.

This paper presents a facial recognition algorithm of ASM-optical flow method and HMM based on the JAFFE database. By the comparison of feature parts motion between the neutral expression and other expressions, the expressions can be recognized and classified. First, we set the feature points on original images, and track these points to form optical flow field of the feature parts. Then, the 6 dimensional feature sequences are extracted, including the motion direction of eyebrows, eyes, nose, and mouth; the variation between eyebrows and eyes; the distance variation between eyes and mouth. According to the training of HMM, we can classify the inputted feature sequences. Compared to some other facial expression recognition methods, this paper has two advantages as follow:

1) Present a facial expression feature extraction method based on optical flow field of the motion direction of feature parts, only with 6-dimensional feature extracted, and the dimension is low.

2) Present a framework of facial expression recognition based on ASM-optical flow method and HMM. The method is strongly robust, and it can provide some ideas for the research of dynamic image sequence and subtle expression recognition.

Basics

ASM and Optical Flow Method. ASM is commonly used to obtain facial feature points, whose main feature is the application of statistical methods to model the target image. ASM relies on the training images, in which the feature points are pre-calibrated to form point distribution model (PDM-Point Distribution Model). And by adjusting the weight parameters of the model to match the model and the image extracted the image feature points, the image feature points extraction can be realized [7,8].

Optical flow method is the apparent movement caused by the brightness pattern, reflecting the actual situation of inter-frame motion, which is a very effective face recognition algorithm. When the object is in motion, the brightness pattern of its corresponding point is also moving. The apparent motion of image brightness pattern is the optical flow. Optical flow contains not only the movement of the object being observed, but also contains a wealth of information of 3-dimensional structure features. Optical flow movement is an important part in computer vision, and it is the application of the time-domain changes and correlation of grayscale data from moving image to determine the movement of the image pixels [9].

For an image, its brightness is expressed as $I(x, y, t)$, assuming that:

1) Image brightness values $I(x, y, t)$ in most regions are only related to their coordinates x and y;

2) Whether active or static, the brightness values on any point do not change with time.

For a moving object or a moving point, assuming that after dt time, its displacement is (dx, dy). Brightness $I(x, y, t)$ can be expanded with the Taylor series expansion as:

$$I(x+dx, y+dy, t+dt) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \ldots. \tag{1}$$

In the equation (1), the ellipsis means the higher-order partial derivatives of x, y and t. According to assumption 2), we know:

$$I(x+dx, y+dy, t+dt) = I(x, y, t). \tag{2}$$

From formulas (1) and (2), we can get:

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \ldots = 0. \tag{3}$$

Suppose the higher-order partial derivatives to be approximately zero and $\frac{dx}{dt} = u$ and $\frac{dy}{dt} = v$. Both sides of Equation (3) are divided by dt, and we get:

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v = -\frac{\partial I}{\partial t}. \tag{4}$$

Equation (4) is the optical flow constraint equation. The basic optical flow equation can determine an optical flow constrain field. On this line, any point (u, v) meets the basic formula and we can obtain the normal velocity from the following formula:

$$\mathrm{v} = \frac{-I_t(\nabla I)}{\|\nabla I\|_2^2}. \tag{5}$$

We can not obtain the unknown variable u and v only through optical flow, so some other constraints must be considered to determine the optical flow field. Researchers from different perspectives introduce different constraints, resulting in different optical flow analysis: Horn & Schunch technology [10], Lucas & Kanada technology and block matching technique. In this paper, H-S optical flow technology based on Horn & Schunch is used.

Hidden Markov Model (HMM). HMM is a double stochastic process and a probabilistic model indicated with the parameters. It consists of two parts: Markov chain and general random process. Markov chain described by the transition probability is used to describe the state transfer. General random process is used to describe the relationship between the state and the observed sequence,

which is described by the observing values probability. HMM mainly consists of the following components:

(1) State set A, $A=\{s_1,s_2,...,s_m\}$ and m is the number of states .

(2)A set of observations B, $B=\{u_1,u_2,...,u_n\}$ and n is the number of observation values.

(3)State transition probability distribution X, $X=\{x_{ij}\}$, $x_{ij}=P\{q_{t+1}=s_j \mid q_t=s_i\}, 1\le i,j\le n$, it shows the transition probability from time t state s_i to time $t+1$ state s_j .

(4) State j observation probability distribution Y, $Y=\{y_j(k)\}$ shows the corresponding observed value probability of state j output, $y_j(k)=P\{B_t=u_k \mid q_t=s_j\}, 1\le j\le m, 1\le k\le n$.

(5) Initial state probability distribution ϕ, $\phi=\{\phi_i\}$, $1\le i\le m$, $\phi_i=P\{q_1=s_i\}$.

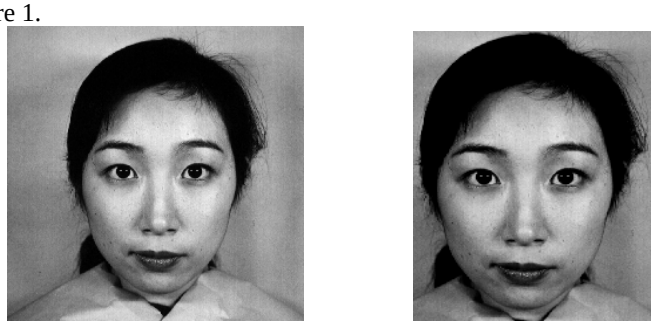
From the above, HMM can be defined as a quintuple $\lambda:\lambda=(A,B,\phi,X,Y)$.

ASM-Optical Flow Method and HMM for Facial Expression Recognition

Facial Expression Recognition Algorithms and Implementation based on ASM-Optical Flow Method and HMM. Variations of facial expression are reflected by the motion of featured facial parts such as eyebrow, nose, and mouth, so the expression can be recognized by extracting the movement direction of these parts. ASM is used to acquire the feature points. Optical flow method is used to track these feature points to get the optical flow motion field of feature points. The motion direction extracted from these feature points is used as the input of HMM. 6-dimensional feature sequences are extracted, including the motion direction of eyebrows, eyes, nose, and mouth; the variation between eyebrows and eyes; the distance variation between eyes and mouth. And then the HMM is used to facial expression sequences recognition and classification. The algorithm is as follows:

1) Pre-processing the FER

Clip the images from the facial expression database into size of $l\times h$ making the tip of the nose (x, y) as the base point; and pre-process the images through gray scale transformation, highlighting the characteristic parts of eyebrows, nose and mouth. The original image and the one after preprocessing are shown in Figure 1.

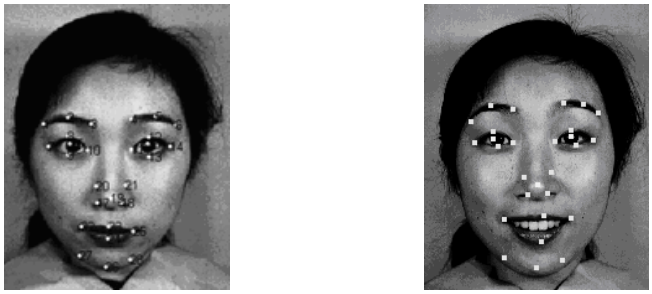


(a) Original image (b) The image after preprocessing

Fig. 1 The original image and the one after preprocessing

2) Acquiring feature points through ASM

Set 28 points in facial expression image as facial expression feature points, and they can be extracted by ASM. Feature extraction images are shown in Figure 2.



(a) Feature points extraction on neutral expression (b) Feature points extraction on happy expression

Fig. 2. feature point extraction results

3) Calculation of the optical flow field

H-S optical flow method is used to calculate optical flow field of the feature points obtained from the images which are processed above.

4) Motion direction extraction

The integral situation of upward and downward of the feature points motion direction are counted, including the pixels numbers and the distance variations, and the ratio between them can be used to estimate the feature part's motion direction. Then the 6-dimensional feature sequences are obtained, which include the motion direction of eyebrows, eyes, nose, and mouth; the variation between eyebrows and eyes; the distance variation between eyes and mouth.

5) Facial expression recognition through HMM

Finally, in this method, we select the states number as 6, namely m=6. The 6-dimensional feature points are inputted into the HMM after normalization.

Experimental Results and Analysis. The experiments are carried out on the computer with 1.6GHz dual-core CPU and 1G RAM memory. And JAFFE face database is used in the experiments. Vertical axis range between 110 ± 5 based on the nose is selected. The width and length of image are cut into $l = 230$, $h = 250$ respectively.

Actually, the high recognition rate of facial expression recognition is based on adequate training of a large number of training sequences. In the experiments the images from JAFFE are divided into 10 groups. Each group has 20 ± 3 images . 9 of the 10 groups are selected randomly every time, and the group left is used as the testing data. Facial expression recognition results are shown in Table 1.

Table 1 The results of facial expression recognition

Expression	Sad	Fear	Disgust	Surprise	Angry	Happy
Recognition rate(%)	82.38%	80.24%	85.79%	86.44%	85.21%	88.79%

As can be seen from Tab.1 that the recognition rates for two facial expressions including sadness and fear through this method is low, while the other facial expression recognition rates are high. So for the expressions of sadness and fear, two HMMs are used, and other types of facial expressions use only one HMM. Improved recognition results are shown in Table 2.

Table 2 The results of the improved facial expression recognition

Expression	Sad	Fear	Disgust	Surprise	Angry	Happy
Recognition rate(%)	84.68%	84.34%	85.79%	86.44%	85.21%	88.79%

Conclusion

This paper presents a facial expression recognition method with combination of ASM-optical flow method and HMM for facial expression recognition. The fusion of ASM and optical flow field is used to acquire motion direction of facial image feature points, which can be used as facial expression features. As the input of HMM, the dimension of 6-dimensional features is low, reducing the computer consumption and improving efficiency; under the condition of low feature dimension, the method achieves a high recognition rate of facial expression. Meantime, the method with strong robustness can provide subsequent research of dynamic expression recognition and subtle facial expression recognition with some useful and helpful ideas.

Acknowledgments

This work is supported by Project of the National Natural Science Foundation (61171131), Project of Natural Science Foundation of Shandong Province (Y2007G32), Project of the Applied Basic Research of Qingdao (09-1-3-52-JCH) and Outstanding Young Scientist Research Award Fund (the Doctor Fund) Project of Shandong Province (BS2009DX001).

References

[1] Liu Xiaomin, Tan Huachun, Zhang Yujin. New Research Advancesin Facial Expression Recognition. Journal of Image and Graphics,2006, 11 (10) : 1359- 1368.

[2] Fasel B, Luettin J. Automatic Facial Expression Analys is: A Survey. Pattern Recognition, 2003, 36 (1) : 259- 275.

[3] Zeng Zhihong, Pantic M, Roisman G I, et a l. A Survey of Affect Recognition Methods: Aud io, Visua,l and Spontaneous Expressions. IEEE Transon Pattern Analysis and Machine Intelligence, 2009, 31(1) : 39- 58.

[4] XUE Yuli, MAO Xia, GUO Ye, LV Shanwei. The Research Advance of Facial Expression Recognition in Human Computer Interaction[J]. Journal of Image and Graphics, 2009, 14(5):764-772.

[5] ZHANG Liwei, ZHANG Hang, ZHANG Yuying. A Survey of Facial Expression Recognition[J]. Techniques of Automation and Applications, 2009, 28(1):94-97.

[6] Feature Extraction Methods on Facial Expression Recognition, Journal of Chongqing Institute of Technology(Natural Science), 2008, 22(6): 118-121.

[7] T Cootes, C Taylor, D Cooper, et al. Active shape models-their training and application[J]. Computer Vision and Image Understanding: 1995, 61(1): 38- 59.

[8] ZHANG Sujun, JIANG Bin, WANG Ting. Facial Expression Recognition Algorithm Based on Active Shape Model and Gabor Wavelet[J]. Journal of Henan University (Natural Science), 2010, 40(5):521-524.

[9] CHENG Yuanhang. Face Feature Tracking Based on Optical Flow Algorithm[J]. Computerand Morden, 2010,7:120-122.

[10] Horn B K P, Shunck B G. Determining Optical Flow[J]. Artificial Intelligence, 1981, 17: 185-203.

Facial Illumination Compensation Based on the Wavelet Transform

Wencang Zhao[1, a],Chengcheng Zhao[1,b]

[1] College of Automation and Electronic Engineering, Qingdao University of Science & Technology, Qingdao 266042, China

[a]wencangzhao@gmail.com, [b]chengcheng19871111@163.com

Keywords: face recognition,wavelet transform,illumination compensation,Image denoising

Abstract. In order to improve the rate of face recognition under different lighting conditions，this paper presents a method of face illumination compensation based on *wavelet transform and edge denoising*.The first step is to do image gray processing, then the wavelet transform in logarithmic domain to get low frequency image,and carries on the light compensation,then high-frequency image de-noising filter,and last reverse the image processing to get the result.The results show the method can solve the problem of uneven illumination and improve face recognition rate effectively.

Introduction

In recent years,because of its natural and friendly properties,face recognition has become a hot spot of pattern recognition, image processing, computer vision, cognitive science and neural networks[1,2].However, the effect of face recognition is very vulnerable affected by posture, facial expression, light factors, especially the effect of light .In face recognition, the test is usually assumed in the same light conditions,But the truth is:the really same light condition is not exist,Different light conditions affect a lot on the face recognition.Therefore,how to process in different light conditions and obtain a good face recognition is a serious problem.

Effects of illumination changes on face recognition was first proposed by Adini and Jacobs[3,4] .In many cases, the same face in different light conditions changes more than the same face in different lighting conditions .The influence of the light on the face image mainly embodied in the intensity of illumination and illuminate Angle.Changes of light intensity will affect the gray of face image,the changes of angle will make facial image produced different light and shade area which has a big effect on the extraction of face information.At present,the methods used to solve the lighting problem can be roughly divided into three categories:first,do illumination compensation on images to reduce the impact of light,but this method is too direct, simple and very difficult to achieve the desired results.such as histogram equalization, the logarithmic transform[5], the exponential transform[6], etc;Second, extract the feature space which was not sensitive to light changes[7,8],This method mainly use the face features which are not sensitive to light as a target,such as the extract of edge image,Gabor transformation of face image[9],etc;The last method is to establish facial model method,This is mainly reflected in the 3D model[10], In different light conditions, there are different models in different light conditions,therefore, this method is quite complex.These three methods adjust the identified images to a unified light and in the light conditions complicated cases are not always effective.

Based on the above issues,this paper puts forward an algorithm of wavelet transform combined with the high-frequency filtering, it achieves effective compensation of illumination and gets a good effect of face recognition.

Image Pre-processing

Due to various constraints and interference,the original images often need to be pretreatment.This image in this paper is a 2-d gray image,so it can be the histogram modifications to enhance image contrast , reduce the interference of light on the image and improve the quality of the image.

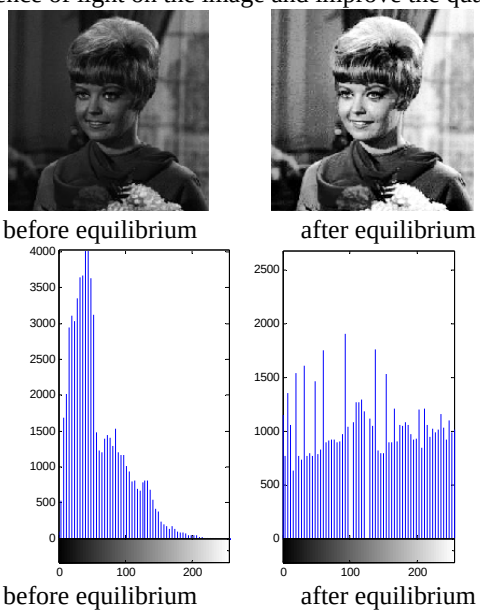


Fig. 1 image histogram equalization

As can be seen from Figure 1,after histogram equalization,the image contrast enhanced, all the details are more obvious. Before image histogram equalization,the ratio of low gray is very large.After adjustment, each part has a balanced proportion. However, histogram is just balance the image ,so the image looks too bright.

Light Compensation

Wavelet Transform. Wavelet transform is a conversion of time and frequency. Therefore, it can effectively extracted information from the image.By using wavelet decomposition and reconstruction algorithm it can achieve the extraction of partial information and dimension reduction functions,wavelet transform can do image frequency decomposition, Face images can be seen as the product of reflecting component and light weight.

$$f(x,y) = f_r(x,y) \times f_e(x,y) \tag{1}$$

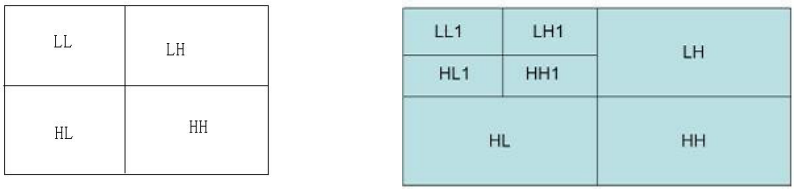
$f_e(x,y)$ changes with continuity,it belongs to low-frequency area in the image spectrum; $f_r(x,y)$ reflects the details of the image and the fluctuations is large,so it belongs to high-frequency area.This calculation is relatively too complex,in order to reduce computation,taking the logarithm on both sides:

$$\log[f(x,y)] = \log[f_r(x,y)] + \log[f_e(x,y)] \tag{2}$$

In this way,multiplication operation was converted into addition operation,greatly reduce the computational,and because of the monotony of the logarithmic function, the transformed relationship the same applies to the original function.

A two-dimensional image after wavelet transform can be decomposed into a low-frequency

component and three high frequency components,and the three high-frequency components respectively represent level detail component, vertical detail component and diagonal detail component.As the low-frequency component own the main image information,if want to do image level 2 wavelet decomposition, just need to decompose the low frequency components,after n times the wavelet decompose, there will be a low-frequency component and n level detail components, n vertical detail components and n diagonal detail components.



a order wavelet decomposition second order wavelet decomposition

Figure 2 wavelet decomposition process

Processing of High-frequency Information. A large part of the high frequency is noise,so the information should be denoising process,commonly used filters are: median filter,Gaussian low-pass filter,Laplace filter,Multi-dimensional filter, etc. Filtering effect of these four filters are:

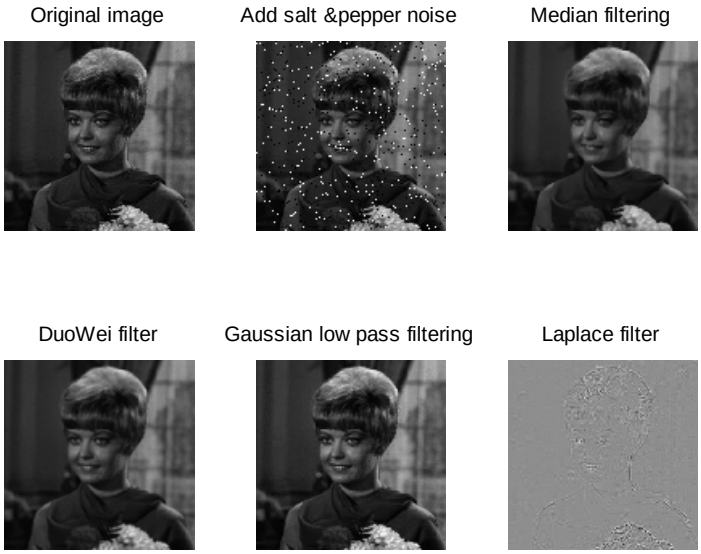


Fig. 3 kinds of filters filter results

As can be seen from figure 3,although the median filter maintains a good edge effect, it easy to lose some details;multi-dimensional filter has rather poor edge-preserving effect , the overall picture rather ambiguous;laplace filter saves little background information;only Gauss low-pass filter maintain the image details and have a good edge effect, it is the best denoising method.

Realize of Face Recognition. After Wavelet transform and image filtering,the process of light compensation has finished,at this time the image shoule be anti-transform to emerge the image visually,The flow chart of face recognition is:

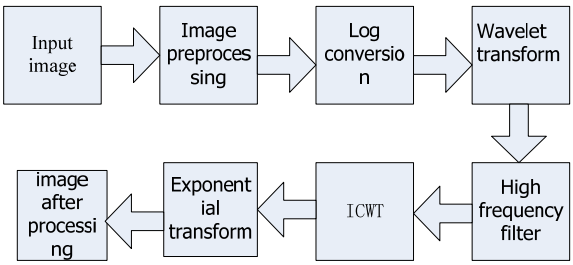


Fig. 4 face recognition flow chart

Experimental Results and Analysis. In order to obtain better results,this paper uses Yale's[11] face database with experiments,since this paper studies the impact of illumination change on the face recognition,so only use the face images which are positive, natural expression, I selected 150 image as the object from the library.

To test the validity of the method of this pape,it respectively compared with histogram modification and discrete cosine,The direction of light are 45 degrees left of the light,45 degrees rightof the light,ordinary light,positive low light.The situation of face recognition is:

a) images before processing

b) images after the histogram equalization

c) images after discrete cosine transform

d) images after The algorithm of this article

Fig. 5 kinds of face recognition results

Tab. 1 the error rate of every method under different illumination

	ordinary light	45 degrees left of the light	45 degrees rightof the light	positive low light
image before processing	0.02	0.45	0.52	0.21
image after the histogram equalization	0.03	0.36	0.31	0.12
image after discrete cosine transform	0.01	0.16	0.19	0.1
image after The algorithm of this article	0.01	0.18	0.12	0.05

As can be seen from Table 1,histogram has a better recognition in low light.In a polarizing,discrete cosine also achieved good recognition results,the proposed method in this article has a slightly lower recognition rate than discrete cosine in the light of 45 degrees left, the rest of the results are better than other algorithms.

Conclusion

Illumination changes has been a problem for face recognition,although the wavelet transform can filtere image,but has a low recognition rate and robustness is poor,the proposed method in this article pretreat the images first,then the wavelet transform combined with the high-frequency filtering methods to effectively solve the problem of illumination changes and get a good recognition rate in the Yale B face database .

References

[1] Che Happa R，Wilson CL，Sirohey S. Human and l' n a.chine recognition of faces：a survey [J].Proceedings of the IEEE,1995，83(5)：705—740.

[2] Bowyer K W，Chang K，Flynn PJ. A survey of approaches and challenges in 3D and multimodal 3D+2D face recognition[J]. Computer Vision and Image Understanding，2006，101 (1)：1—15.

[3] Moses Y,Ullman S. Limitation of Non-model-based Recognition Schemes[C]. Proceeding of ECCV-92,Berlin:Springer-Verlag,1992:820- 828.

[4] Adini Y,Moses Y，Ullman S. Face Recognition:The Problem of Compensating for Changes in illumination Direction[J] . IEEE Transactions on Pattern Analysis and Machine Intelligence ,1997,19 (7) ： 721 - 732.

[5] Shan S,Gao W,Cao B,eta1..Illumination normalization for robust face recognition against varying lighting conditions[J]. Proc IEEE Workshop on AMFG,2003：157—164.

[6] Savvides M,Kumar V. Illumination normalization using logarithm transforms for face authentication[C].Proc IAPR AVBPA,2003:549—556.

[7] P.N.Belhumeur，J.P.Hespanha and D.J.Kriegman. Eigenfaces vs. Fisherfaces：recognition using class specific linear projection[J].IEEE Trans on Pattern Analysis and Machine Intelligence,1997,l 9 (7) :717—720.

[8] J Yang,J Y Yang,From image vector to matrix:a straightforward image progression technique-MPCA vs.PCA[J].Pattern Recognition. 2002, 35(9):1997-1999.

[9] LIU D H，LAM K M，S HEN L S. Illumination invariant face recognition[J]. Pattern Recognition, 2005,38：1705 — 1706.

[10] BIANZ V,VEITTER T.Face recognition based on fitting a 3D morphable model [J] .IEEE Trans.on PAMI , 2003,25(9)：1063—1074.

[11] A.S. Georghiades, P.N. Belhumeur, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 643– 660.

Robust Cognitive System Engineering Based on Control Frame of Cognition

Rui Wang[1] and Keiji Watanabe[2]

[1] Northeastern University, Shenyang, China

[2] Higashi 2-7-123, Yonezawa, Japan

wangruisys@hotmail.com

Keywords: cognitive system engineering (CSE); cognition; model-based control; robust cognitive system; fuzzy neural network (FNN)

Abstract. In this paper, a robust cognitive system framework that derives from control frame of cognition is presented. Through reviewing structure of control system, cognitive system can be illustrated in some forms of model-based control, which is a crucial issue of the system engineering. For robust cognitive system, fuzzy neural network is introduced into the system for robustness and adaptive performance. The structural modeling of cognitive system is also proposed to give an analytical method for system analyses and syntheses. Since the principle concerns the knowledge of cognitive science and control engineering, it shows advantages of inter-discipline in system theory and engineering.

Introduction

Modern day a relatively new sub-filed of science gained a significant attention with the advancement in technology which known as cognitive system engineering (CSE). The focus of CSE is how humans can cope with and master the complexity of processes and technological environments, and the main thrust of CSE is how control can be improved. Consider cognition is the scientific term for the “process of thought” to knowing, it usually refers to an information processing view of an individual's psychological functions. Further, adaptive performance of the whole process is an important feature which implies the robustness of it.

A frame of control system that named as perceptual control presented by W.T.Powers only shows a simple structure and lack of detail description. It is known that for process control, the “controller” usually is considered to “understand” the “external environment”. After the perceptual control theory was presented, cognitive science had a significant development, which provides us some new way to study the “model” in the “controller”, which means it is necessary to do some more analysis on the control frame of system. Further, many control engineering studies of bio-system control system are seemed lack in the knowledge of cognitive science obviously since they could not give some details that one observed and recorded a clear and reasonable interpretation, which just made the presented studies difficult to integrate the control inter-discipline and cognitive science in high level so that difficult to provide more developments of cognitive system engineering and control theory.

By reviewing the frameworks of cognitive system engineering, this paper presents a frame of robust cognitive system which can be used to improve the robustness and adaptive performance.

Cognition and Process Control

Cognitive science is the interdisciplinary study of how information is represented and transformed in the brain, which consists of multiple research disciplines, including psychology, philosophy, neuroscience, linguistics, anthropology, sociology, etc., and it spans many levels of analysis, from low-level learning and decision mechanisms to high-level logic and planning; from neural circuitry to modular brain organization. Since we focus on the control algorithm of cognition, in this paper, we analyze the robust cognitive system at control level.

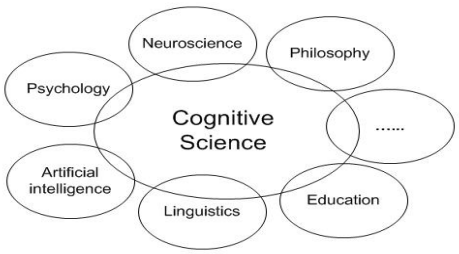


Fig.1 Interdisciplinary ─ cognitive science

Among the items concerned with cognitive science listed in Fig.1, psychology and neuroscience are the most important and functional items. The two issues are the basic terms for "the process of thought", which implies the analytic relationship with brain and mind.

Brains, the contents of our skulls, are composed of extraordinarily intricate, self-organizing, physical structures, performing many tasks in parallel at many scales, from individual molecules to large collections of cooperating neurons or chemical transport system.

Minds are more abstract and contain ideas, perceptions, thoughts, feelings, memories, mathematical knowledge, motives, moods, emotions, reasoning processes, decisions, motor control skills and other things that cannot be seen by opening up skulls. Yet their existence and their power to do things depend on all the "wetware" components that make up brains.

Consider that psychology and neuroscience are the science of the mind and brain respectively, the transform of which is shown in Fig.2.

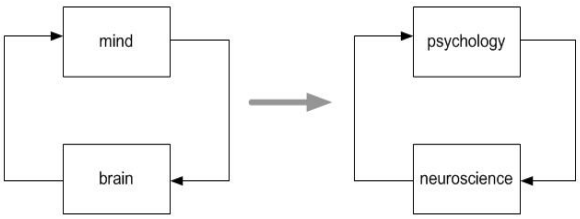


Fig.2 From "brain &mind" to "psychology &neuroscience"

The perceptual control system is an open system, which is a system that regularly exchanges feedback with its external environment. Cognitive system also can be open system and its evaluation of cognition derives from Fig.2. Figure 3 gives the frame of cognition in open system.

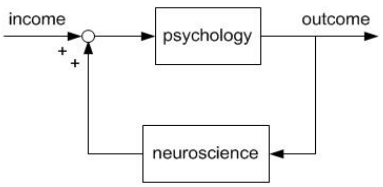


Fig.3 Cognition in open system

Cognition has to work on better models that explain natural processes and that are reliably able to make predictions. Neuroscience and psychology are able to analyze the relation between the physiology of the brain and mental processes. According to the pragmatic definition, any system that performs in an orderly manner by showing evidence of goal directed or controlled actions is said to be cognitive. This means that humans are cognitive, but also that machines or technological artefacts can be cognitive. If we combine the control theory and cognitive science, the algorithm of human control system can be given in Fig.4.

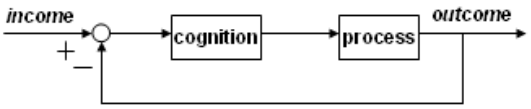


Fig.4 Cognition control system

Control Structure of Cognitive System

The concept of control is essential to cognitive system engineering. According to control theory, a "good controller" must involve a process model or some information about the process. For this, internal model control (IMC) algorithm also can be used in "human control system" to make it more scientific and systematic. Further, in terms of control modes this can now be described as the ability to maintain a control mode despite disturbing influences. For a given domain it is possible to develop more precise criteria for control mode transitions, and use these as the basis for a model.

Cognition in Control System

Cognition that involves "mind" and "brain" mainly formulates the "control process" shown in Fig.5. In the control system description, information of outcome is sensed, and changes are effected in response to the information. In human process, "sensor" can be perception; feeling; etc.

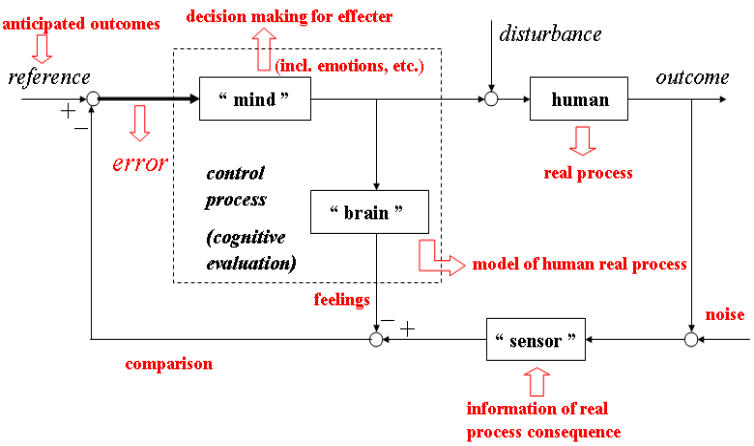


Fig.5 Cognition in control system

Cognitive System based on IMC

Consider cognitive science seeks to unify neuroscience and psychology with other fields that concern themselves with the brain and mind, a basic framework of the system can be given in Fig.6, in which

psychology is a “free parameter” which tunes the human brain/mind system and can be used to give some interpretation of the output performance of “cognition control system”, so that which can be regarded as “controller”. On the other hand, “neuroscience” yields a model of human brain process in the controller. Based on the idea, control frame of CSE is given by Fig.6 derives from internal model control (IMC).

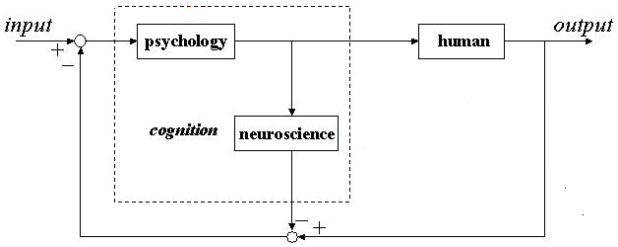


Fig.6 Cognitive system based on IMC

Robust Cognitive System based on Model Bridge Control

In common, it is impossible to modeling exactly and neuroscience is not always able to explain all the observations made in laboratories, for this, the system towards cognitive psychology in order to find models of brain and behaviour on an interdisciplinary level – Cognitive Neuropsychology. This “inter-science” as a bridge connects and integrates the two most important domains and their methods of research of the human mind, which just like a “compensator model” for robustness of brain/mind. The frame shown in Fig.7 derives from model bridge control (MBC). Research at one level provides constraints, correlations and inspirations for research at another level.

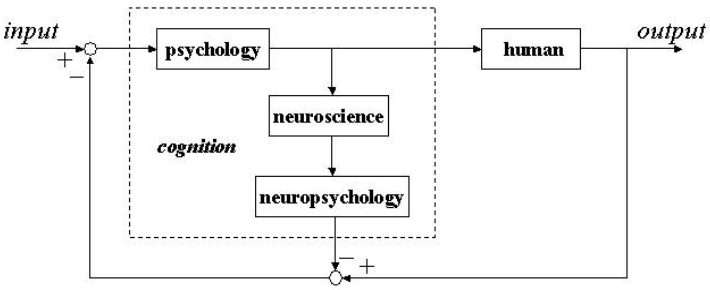


Fig.7 Robust cognitive system based on MBC

Structural Modeling of Robust Cognitive System Engineering

Modeling of cognitive system engineering in this study is based on general system theory (GST). GST was originally proposed by Hungarian biologist Ludwig von Bertalanffy in 1928, He proposed that ‘a system is characterized by the interactions of its components and the nonlinearity of those interactions.’ Cognitive system can be modeled as a ‘general system’. And CSE is developed based on general system theory.

Since we have known the fact about the brain that different parts of the brain perform different jobs, system can be considered into each element or section and relations of them, i.e., many topics in cognition— remembering, attention, making decisions, emotions, etc. In general system theory, a system variable is any element in an acting system that can take on at least two different states, which can be continuous and the condition of a variable in a system is known as the system state.

CSE is interested in developing better ways of analyzing and control performance of cognitive systems. Modeling of cognition, has willingly accepted as its purpose to account for what goes on in the human mind, specifically as the information processes that go on in the mind between input and output. The structural modeling provides an analytical method for study on performance of CSE.

Consider fuzzy neural network combining both the advantage of neural network and fuzzy control, which can be introduced into the design of CSE as shown in Fig.8. In this sense, robust cognitive system can be modeled as follows:

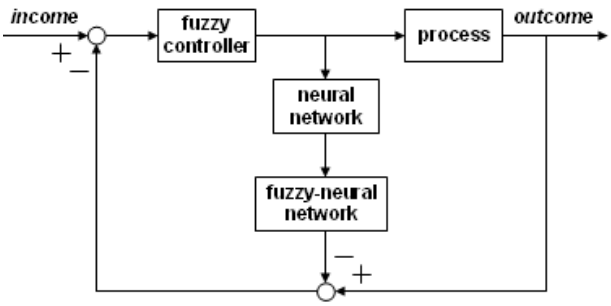


Fig.8 Control framework of robust cognitive system

System Structural Description

X and E denote the set of all elements and relation set of them in the cognitive system respectively which given in

$$X=\{S_1, S_2, \ldots\ldots, S_n\} \tag{1}$$

$$E=\{e_{ij}\} \ ; \ j=1,2, \ldots\ldots, n \ ; \ e_{ij}=(S_i,S_j) \tag{2}$$

where S_i denotes elements vector and e_{ij} represents the relation set between vectors S_i with S_j.

Then, the system S can be defined as

$$S \triangleq (X,E) \tag{3}$$

$$M = I \cup A^1 \cup A^2 \cup A^3 \cup \cdots \cup A^n \tag{4}$$

where A is a matrix of X, M represents the reachability matrix of system, which yields structural modeling.

In addition, the structural model matrix and its transformation make it possible to analyze the system integrating cognition and control theory.

System State Description

If we refer to the basic principles of the feedback loop model and assume that R and S denote income and system state (construct) respectively, Y is the action used to compensate the error in cognitive system. The system can be described as:

$$S(k) \leftarrow R(k) \mid S(k-1) \tag{5}$$

which means the state of system at time k is determined by the income/feedback given the system at time k-1. This relation provides a link from the past to the present and represents the reactive aspects of the model.

$$Y(k) \leftarrow S(k) \mid R(k+1) \tag{6}$$

which means the action at time k is determined by the current system state (construct) given the expected outcome (feedback) of the action. This relation provides a link from the present to the future and represents the proactive aspects of the model.

Summary

In summary, this paper has reviewed some foundations of control frame in CSE. Structural modeling has been given by combining the knowledge of cognitive science and theory of control system, which can be used to make cognitive science more systematic and scientific. By introducing fuzzy neural network into the system design, the robustness and adaptive performance of cognitive system can be improved. Based on the system theory and control frame of CSE, it is possible to refine the control system design method for more intelligence.

References

[1] Powers, W. T., *Behaviour: The control of perception*. Chicago: Aldine Transaction, 1973.

[2] R.Wang, K. Watanabe, E. Muramatsu, Y. Ariga. “Phase Compensated Robust Control via IMP”. SICE Annual Conf. 2005, Okayama, Japan, pp.3007-3011.

[3] K. Watanabe, R.Wang, E. Muramatsu, Y. Ariga. “Model Bridge Control ─ Multi-Degree-of-Freedom Design for High Robustness and High Performances”. IFAC08, Seoul, Korea, pp.6166-6171

[4] Erik Hollnagel, and David D. Woods, *Joint Cognitive Systems Engineering*, CRC Press Taylor & Francis Group, 2005.

[5] http://en.wikipedia.org/wiki/Mind#Brain

[6] http://en.wikipedia.org/wiki/Cognitive_science

[7] http://en.wikipedia.org/wiki/Cognition

[8]David S. Walonick, “General System Theory”, http:// www.survey-software-solutions.com/walonick/systems-theory.htm, 1993.

[9] Daniel Reisberg, *cognition: exporing the science of the mind*, W.W. Norton&Company, Inc., 2005.

CT Liver Segmentation Based on Fuzzy Cellular Neural Networks and Its Stability

P. Balasubramaniam[1, a], M. Kalpana [*2, b]

[1]Institute of Mathematical Sciences, University of Malaya, 50603, Kuala Lumpur, Malasiya

[2]Department of Mathematics, Gandhigram Rural Institute - Deemed University,

Gandhigram - 624 302, Tamilnadu, India

[a]balugru@gmail.com, [b]kalpana.nitt@gmail.com

Keywords: Image segmentation; Simple thresholding; Region of interest; Fuzzy cellular neural network; CT liver image

Abstract: Computed tomography or Computer tomography (CT) liver images acquired by CT imaging mechanisms often contain uncertainties due to electrical/hot noise, the diversity of the human organ and the partial volume effect. Such uncertainties make boundary segmentations very difficult among different organs appearing in liver images. The aim of this paper is to determine a well defined CT liver segmentation. In this paper, a simple thresholding (ST) and region of interest (ROI) methods are used with fuzzy cellular neural network (FCNN) to obtain a well defined CT liver segmentation. Finally, the experimental and comparison results are provided to illustrate the effectiveness of our proposed simple technique.

1. Introduction

Chua and Yang first introduced cellular neural networks (CNNs) in 1988 [1, 2]. A set of CNNs in parallel local connectivity are widely accepted in both practical applications and in biologies to achieve higher-level information processing and reasoning functions. This integrated CNN system helps to solve more complex intelligence problems [3, 4]. However, in every phase of image processing, there exist many uncertainties. Fuzzy set theory provides the mathematical strength to capture these uncertainties. FCNN was introduced by Yang et al. [5, 6] and has been used in [7, 8] as an application tool in image processing and pattern recognition.

Current medical practice in different diagnosis of diseases and therapy require standard medical tests that could be performed accurately, efficiently and rapidly. Thus, the need for accurate, efficient and economical performance of this test makes it a perfect target for research in automation.

Liver diseases, including liver cancer, are among the leading causes of death in developing countries. The most useful approach for reducing deaths due to liver diseases is to treat these diseases in the early stages. Early treatment requires early diagnosis, which requires an accurate and reliable diagnostic procedure. Medical image processing plays a pivotal role in early diagnosis. Some of the research works on liver images can be seen in [9]-[12].

Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images. The ROI is a selected subset of samples within a data set identified for a particular purpose. The concept of ROI is commonly used in medical

[*] The author was supported by No. DST/INSPIRE Fellowship/2010/[293]/dt. 18/03/2011.

imaging. Motivated by the above discussion, in this paper a ST and ROI methods are used with FCNN to get a well defined CT liver segmentation. Comparison of advanced fuzzy cellular neural network (AFCNN) in [9] and FCNN for CT liver segmentation is also provided to show the effectiveness of the proposed technique.

This paper is organized as follows. Section 2 deals with the model description of FCNN. Section 3 deals with the global stability. The experimental studies and comparison are discussed in Section 4. Section 5 concludes the paper.

2. Model Description of FCNN

The locally connected network consists of $M \times N$ neurons. The output of a neuron is connected to all the inputs of every neuron in its $r \times r$ neighborhood, and similarly all the inputs of a neuron are only connected to the outputs of each neuron in its $r \times r$ neighborhood. Each neuron in this $M \times N$ FCNN can be described in the following way.

The state equation of a cell C_{ij} is given by

$$\begin{aligned} C\frac{dx_{ij}}{dt} = &-\frac{1}{R_x}x_{ij} + \sum_{C_{kl}\in N_r(i,j)} A(i,j;k,l)\, y_{kl} + \sum_{C_{kl}\in N_r(i,j)} B(i,j;k,l)\, u_{kl} + I_{ij} \\ &+\widetilde{\Lambda}_{C_{kl}\in N_r(i,j)}\big(A_{fmin}(i,j;k,l) + y_{kl}\big) + \widetilde{\mathrm{V}}_{C_{kl}\in N_r(i,j)}\big(A_{fmax}(i,j;k,l) + y_{kl}\big) \\ &+\widetilde{\Lambda}_{C_{kl}\in N_r(i,j)}\big(B_{fmin}(i,j;k,l) + u_{kl}\big) + \widetilde{\mathrm{V}}_{C_{kl}\in N_r(i,j)}\big(B_{fmax}(i,j;k,l) + u_{kl}\big). \end{aligned} \tag{1}$$

Input equation of C_{ij} :

$$u_{ij} = E_{ij},\ \ 1 \le i \le M,\ \ 1 \le j \le N. \tag{2}$$

Output equation of C_{ij} :

$$y_{ij} = f(x_{ij}) = \tfrac{1}{2}\big(|x_{ij}+1| - |x_{ij}-1|\big). \tag{3}$$

Constraint conditions:

$$\begin{cases} A(i,j;k,l) = A(k,l;i,j),\ A_{fmin}(i,j;k,l) = A_{fmin}(k,l;i,j), \\ A_{fmax}(i,j;k,l) = A_{fmax}(k,l;i,j),\ \ 1 \le i \le M,\ 1 \le j \le N, \end{cases} \tag{4}$$

where x_{ij}, u_{ij} and y_{ij} denote respectively the state, input, and output variables of cell C_{ij}. $A(i,j;k,l)$ and $B(i,j;k,l)$ denote the feedback and feedforward synaptic weights between cells C_{ij} and C_{kl}, respectively. The bias is I_{ij} (also called threshold) of cell C_{ij}, which may be static, time-varying, space-invariant, or space-varying; $A_{fmin}(i,j;k,l)$ and $A_{fmax}(i,j;k,l)$ are elements of the fuzzy feedback MIN and MAX templates, respectively; $B_{fmin}(i,j;k,l)$ and $B_{fmax}(i,j;k,l)$ are elements of the fuzzy feedforward MIN and MAX templates, respectively; $C > 0$ and $R_x > 0$ are the values of the capacitor and the resistor, respectively.

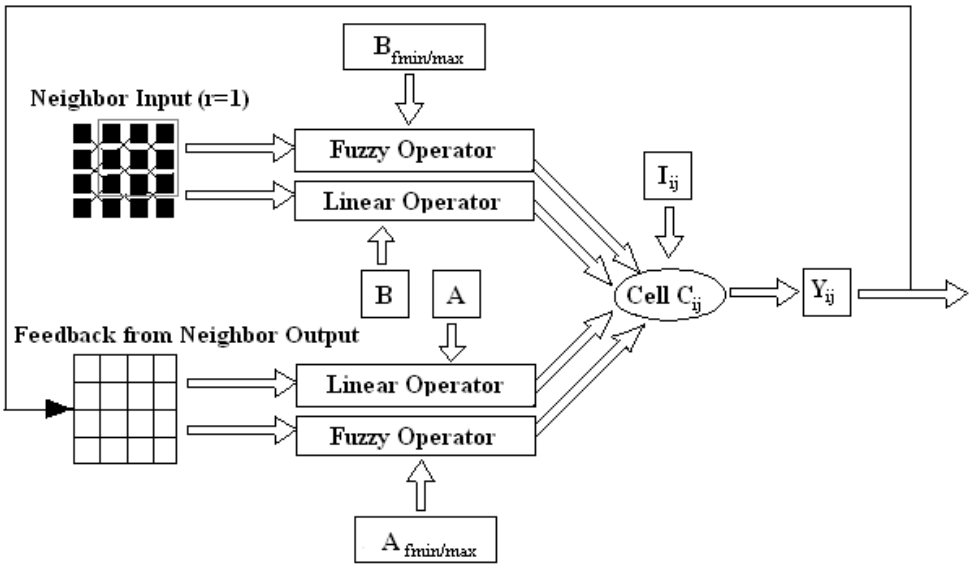


Fig. 1 Architecture of FCNN.

Definition 2.1 ([1]). *r-neighborhood*

The *r-neighborhood* of a cell $C(i,j)$, in a CNN is defined by

$$N_r(i,j) = \{C(k,l) \mid max\,\{|k-i|,|l-j|\} \le r,\ \ 1 \le k \le M;\ \ 1 \le l \le N\},$$

where r is a positive integer number.

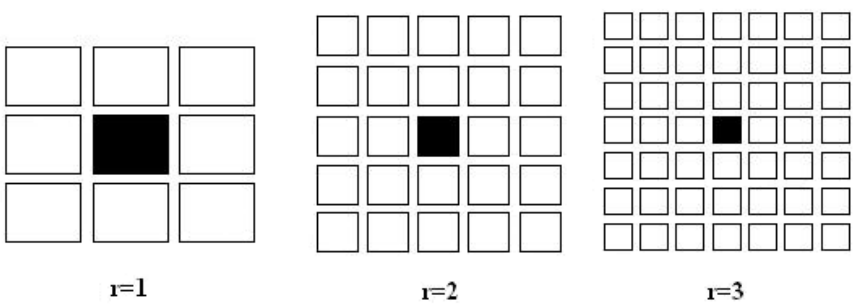


Fig. 2 The neighborhood of cell C_{ij} defined by Eq. 1 for $r = 1, r = 2$ and $r = 3$, respectively.

3. Global stability

Before reviewing our experimental studies, we will prove the global stability of the FCNN in this section.

Definition 3.1 The fuzzy feedback MIN/MAX templates and fuzzy feedforward MIN/MAX templates $A_{fmin}(i,j;k,l)$, $A_{fmax}(i,j;k,l)$, $B_{fmin}(i,j;k,l)$, and $B_{fmax}(i,j;k,l)$ take the constant values α,β,γ and δ respectively; if for all $C_{kl} \in N_r(i,j)$ and there exists a connection between neurons C_{ij} and C_{kl} .

Thus, in terms of the above definition, Eq. 1 can be reformulated as

$$C\frac{dx_i}{dt} = -\frac{1}{R_x}x_i + \sum_{j=1}^{MN} a_{ij} f_j(x_j) + \sum_{j=1}^{MN} b_{ij} u_j + I_i + \widetilde{\bigwedge}_{j=1}^{MN}\alpha_{ij} f_j(x_j) + \widetilde{\bigvee}_{j=1}^{MN}\beta_{ij} f_j(x_j)$$
$$+\widetilde{\bigwedge}_{j=1}^{MN}\gamma_{ij} u_j + \widetilde{\bigvee}_{j=1}^{MN}\delta_{ij} u_j, \quad i = 1,2,\dots,MN, \tag{5}$$

where $\alpha_{ij} = \begin{cases} \alpha, & if\ there\ is\ a\ connection \\ undefined, & otherwise, \end{cases}$ $\beta_{ij} = \begin{cases} \beta, & if\ there\ is\ a\ connection \\ undefined, & otherwise, \end{cases}$

$\gamma_{ij} = \begin{cases} \gamma, & if\ there\ is\ a\ connection \\ undefined, & otherwise, \end{cases}$ $\delta_{ij} = \begin{cases} \delta, & if\ there\ is\ a\ connection \\ undefined, & otherwise. \end{cases}$

According to (4), we have immediately set $a_{ij} = a_{ji}$, $\alpha_{ij} = \alpha_{ji}$, and $\beta_{ij} = \beta_{ji}$.

Lemma 3.1 ([13]). Let x, x' be two states of Eq. 5, then

$$\left|\widetilde{\bigwedge}_{j=1}^{MN}\alpha_{ij} f_j(x) - \widetilde{\bigwedge}_{j=1}^{MN}\alpha_{ij} f_j(x')\right| \le \sum_{j=1}^{MN}\left|\alpha_{ij}\right|\left|f_j(x) - f_j(x')\right|,$$
$$\left|\widetilde{\bigvee}_{j=1}^{MN}\beta_{ij} f_j(x) - \widetilde{\bigvee}_{j=1}^{MN}\beta_{ij} f_j(x')\right| \le \sum_{j=1}^{MN}\left|\beta_{ij}\right|\left|f_j(x) - f_j(x')\right|.$$

(A_1) The neuron activation function $f_j(\cdot)$ are Lipschitz continuous; that is, there exist constants $l_j > 0$ such that

$$\left|f_j(\xi_1) - f_j(\xi_2)\right| \le l_j|\xi_1 - \xi_2|, \quad \text{for all } \xi_1,\ \xi_2 \in \mathbb{R},\ \xi_1 \ne \xi_2.$$

Let the $MN \times MN$ matrix be $|A| \triangleq \left(\left|a_{ij}\right| + \left|\alpha_{ij}\right| + \left|\beta_{ij}\right|\right)_{MN\times MN}$, then we have Lemma 3.2.

Lemma 3.2 ([9]). If the spectral radius $\rho(R_x|A|)$ of the matrix $R_x|A|$ is less than 1, then there exists only one globally stable equilibrium point in Eq. 5.

Let x^* be an equilibrium state of Eq. 5. In terms of Eq. 5, we have

$$C\frac{d(x_i - x_i^*)}{dt} = -\frac{1}{R_x}(x_i - x_i^*) + \sum_{j=1}^{MN} a_{ij}\left(f_j(x_j) - f_j(x_j^*)\right) + \left[\widetilde{\bigwedge}_{j=1}^{MN}\alpha_{ij} f_j(x_j)\right.$$
$$\left. -\widetilde{\bigwedge}_{j=1}^{MN}\alpha_{ij} f_j(x_j^*)\right] + \left[\widetilde{\bigvee}_{j=1}^{MN}\beta_{ij} f_j(x_j) - \widetilde{\bigvee}_{j=1}^{MN}\beta_{ij} f_j(x_j^*)\right], \ i = 1,2,\dots,MN. \tag{6}$$

Theorem 3.1 Suppose the spectral radius $\rho(R_x|A|)$ of the matrix $R_x|A|$ is less than 1, under assumption (A_1), and

$$\sum_{j=1}^{MN}(|a_{ij}| + |\alpha_{ij}| + |\beta_{ij}|)l_j + \sum_{j=1}^{MN}(|a_{ji}| + |\alpha_{ji}| + |\beta_{ji}|)l_i < \frac{2}{R_x}, \quad i = 1,2,\dots,MN.$$

Then, the equilibrium state x^* in Eq. 6 is globally stable.

Proof Since $\rho(R_x|A|) < 1$, there exists only one equilibrium state in Eq. 6. Let us take the following Lyapunov function

$$V(x) = \sum_{i=1}^{MN} C(x_i - x_i^*)^2.$$

Obviously, $V(x^*) = 0$, and $V(x) \to +\infty$ ($|x_i - x_i^*| \to +\infty$). Let us observe

$$\begin{aligned}
\frac{dV(x_i)}{dt} &= \sum_{i=1}^{MN} 2C(x_i - x_i^*)\frac{d(x_i - x_i^*)}{dt} \\
&= \sum_{i=1}^{MN} 2(x_i - x_i^*)\left[-\frac{1}{R_x}(x_i - x_i^*) + \sum_{j=1}^{MN} a_{ij}\left(f_j(x_j) - f_j(x_j^*)\right) + \widetilde{\wedge}_{j=1}^{MN}\alpha_{ij}f_j(x_j)\right. \\
&\quad \left. -\widetilde{\wedge}_{j=1}^{MN}\alpha_{ij}f_j(x_j^*) + \widetilde{\vee}_{j=1}^{MN}\beta_{ij}f_j(x_j) - \widetilde{\vee}_{j=1}^{MN}\beta_{ij}f_j(x_j^*)\right] \\
&\le \sum_{i=1}^{MN}\left(\frac{-2}{R_x}\right)(x_i - x_i^*)^2 + \sum_{i=1}^{MN} 2|x_i - x_i^*|\left[\sum_{j=1}^{MN}|a_{ij}||f_j(x_j) - f_j(x_j^*)|\right. \\
&\quad \left. + \sum_{j=1}^{MN}|\alpha_{ij}|\,|f_j(x_j) - f_j(x_j^*)| + \sum_{j=1}^{MN}|\beta_{ij}|\,|f_j(x_j) - f_j(x_j^*)|\right] \\
&\le \sum_{i=1}^{MN}\left(\frac{-2}{R_x}\right)(x_i - x_i^*)^2 + \sum_{i=1}^{MN} 2|x_i - x_i^*|\left[\sum_{j=1}^{MN}|a_{ij}|\,l_j\,|x_j - x_j^*| + \sum_{j=1}^{MN}|\alpha_{ij}|\,l_j\,|x_j - x_j^*|\right. \\
&\quad \left. + \sum_{j=1}^{MN}|\beta_{ij}|\,l_j\,|x_j - x_j^*|\right] \\
&\le \sum_{i=1}^{MN}\left(\frac{-2}{R_x}\right)(x_i - x_i^*)^2 + \sum_{i=1}^{MN} 2|x_i - x_i^*|\left[\sum_{j=1}^{MN}(|a_{ij}| + |\alpha_{ij}| + |\beta_{ij}|)\,l_j\,|x_j - x_j^*|\right] \\
&\le \sum_{i=1}^{MN}\left(\frac{-2}{R_x}\right)(x_i - x_i^*)^2 + \sum_{i=1}^{MN}\left\{\sum_{j=1}^{MN}(|a_{ij}| + |\alpha_{ij}| + |\beta_{ij}|)\,l_j\,\left[(x_i - x_i^*)^2 + (x_j - x_j^*)^2\right]\right\} \\
&\le \sum_{i=1}^{MN}\left(\frac{-2}{R_x}\right)(x_i - x_i^*)^2 + \sum_{i=1}^{MN}\left[\sum_{j=1}^{MN}(|a_{ij}| + |\alpha_{ij}| + |\beta_{ij}|)\right]l_j\,(x_i - x_i^*)^2 \\
&\quad + \sum_{i=1}^{MN}\left[\sum_{j=1}^{MN}(|a_{ji}| + |\alpha_{ji}| + |\beta_{ji}|)\right]l_i\,(x_i - x_i^*)^2 \\
&= \sum_{i=1}^{MN}\left\{\frac{-2}{R_x} + \sum_{j=1}^{MN}(|a_{ij}| + |\alpha_{ij}| + |\beta_{ij}|)\,l_j + \sum_{j=1}^{MN}(|a_{ji}| + |\alpha_{ji}| + |\beta_{ji}|)\,l_i\right\}(x_i - x_i^*)^2 < 0.
\end{aligned}$$

Then, the equilibrium state x^* in Eq. 6 is globally stable. Its global stability shows that the state of a neuron will finally find a stable equilibrium point.

When applying FCNN to an $M \times N$ liver image, we should consider $M \times N$ neurons in FCNN, in which each neuron corresponds to a pixel. Each neuron will change its state iteratively according to Eq. 1, until the entire FCNN network converges. In other words, we transform a liver image into a dynamic system (that is FCNN), and its state equations will continuously change to approach the minimum energy until the final convergence is achieved. We will demonstrate the power of FCNN with ST and ROI methods by our experimental results on the segmentation of CT liver image.

4. Experimental studies

In this section, we will present few experimental results to demonstrate the effectiveness of our FCNN based approach.

4.1 CT liver segmentation by using FCNN

The original templates for CT liver image of Fig. 3 are

$$A = \begin{bmatrix} 0.2 & 0.1 & 0.2 \\ 0.2 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.2 \end{bmatrix}, \quad B = \begin{bmatrix} 0.1 & 0.3 & 0.1 \\ 0.1 & 0.2 & 0.1 \\ 0.1 & 0.4 & 0.1 \end{bmatrix}, \quad A_{fmin} = \begin{bmatrix} 0 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0 \\ 0.1 & 0.2 & 0.2 \end{bmatrix},$$

$$B_{fmin} = \begin{bmatrix} 0.3 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.3 \\ 0.5 & 0.1 & 0.2 \end{bmatrix}, \quad A_{fmax} = B_{fmax} = \begin{bmatrix} 0 & 0.1 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0.1 & 0 \end{bmatrix}.$$

Letting $R_x = 1,\ C = 1,\ I = 0,\ x_0 = undefined,$ and $f_j(x_j) = \frac{1}{2}(|x_j + 1| - |x_j - 1|), \quad j = 1,2, \dots, MN,$ which satisfies assumption (A_1), we get $l_j = 1$. By using Matlab, the results are shown

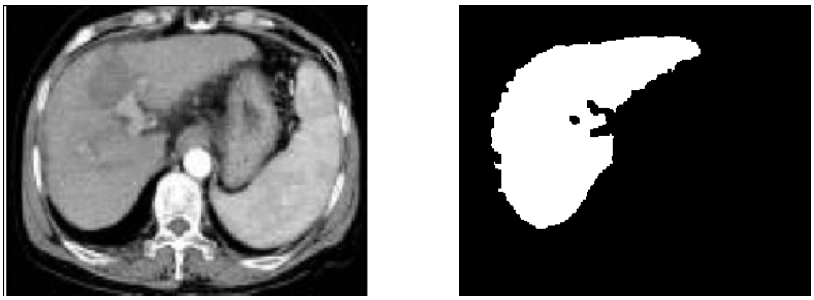


Fig. 3 Segmented CT liver image.

in Fig. 3 for the given input CT liver image. The Fig. 3 demonstrates the very acceptable segmentation result. However, the result obtained successfully by only one iteration.

4.2 CT liver segmentation algorithm

The following algorithm is formulated for CT liver image of segmentation by using FCNN:

1) Transform the color image from *(R,G,B)* to grayscale image. A grayscale digital image is an image in which the value of each pixel is a single sample, that is, it carries only intensity information. Images of this sort, also known as black-and-white, are composed exclusively of shades of gray, varying from black at the weakest intensity to white at the strongest.
2) From a grayscale image, ST method is used to create binary images.
3) Obtain the subset of a CT liver region by FCNN with the above given original template.
4) Finally, ROI is applied to the obtained subset of a CT liver region by FCNN.

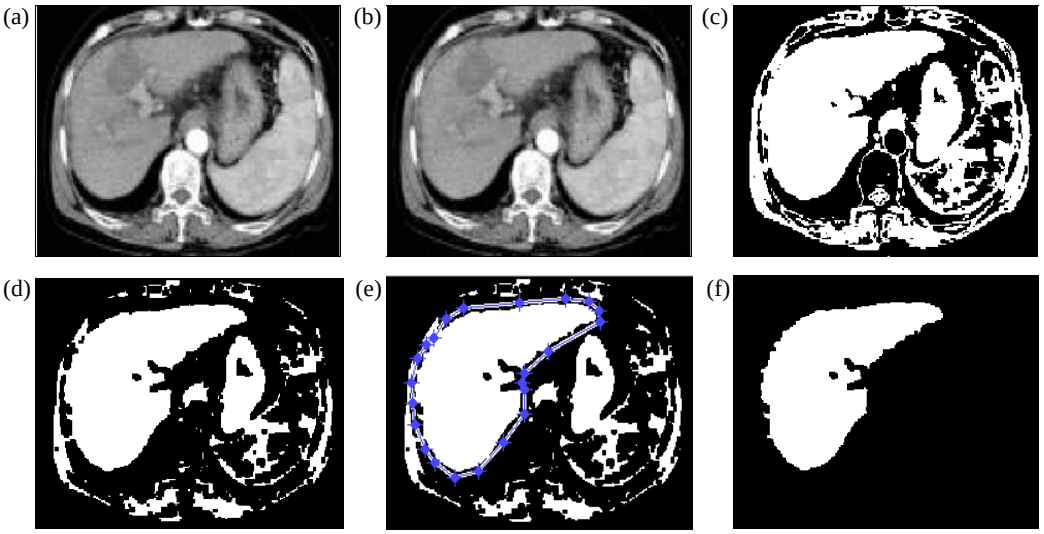


Fig. 4 Segmentation of CT liver image. (a) CT liver *(R,G,B)* color image; (b) CT liver after converting as grayscale image; (c) CT liver after applying ST; (d) CT liver after applying FCNN method; (e) CT liver after applying ROI; (f) the very acceptable segmented CT liver image.

Remark 4.1 When compared to AFCNN in [9], the proposed algorithm is very simple and more efficient. The Fig. 5(a) represents CT liver *(R,G,B)* color image. The Fig. 5(b) represents CT liver segmentation obtained by using AFCNN in [9]. The Fig. 5(c) represents CT liver segmentation obtained based on FCNN with ST and ROI methods. From Fig. 5, one can easily identify that the circle marked in Fig. 5(a) and 5(c) are similar. However, in Fig. 5(b) some information is lacked. Moreover, the result has been obtained by using AFCNN in [9] with fifty iterations, but in our paper, the more efficient result is obtained by an iteration using FCNN with ST and ROI methods.

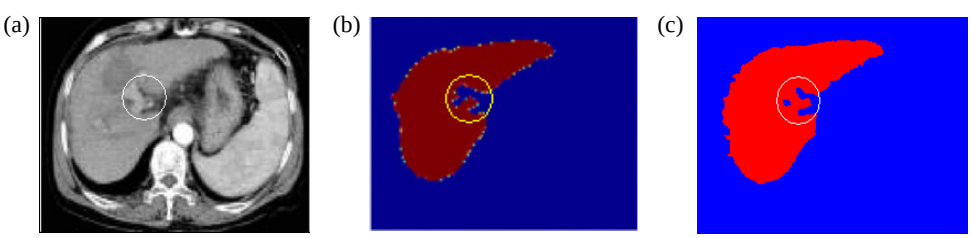


Fig. 5 Comparison of CT liver segmentation using AFCNN (see in [9]) and FCNN.

5. Conclusions and future work

In this paper, FCNN with ST and ROI methods have been intend to obtain a well defined CT liver boundary. The proposed algorithm is simpler and well organized compared to AFCNN in [9]. The experimental results in this paper illustrate that the proposed method effectively accomplishes the desired CT liver segmentation. In [9], after 50 iterations they obtained the result with some hidden information. However, based on FCNN with ST and ROI methods yield almost correct result for the liver boundary in one iteration. Moreover, another direction worthy of future study may possibly to integrate the proposed method for CT liver edge detection.

References

[1] L.O. Chua, L. Yang, IEEE Trans. Circuits Syst. Vol. 35(10) (1988), p. 1257

[2] L.O. Chua, L. Yang, IEEE Trans. Circuits Syst. Vol. 35(10) (1988), p. 1273

[3] T. Su, M. Huang, C. Hou, Y. Lin, Neural Process. Lett. Vol. 32(2) (2010), p. 147

[4] Huaqing Li, Xiaofeng Liao, Chuandong Li, Hongyu Huang, Chaojie Li, Commun. Nonlinear Sci. Numer. Simul. Vol. 16(9) (2011), p. 3746

[5] T. Yang, L.B. Yang, C.W. Wu, L.O. Chua, in: *Proceedings of the IEEE International Workshop on Cellular Neural Networks and Applications* (1996), p. 181

[6] T. Yang, L.B. Yang, C.W. Wu, L.O. Chua, in: *Proceedings of the IEEE International Workshop on Cellular Neural Networks and Applications* (1996), p. 225

[7] K. Nishizono, Y. Nishio, in: *RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'06), Waikiki Beach Marriott, Honolulu, Hawaii, USA* (2006), p. 90

[8] W. Shitong, W. Min, IEEE Trans. Inform. Tech. Biomed. Vol. 10(1) (2006), p. 5

[9] W. Shitong, F. Duan, X. Min, H. Dewen, Artif. Intell. Medicine Vol. 39 (2007), p. 65

[10] Y. Chi, J. Liu, S.K. Venkatesh, S. Huang, J. Zhou, Q. Tian, W.L. Nowinski, IEEE Trans. Biomed. Eng. Vol. 58(8) (2011), p. 2144

[11] S. Lim, Y. Jeong, Y. Ho, J. Vis. Commun. Image R. Vol. 17(4) (2006), p. 860

[12] J. Yan, T. Zhuang, B. Zhao, L.H. Schwartz, Comput. Medical Imaging Graphics Vol. 28(1-2) (2004), p. 33

[13] T. Yang, L.B. Yang, IEEE Trans. Circuits Syst. I Vol. 43(10) (1996), p. 880

2012 International Conference on Affective Computing and Intelligent Interaction
Lecture Notes in Information Technology, Vol.10

GQ2 vs. ECC: A Comparative Study of Two Efficient Authentication Technologies

(Extended abstract)*

Louis C. Guillou[1, a] and Marc Joye[2, b]

[1]Independent researcher, Bourgbarré, France

[2]Technicolor, Security & Content Protection Labs, Cesson-Sévigné, France

[a]louisguillou@orange.fr, [b]marc.joye@technicolor.com

Keywords: Authentication protocols, GQ2, ECDSA, PKI, Implementation.

Abstract. The now classical GQ scheme is a zero-knowledge proof of knowledge of a (plain) RSA signature. With a new set of keys but with exactly the same protocol, the GQ2 scheme is a zero-knowledge proof of knowledge of a decomposition of the RSA modulus in use. For dynamic authentication, efficiency reasons suggest to use GQ2 rather than RSA or GQ or even, in certain cases, ECC. This paper provides a thorough comparison between GQ2 and ECC and specifies when a given technology should be preferred over the other one.

Introduction

Authentication schemes are cryptographic schemes enabling a prover holding a private key to authenticate herself to a verifier holding the matching public key. In their seminal paper, Fiat and Shamir [1] showed how zero-knowledge techniques can help in the design of such schemes. The prover convinces the verifier in a 3-pass protocol that she possesses some secret information without revealing information whatsoever about her secret.

The GQ scheme by Guillou and Quisquater [2] is one of the most efficient extensions of the original Fiat-Shamir technique. The GQ scheme is RSA based, it uses e-th modulo N with e co-prime to $\varphi(N)$ (whereas the Fiat-Shamir scheme uses square roots modulo N). It is known to be (honest-verifier) zero-knowledge and a proof of knowledge of a plain RSA signature (i.e., an e-th root modulo N of the prover's public key) [3]. What is less known is that the GQ authentication protocol can be used with 2^k-roots modulo N. The corresponding protocol is then referred to as the GQ2 authentication protocol. It appears in ISO standard ISO/IEC 9798-5 [4]. GQ2 is also a signature scheme using the Fiat-Shamir heuristic. GQ2 signature appears in ISO standard ISO/IEC 14888-2 [5]. The main advantage of GQ or GQ2 signature over RSA signature is that it is typically one order of magnitude faster. Compared to GQ, GQ2 offers the advantage of being compatible with RSA. A user possessing an RSA key pair can use it with GQ2 for dynamic authentication or signature.

Yet another way to obtain efficient schemes is to rely on elliptic curve cryptography (ECC). There is no standardized authentication protocol based on elliptic curves. But there is a standard for digital signature, the Elliptic Curve Digital Signature Algorithm (ECDSA) [6,7,8], which can be used for dynamic authentication purposes as follows. The verifier chooses a message uniformly at random and requests the prover to produce a valid signature on the message.

Elliptic curve cryptography is becoming more and more popular. Nevertheless, RSA is still dominating the security marketplace for public-key cryptography. Moving from RSA to ECC implies

* The full version of this paper is available on the Cryptology Eprint Archive at URL http://eprint.iacr.org/.

changing the underlying public-key infrastructure (PKI). This has a cost. A simpler alternative would be to move from RSA to GQ2. But what is the expected performance? How does GQ2 compare to ECC? This paper answers this latter question and provides a detailed analysis of the two technologies with concrete implementations.

GQ2 Authentication and Signature

Squares and Square Roots. Consider a prime number p. The multiplicative group of integers modulo p, namely $\mathbf{Z}_p^*$, has $p-1$ elements: $\{1, \ldots, p-1\}$. An element a in $\mathbf{Z}_p^*$ is a square if and only if there exists some α in $\mathbf{Z}_p^*$ that $a = \alpha^2$. Fermat theorem says that

$$a^{p-1} = 1 \pmod{p}$$

for any a in $\mathbf{Z}_p^*$. Therefore, an element a is a square in $\mathbf{Z}_p^*$ (or a quadratic residue modulo p) if and only if $a^{(p-1)/2} = 1 \pmod{p}$. This is known as Euler's criterion and gives rise to the Legendre symbol:

$$(a|p) := a^{(p-1)/2} \bmod p = \begin{array}{l} 1 \text{ if } a \text{ is a quadratic residue modulo } p \\ -1 \text{ if } a \text{ is a quadratic non-residue modulo } p \end{array}$$

Suppose further that prime p is odd (i.e., $p \neq 2$). Then we can write

$$p = 2^b t + 1 \text{ with } t \text{ odd and } b \geq 1 .$$

When b = 1 —or equivalently when $p \equiv 3 \pmod 4$, if a is a quadratic residue modulo p then its two square roots are

$$\alpha = a^{(p+1)/4} \pmod{p} \text{ and } -\alpha \pmod{p}$$

Among these two roots, it is easily verified that α is itself a quadratic residue modulo p and that $-\alpha$ is not. In other words, square roots exist and are unique in the subgroup of quadratic residues modulo p. More generally, if we define the cyclic subgroup

$$\mathbf{S}_p = \{x^{**}2^b \mid x \text{ in } \mathbf{Z}_p^* \text{ where } p = 2^b t + 1 \text{ with } t \text{ odd}\}$$

then square roots exist and are unique in $\mathbf{S}_p$. Given an element a in $\mathbf{S}_p$, its square root is given by

$$\alpha = a^{(t+1)/2} \pmod{p} .$$

Indeed, since any element a in $\mathbf{S}_p$ is 2^b-power residue, it follows that $a^t = a^{(p-1)/2^{**}b} = 1$, and consequently $\alpha^2 = a^{t+1} = a^t a = a$. Iterating the process, for any $k \geq 1$, 2^k-th roots exist and are unique in $\mathbf{S}_p$: $a = \beta^{**}2^k$ is equivalent to $\beta = a^x$ with $x = ((t+1)/2)^k \bmod t$.

Basic Protocol.

Key generation: Any GQ scheme makes use of a generic equation

$$GQ^v \equiv 1 \pmod{N} \qquad (1)$$

where G is a public number, Q a private number, N an RSA-type modulus, and v a verification exponent.

Let $N = p_1 p_2$ where $p_1 = 2^{b_1} t_1 + 1$ and $p_2 = 2^{b_2} t_2 + 1$ are prime, with t_1, t_2 odd and b_1, $b_2 \geq 1$. Let g be a small prime number such that

$$\begin{array}{l} (g|p_1) = -1, (g|p_2) = 1 \text{ if } b_1 > b_2 \\ (g|p_1) = 1, (g|p_2) = -1 \text{ if } b_1 < b_2 \\ (g|p_1) = -(g|p_2) \qquad \text{if } b_1 = b_2 \end{array} .$$

For example, the following values are pertinent:

- $g = 2$ with $p_1 \equiv 3 \pmod 8$ and $p_2 \equiv 7 \pmod 8$, i.e., $N \equiv 5 \pmod 8$;
- $g = 3$ with $p_1 \equiv 1 \pmod 3$ and $p_2 \equiv 2 \pmod 3$, i.e., $N \equiv 2 \pmod 3$;

– $g = 5$ with $p_1 \equiv \pm 2 \pmod 5$ and $p_2 \equiv \pm 1 \pmod 5$, i.e., $N \equiv \pm 2 \pmod 5$; and so on.

Define $b = \max(b_1, b_2)$ and let $G = g^{**}2^b$. Remark that G is in $\mathbf{S}_{p1} \times \mathbf{S}_{p2}$ while g is not since g is a quadratic non-residue in $\mathbf{Z}_N^*$. Let $v = 2^{k+b}$ be the verification exponent where $k \geq 1$ is a security parameter. Finally, let $Q_i = G^{**}x_i \bmod p_i$ with $x_i = -((t_i + 1)/2)^{k+b} \bmod t_i$ ($i = 1,2$). By construction, letting now $Q = \mathrm{CRT}(Q_1,Q_2)$ (i.e., $Q \equiv Q_1 \pmod{p_1}$ and $Q \equiv Q_2 \pmod{p_2}$) yields $GQ^v \equiv 1 \pmod N$ as per Eq. (1).

The public key is $\{N, b, k, g\}$ and the private key is $\{p_1, p_2, Q_1, Q_2\}$.

Authentication: GQ2 authentication protocol between a prover, say Peggy, and a verifier, say Victor, is given by the following 3-pass scheme.

Peggy		**Victor**
random r in $\{1, \ldots, N-1\}$		
$W = r^v \bmod N$	$— W →$	
	$← d —$	random d in $\{0,1\}^k$
$D = rQ^d \bmod N$	$— D →$	
		$W =^? D^v G^d \bmod N$

The exchanged quantities, W, d, D, are called witness, challenge and response, respectively. The protocol can be iterated several times.

Signature: Companion GQ2 signature scheme is obtained thanks to the Fiat-Shamir heuristic [1] (see also [9]). The challenge is replaced with the hash value of the witness and of the message M to be signed. So, the signature σ on a message M is given by

$\sigma = (d, D)$ with $d = H(W, M)$ where $W = r^v \bmod N$, and $D = rQ^d \bmod N$

for some cryptographic hash function $H : \{0,1\}^* \rightarrow \{0,1\}^k$. The validity of signature $\sigma = (d, D)$ is then verified by checking whether $d = H(W', M)$ with $W' = D^v G^d \bmod N$.

Security Considerations. In the authentication protocol, Peggy proves in a zero-knowledge fashion that she knows a solution Q to Eq. (1), which, in turn, implies the knowledge of the factorization of N. Indeed, suppose without loss of generality that $(g|p_1) = -1$ and $(g|p_2) = 1$ (and thus $b_2 \leq b_1$ and $b = b_1$). Letting $h = g^{**}2^{b-1} \bmod N$ and $z := (Q^{-1})^{**}2^{k+b-1} \bmod N$, Equation (1) yields

$h^2 \equiv z^2 \pmod N$ or equivalently $(h - z)(h + z) \equiv 0 \pmod{\{p_1, p_2\}}$.

Since $k \geq 1$, first note that it is a 2^b-power residue modulo N. But $h \neq z \pmod{p_1}$ because h is not a 2^b-power residue modulo p_1 (i.e., h not in $\mathbf{S}_{p1}$) since $(g|p_1) = -1$, and $h \equiv z \pmod{p_2}$ because h in $\mathbf{S}_{p2}$ since $(g|p_2) = 1$. Hence,

$$\gcd(h - z, N)$$

will disclose p_2.

The three security properties zero-knowledge proofs of knowledge must fulfill are completeness, soundness, and zero-knowledge.

Completeness: Completeness means that given an honest prover and an honest verifier, the protocol succeeds with overwhelming probability. For GQ2, it is easy to see that $GQ^v \equiv 1 \pmod N$ implies that $r^v \equiv (rQ^d)^v G^d \pmod N$ for any challenge d.

Soundness: Soundness means that a dishonest prover should not be able to convince an honest prover. More formally, soundness means there exists a (polynomial-time) knowledge extractor K such that if a dishonest prover successfully executes the protocol with non-negligible probability then K can recover some information allowing successful subsequent protocol executions.

From two accepting conversations sharing the same witness in GQ2 authentication, say (W, d, D) and (W, d', D'), knowledge extractor K can factor RSA modulus N as follows.

1. Let $\gcd(v, d - d') = 2^{\delta}$ where $0 \leq \delta \leq k - 1$ (since d, d' in $\{0, 1\}^k$ and $d \neq d'$) and write $d - d' = 2^{\delta}\tau$ where τ is odd;
2. Observing that $W \equiv D^v G^d \equiv (D')^v G^{d'} \pmod{N}$, compute $H = (g^{\tau})^{**}2^{b-1} \bmod N$ and $Z = (D'/D)^{**}2^{k+b-\delta-1}$, and $\gcd(H - Z, N)$, which yields a non-trivial factor of N. It is worth noting here that H^2 and Z (and thus any power thereof) are in $\mathbf{S}_{p1} \times \mathbf{S}_{p2}$.

Zero-knowledge: The property of zero-knowledge ensures that the prover does not disclose any information about its secret knowledge while interacting with the verifier. For GQ2, it is possible to simulate conversations that are indistinguishable from real conversations (i.e., from those resulting from protocol executions with the real prover). This can be achieved by first randomly picking d^* randomly in $\{0,1\}^k$ and D^* randomly in $\mathbf{Z}_N^*$, and next by computing $W^* = (D^*)^v G^{d^*} \bmod N$. If ($W$, d, D) denotes a real conversation, it is easily verified that the two distributions (W^*, d^*, D^*) and (W, d, D) are indistinguishable.

Compatibility with RSA. One of the main advantages of GQ2 technology resides in its compatibility with RSA. Compatibility with RSA is important as it allows GQ2 to rely on the public-key infrastructure (PKI) already deployed for RSA.

Public parameters: Any RSA modulus N and its associated certificate can be used as it is within GQ2. The other public parameters, namely (b, k, g), are either given by the user or defined by the application.

Remark that one can verify that g is a quadratic non-residue modulo N from its Jacobi symbol; i.e., by checking that $(g|N) = -1$. If $N = p_1 p_2$, this ensures that $(g|p_1) = -(g|p_2)$. Remark also that $(g|N) = -1$ implies that the very existence of a solution Q to Eq. (1) and that N is composite.

Private parameters: When RSA is implemented with Chinese remaindering (a.k.a. RSA in CRT mode), the user knows secret primes p_1 and p_2 of RSA modulus N and can therefore derive the corresponding private key, $\{p_1, p_2, Q_1, Q_2\}$.

In contrast, when RSA is implemented in standard mode, the user has only access to the triple (N, e, d) where $ed \equiv 1 \pmod{\varphi(N)}$. In this case, the user does not have directly the knowledge of primes p_1 and p_2. But the user can easily recover them thanks to Miller's algorithm [10] because a multiple of the group order, $\#\mathbf{Z}_N^* = \varphi(N)$, is known; i.e., $ed - 1$.

Extended Protocol. The efficiency of the basic protocol can be improved using a multi-prime RSA-type modulus. Define $N = p_1 \ldots p_{m+1}$ for large (equal-size) primes $p_i = 2^{b_i} q_i + 1$ with q_i odd and $b_i \geq 1$. For better efficiency, it is advantageous to choose $b_i = 1$. Parameter $m \geq 1$ is an additional security parameter. The global security is given by the product $k \times m$. Typical values, depending on the application context, are:

- $k \times m = 8$: weak authentication mode;
- $k \times m = 32$: strong authentication mode;
- $k \times m = 80$: signature mode (or more generally, $k \times m \geq 80$).

Parameter b is now defined as $b = \max_{1 \leq i \leq m} b_i$. There is also a set of m quadratic non-residues, $\{g_1, \ldots, g_m\}$, that are selected in a way similar as for the basic protocol.

Elliptic Curve Cryptography

Discrete-log based cryptosystems can be transposed in the setting of elliptic curves over a finite field. Since the underlying problem is substantially harder in the case of elliptic curve cryptography, smaller parameters can be used but with the same expected security level [11,12]. The main schemes

based on elliptic curves are ECDSA for digital signatures, ECDH for key exchange and ECIES for encryption.

Let E be an elliptic curve defined over the prime field GF(p), of order $\#E(\mathrm{GF}(p)) = hn$ for some co-factor h in {1,2,3,4} and prime n. Let also G in E be a base point of order n. The domain parameters for ECDSA are $\{E, p, G, n, h\}$. An ECDSA key pair is associated with a given set of domain parameters. Each user chooses a random element d in $\{1, \ldots, n-1\}$ and computes the point $Q = [d]G$ in E. The public key is Q and the private key is d.

Before detailing the signing/verification processes, it is useful to introduce some notation. The neutral element of elliptic curve E (namely, the point at infinity) is denoted by O. Given a point P in $E \setminus \{O\}$, its x-coordinate is denoted by $\mathrm{x}(P)$.

The ECDSA signature on a message M is obtained as follows:

1. randomly choose k in $\{1, \ldots, n\text{-}1\}$;
2. compute $r = \mathrm{x}([k]G) \bmod n$; if $r = 0$ go to Step 1;
3. compute $s = k^{-1}(H(M) + dr)$; if $s = 0$ go to Step 1;
4. return $\sigma = (r, s)$.

The validity of $\sigma = (r, s)$ is then checked using public key Q as follows:

1. verify that r, s in $\{1, \ldots, n\text{-}1\}$;
2. compute $w = s^{-1} \bmod n$, $u_1 = H(M)w \bmod n$ and $u_1 = rw \bmod n$;
3. compute $X = [u_1]G + [u_2]Q$;
4. accept the signature if and only if $X \neq O$ and $\mathrm{x}(X) \equiv r \pmod n$.

Comparison

The next three tables give a comparative of 1024-bit GQ2 and 160-bit ECC technologies on 8-, 16- and 32-bit processors. For a given amount of RAM, they list the number of needed CPU multiplications, under certain implementation assumptions (see the full paper for details).

Table 1. GQ2 vs. ECC: 8-bit processor

RAM (bytes)	GQ2 technology			ECC technology
	Weak auth.	Strong auth.	Signature	signature
240	—	—	—	2,344,860
450	399,232	1,397,632	—	1,988,175
460	↓	↓	3,394,432	↓
520	↓	↓	↓	1,867,582

Table 2. GQ2 vs. ECC: 16-bit processor

RAM (bytes)	GQ2 technology			ECC technology
	Weak auth.	Strong auth.	Signature	Signature
240	—	—	—	614,130
450	101,312	354,752	—	520,713
460	↓	↓	861,132	↓
520	↓	↓	↓	489,129

Table 3. GQ2 vs. ECC: 32-bit processor

RAM (bytes)	GQ2 technology			ECC technology
	Weak auth.	Strong auth.	Signature	Signature
240	—	—	—	167,490
450	26,080	91,360	—	142,013
460	↓	↓	221,920	↓
520	↓	↓	↓	133,399

GQ2 technology requires at least 450 bytes of RAM memory for enabling its implementation. In environments with fewer RAM memory, only ECC technology is available (the implementation we gave requires at least 240 bytes of RAM memory). When the amount of available RAM memory is of the order of 500 bytes (or more), GQ2 in strong authentication mode outperforms ECC. This may appear counter-intuitive as GQ2 makes use of much larger key sizes. This is especially true when the processor size (i.e.,) is large. Besides, the table assumes the basic GQ2 protocol, which is compatible with RSA. Better performance (in both memory and speed) can be obtained using the extended GQ2 protocol. Yet another advantage of GQ2 is that it offers a fine-grained control on the authentication strength (parameter *k*) for better efficiency/security trade-offs.

In summary, GQ2 technology is particularly well suited to environments featuring moderate size RAM memory (500 bytes or more). Moreover, it can make use of the already-deployed PKI for RSA. ECC technology is more intended to very constrained environments.

References

[1] A. Fiat and A. Shamir. How to prove yourself: Practical solutions to identification and signature problems. In: *Advances in Cryptology – CRYPTO '86*, vol. 263 of *Lecture Notes in Computer Science*, pp. 186–194. Springer, 1987.

[2] L. C. Guillou and J.-J. Quisquater. A practical zero-knowledge protocol fitted to security microprocessor minimizing both transmission and memory. In: *Advances in Cryptology – EUROCRYPT '88*, vol. 330 of *Lecture Notes in Computer Science*, pp. 123–128. Springer, 1988.

[3] L. C. Guillou and J.-J. Quisquater. A"paradoxical" identity-based signature scheme resulting from zero-knowledge. In: *Advances in Cryptology – CRYPTO '88*, vol. 403 of *Lecture Notes in Computer Science*, pp. 216–231. Springer, 1990.

[4] ISO/IEC 9798-5. Information technology – Security techniques – Entity authentication – Part 5: Mechanisms using zero-knowledge techniques. International Organization for Standardization, 2009.

[5] ISO/IEC 14888-2. Information technology – Security techniques – Digital signatures with appendix – Part 2: Integer factorization based mechanisms. International Organization for Standardization, 2008.

[6] ANSI X9.62. Public key cryptography for the financial services industry, the elliptic curve digital signature algorithm (ECDSA). American National Standards Institute, 2005.

[7] FIPS 186-2. Digital signature standard (DSS). National Institute for Standards and Technology, 2000.

[8] IEEE P1363. Standard specifications for public-key cryptography. IEEE Standards Association, 2000.

[9] D. Pointcheval and J. Stern. Security arguments for digital signatures and blind signatures. *Journal of Cryptology*, 13(3):361–396, 2000.

[10] G. L. Miller. Riemann's hypothesis and tests for primality. *Journal of Computer and System Sciences*, 13(3):300–317, 1976.

[11] N. Koblitz. Elliptic curve cryptosystems. *Mathematics of Computation*, 48(177):203–209, 1987.

[12] V. S. Miller. Use of elliptic curves in cryptography. In: *Advances in Cryptology – CRYPTO '85*, vol. 218 of *Lecture Notes in Computer Science*, pp. 417–426. Springer, 1985.

Design and Implementation of an Adaptive Control Mechanism for Standby Power Detection and Saving

Shun-Chieh Lin, Huan-Wen Tsai, Yi-Lin Chiang and Tsung-Lin Tsai

Cloud Service Technology Center, Industrial Technology Research Institute,

No.31, Gongye 2nd Rd., Tainan, Taiwan (R.O.C.)

{jason.lin, hwtsai, yilin, will}@itri.org.tw

Keywords: Power saving, standby detection, adaptive method, household appliance.

Abstract. In this paper, an adaptive method for monitoring and controlling household appliances is proposed and can be implemented on resource-limited embedded systems. This method contains the following steps: monitoring and sampling the power of the household appliances to obtain a first standby power range; monitoring and sampling the power of household appliances to obtain a set of first real-time power data; calculating a first standby confidence level based on the number of times that the first real-time power data falls within the first standby power range; and determining that the household appliance is in a standby state if the first standby confidence level is greater than or equal to a standby confidence level threshold. The proposed method is implemented on Silicon Lab MCU F930 and the experiment results show that the proposed method can be applied to different appliances and perform well on standby detection and power saving.

Introduction

A large number of household appliances and office equipments such as televisions (TVs), video recorders, audio players, telephone answering and facsimile machines, computers, printers and copiers cannot be switched off completely without being unplugged. These appliances consume power 24 hours a day, often without the knowledge of the consumer with typical loss per appliance ranging from less than 1W to as much as 25W. This is called ‘standby power’, also known variously as vampire power, vampire draw, phantom load, or leaking electricity. Lee and Yun [4] show that Set-top Box (STB) which receives broadcasting signal and delivers it to display device such as TV usually does not have low-power mode, or standby power mode. The main reasons come from technical barriers and operational stability and STB normally consumes 80~90% power of active mode even when the STB is off. In addition, according to the IEA (International Energy Agency), 3%~13% of a total household (OECD) power consumption is being wasted in the form of standby power and IEA estimates that standby produces 1% of the world's CO2 emissions. Therefore, to meet energy saving requirements, saving power during a standby state has become an important issue for those skilled in the household appliances field.

In the prior art, a lot of efforts were made to reduce power consumption in the field of hardware, software, and chips [1,2,3,5]. Heo et al. [3] presented a reduction control mechanism which has the host-Agent based structure and uses the IEEE 802.15.4 based ZigBee protocol for communication and security between Host and Agents. Ju et al. [5] developed a home energy management system to plain energy management mechanism for power consumption reduction. It requires user's minimal decision and action for correct judgment. It also uses Automatic Meter Reading (AMR) network based on power line communication (PLC). In these previous systems, manual switches or timer switches are

usually used to save power during a standby state. Although some automatic techniques are also developed to detect the standby state, e.g., using a fixed threshold of current/power or a specific ratio of a variation of current/power, their work still can not detect properly while the appliance is old or when applies to different household appliances with the same threshold strategy.

Figure 1 shows that four different types of appliances have different power consumption and different standby power. According to [6], the average standby power of set top box, washer, and inkjet printer is 15.24W, 4.19W, and 4.48W with the standby duration of 19.3 hours, 23.2 hours, and 23.5 hours, respectively. Therefore, systems for monitoring and controlling a standby state of various household appliances are needed. In this paper, we propose an automatic adaptation method for monitoring and controlling the standby status of a household appliance, comprising the steps of: monitoring and sampling the power of the household appliance to obtain a first standby power range; monitoring and sampling the power of the household appliance to obtain a set of first real-time power data; calculating a first standby confidence level based on the number of times that the first real-time power data fall within the first standby power range; and determining that the household appliance is in a standby state if the first standby confidence level is greater than or equal to a standby confidence level threshold. The proposed method can be easily implemented on resource-limited embedded systems. With the proposed system, it is not necessary to manually set the standby confidence level threshold and can be applied to different appliances for standby detection and power saving.

This paper is organized as follows. Section 2 gives an overview of the proposed system and explains about related works such as the type of standby power and the phenomenons in power usage. Section 3 describes the proposed standby status detection system; this section also explains three processing modes of standby status detection system. Implementation results are discussed in section 4. In the end, we have given some concluding remarks in Section 5.

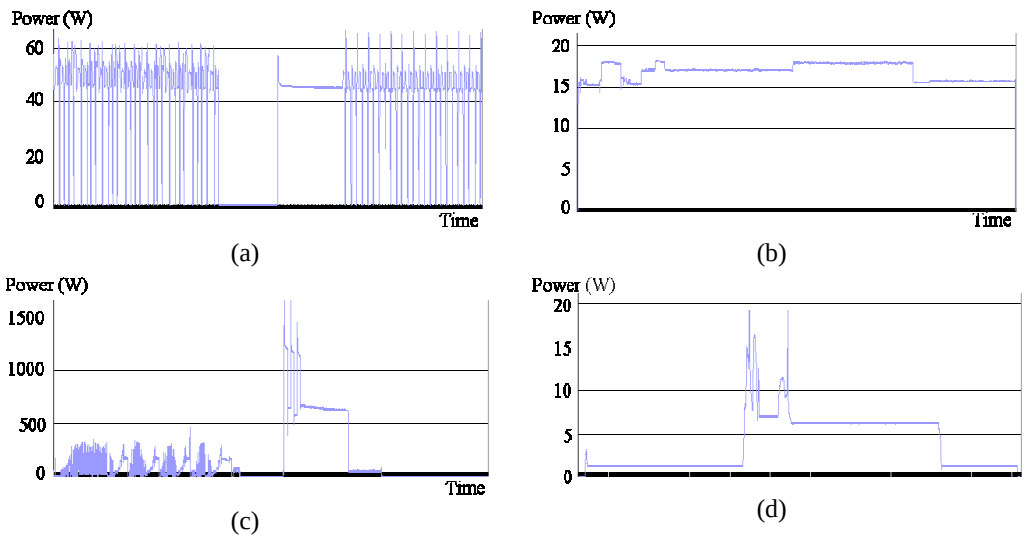


Fig. 1. The power consumption and standby power usage of: (a) an electric fan in a “natural” option operation; (b) a Set-top Box (STB) in data delivering; (c) a combination all-in-one washer and dryer machine from washing to drying; (d) an inkjet printer while printing a document.

System Overview

Several types of standby power are defined in [3]. As shown in Table 1, standby power of many

devices can further be classified into no-load mode, off mode, passive standby mode, active standby mode, and sleep mode. The emergence of active standby power started with the introduction of set top boxes. It is a power mode where the consumer switches off the power (the consumer thinks the power is switched off completely) but the internal circuit still consumes standby power to wait for external cord/cordless signals. As shown in Fig. 2, a set top box in active standby status is a short period and can be distinguished from passive standby status.

Table 1. Definition and types of standby power.

Category	Description	Power Status	Products
No Load	State of the power supply when no power is being provided to the rest of the appliances	-	External power supply(DC/AC power supply, phone battery charger)
Off	The appliances is switched off and has no capacity	Put-Off	TV, VCR, audio, DVD player, PC, monitor, printer
Passive Standby (PS)	The appliances is off, but can be powered up remotely	Put-Off	TV, VCR, audio, DVD player, Set top box
Active Standby (AS)	The appliances is on, but is not providing a primary function	Put-Off	Set top box, homenetwork system
Sleep	Mode entered after a period inactivity	Put-On	PC, monitor, printer, facsimile, scanner

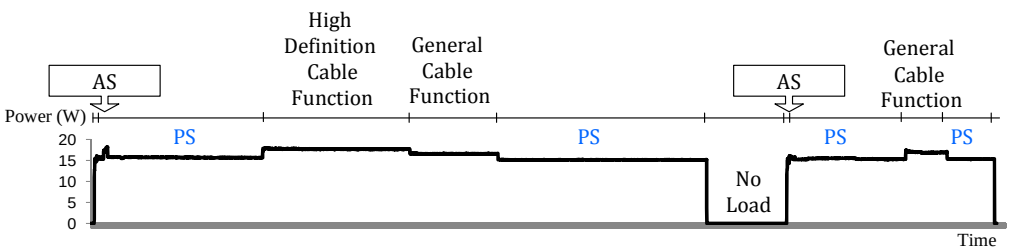


Fig. 2. Power usage of a set top box while in status of “active standby (AS)” mode, “passive standby (PS)” mode, high-definition (digital) cable function, general cable function, and “No Load” mode.

Therefore, referring to the examples of power usage shown in Fig. 1 and Fig. 2, three phenomenons can be discovered as shown in Fig. 3 and followed by descriptions:

I. Volatility in power usage: as showed in Fig. 1 (a), (c), and (d). Home appliances power consumptions have high volatility depends on function provided or the power system design such as active standby mode in STB. Therefore, the proposed system uses a confidence level verification to filter high volatility power usage.

II. Nonstationary standby power after operation: Temperature raise in certain appliance after operating for a period of time will lead to higher power consumption as well. When appliances are forced into standby mode by high temperature, the standby power consumption will be different from the original. To enhance the detection of standby status, the proposed system uses an adaptation method to apply detection in different power consumption status of a home appliance.

III. Similar power consumption in standby mode and non-standby mode: As shown in Fig. 1(b) and Fig. 2, STB has this phenomenon which the power consumption range is 16.5~18.3W in use and 15.1~15.8W when standby. The difference of power usage between non-standby and standby is less than 3W. In this work, the proposed system initializes standby power model training process to obtain the standby power range.

Figure 4 shows the flowchart of the proposed system. The method for monitoring and controlling the standby state of the household appliance includes three modes: training mode, detection mode, and adaptation mode. After the coupling the proposed system to the household appliance, the Training Mode is initiated. As the training mode is completed, the Detection Mode and the Adaptation Mode are performed to monitor the power usage of household appliance for standby status detection and control adaptation.

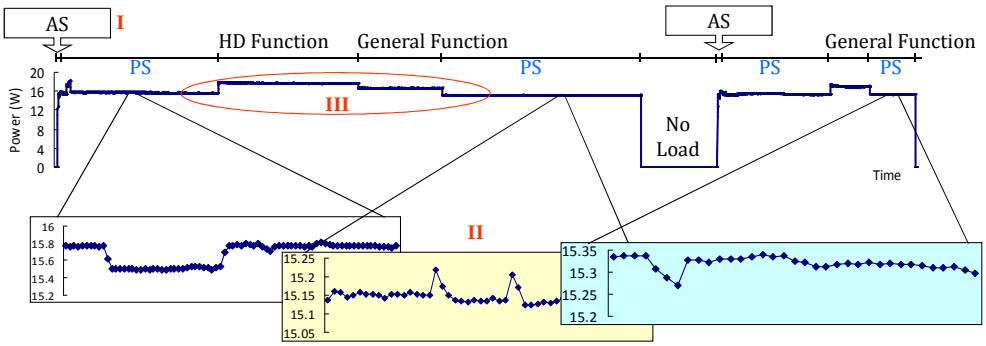


Fig. 3. A STB example of three phenomenons in power usage: I) Volatility in power usage, II) Nonstationary standby power after operation, and III) Similar power consumption in standby mode and non-standby mode.

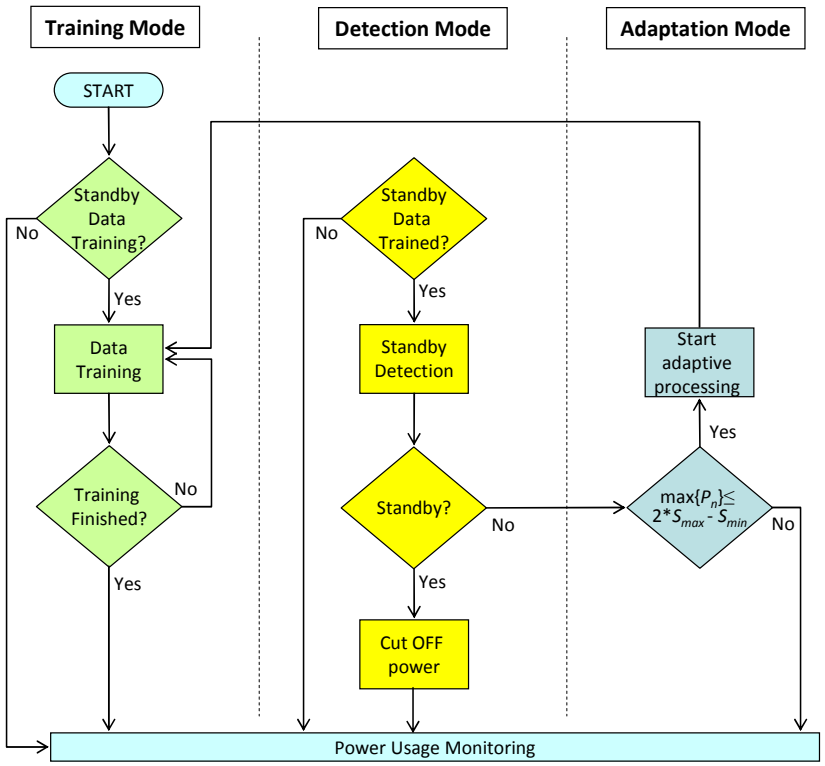


Fig. 4. The system flowchart of the proposed system comprising of three modes: training mode, detection mode, and adaptation mode.

The Proposed Standby Status Detection System

The Process of Training Mode. After coupling the system to the household appliance, the Training Mode is initiated. The process of Training Mode comprises: a) monitoring and sampling the power of the household appliance to obtain a set of first power data, b) establishing a first power range based on the first power data, c) monitoring and sampling the power of the household appliance to obtain a set of second power data, d) calculating a trained confidence level based on the number of times that the second power data fall within the first power range, and e) determining the first power range as the first standby power range if the trained confidence level is greater than or equal to a trained confidence level threshold.

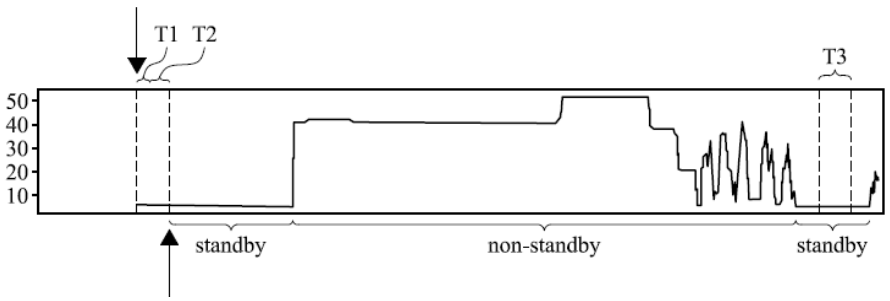


Fig. 5. An example of the power consumed by a household appliance.

As shown in Fig. 5, the household appliance is usually in the standby state at the beginning of operation. Therefore, in this case, when detecting a connection between the system and the household appliance, monitoring and sampling the power of the household appliance is initiated. After monitoring and sampling the power of the household appliance for a period T_1, a plurality of first power data, for example, P_{11}~P_{1N}, is obtained. Then use the first power data P_{11}~P_{1N} obtained to establish a first power range S_{min}~S_{max} of the household appliance. Specifically, the upper limit S_{max} of the first power range P_{11}~P_{1N} is the average μ of the first power data P_{11}~P_{1N} plus the product of a system-determined constant Z and the standard deviation σ of the first power data P_{11}~P_{1N}, and the bottom limit S_{min} of the first power range P_{11}~P_{1N} is the average μ of the first power data P_{11}~P_{1N} minus the product of a system-determined constant Z and the standard deviation σ of the first power data P_{11}~P_{1N}, it also can be shown as follows:

$$S_{\max} = \mu + \mathrm{Z} \times \sigma \tag{1}$$

$$S_{\min} = \mu - \mathrm{Z} \times \sigma \tag{2}$$

wherein the average μ and the standard deviation σ of the first power data P_{11}~P_{1N} can be obtained by the following Equations:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} P_{1i} \tag{3}$$

$$\sigma = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (P_{1i} - \mu)^2} \tag{4}$$

Note that the system-determined constant Z is the (1-$\alpha/2$)-quantile of a unit normal variate, and α is a significance level, which can be referred to Table 2.

Table 2. Quantile of a unit normal variate.

Confidence Level (%)	α	α/2	$Z_{1-\alpha/2}$
20	0.8	0.4	0.253
40	0.6	0.3	0.524
60	0.4	0.2	0.842
68.26	0.3174	0.1587	1.000
80	0.2	0.1	1.282
90	0.1	0.5	1.645
95	0.05	0.025	1.960
95.46	0.0454	0.0228	2.000
98	0.02	0.01	2.326
99	0.01	0.005	2.576
99.74	0.0026	0.0013	3.000
99.8	0.002	0.001	3.090
99.9	0.001	0.0005	3.29
99.98	0.0002	0.0001	3.72

After monitoring and sampling the power of the household appliance for a period T_2, a plurality of second power data, for example, P_{21}~P_{2M}, is obtained. The proposed method calculates a trained confidence level r_s based on the number of times that the second power data P_{21}~P_{2M} falls within the first power range S_{min}~S_{max}. The trained confidence level r_s, for example, is a ratio between the number of times that the second power data P_{21}~P_{2M} falls within the first power range S_{min}~S_{max} and the number of times the second power data P_{21}~P_{2M} was sampled, as follows:

$$r_s = M'/M \tag{5}$$

$$M' = \sum_{n=1}^{M} g_n \text{, where } g_n = \begin{cases} 1, & \text{if } S_{\min} < P_{2M} < S_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In this case, the trained confidence level threshold can be predetermined as 0.95. If the trained confidence level calculated in (5) is greater than or equal to the trained confidence level threshold 0.95, the first power range S_{max}~S_{min} can be regarded as the "standby power range" of the household appliance. Note that, at the same time, the variation threshold required in the Adaptation Mode can also be obtained. For example, the variation threshold δ can be obtained based on the upper limit of the first power range S_{max} and the bottom limit of the first power range S_{min} as follows:

$$\delta = 2S_{\max} - S_{\min} \tag{7}$$

The Process of Detection Mode. After the training mode is completed, the Detection Mode is performed to monitor the household appliance. Referring to Fig. 5, the power of the household appliance is monitored for a period T_3 to sample a set of real-time power data. A standby confidence level can be calculated based on the number of times that the real-time power data falls within the standby power range. Similar to the calculation for the trained confidence level in (5) and (6), the standby confidence level can be the ratio between the number of times that the real-time power data falls within the standby power range and the number of times that the real-time power data is sampled.

In this work, a standby confidence level threshold of 0.95 is predetermined. For detection process, it is determined whether the standby confidence level is greater than or equal to the standby confidence level threshold 0.95. If the determination result is “yes” (i.e., the standby confidence level is greater than or equal to the standby confidence level threshold of 0.95), the control of power saving will be performed. For the purpose of energy conservation, when the household appliance is determined to be in the standby state, the power of the household appliance should be cut off to prevent the household appliance from consuming power.

The Process of Adaptation Mode. In Detection Mode, if it is determined that the standby confidence level is smaller than the standby confidence level threshold 0.95, the proposed system will be performed to determine whether the real-time power data is less than the variation threshold. If the determination result is “No” (i.e., the real-time power data is greater than or equal to the variation threshold), it means that the power consuming situation of the household appliance is normal, and the household appliance is determined to be in non-standby state and is qualified to be provided with power. If the determination result is “Yes” (i.e., the real-time power data is less than the variation threshold), it means that the power consuming situation of the household appliance may have probably changed, and the standby power range obtained in the Training Mode is not precise and should be modified. Then, the proposed system will be performed to re-obtain a new standby power range of the household appliance.

In summary, referring to the Training Mode of this work, the power of the household appliance is monitored and sampled to train a first standby power range and a variation threshold. And with the use of the Detection Mode, the power of the household appliance is monitored and sampled to obtain a set of first real-time power data, and a first standby confidence level is calculated based on the number of times that the first real-time power data falls within the first standby power range. Note that the Training Mode and the Detection Mode will be performed again when the first standby confidence level determined in the Detection Mode is to be less than the standby confidence level threshold and the first real-time power data determined in the Adaptation Mode is to be less than the variation threshold. In the Detection Mode, the second standby confidence level is greater than or equal to the standby confidence level threshold, and the household appliance is determined to be in the standby state. And the second standby confidence level is less than the standby confidence level threshold, wherein it is further determined whether the power consuming situation of the household appliance has changed according to the second real-time power data.

Experimental Results

In this work, standby power detection of home appliances has been performed using our proposed method and the ratio-based method [7]. Referring to [7], the method permits the detection of `standby` state` in linear and non-linear loads connected to the power grid and their automatic disconnection, and comprises 4 main stages: a) detection of the normal operating state of at least one load by means of detection, obtaining the maximum value of the current in said operating state, b) detection of entry into `standby` mode` of the at least one load by the means of detection establishing said `standby` state` when the existing current value obtained in each load is less than a percentage *P* of the maximum value of the current of each load in normal operating state, c) start of means of timing at a determined time *T* for each load when it goes into `standby` mode`, and d) disconnection of at least one load and of the means of detection when value *T* is reached of the means of timing without the at least one load having returned to normal operating state. The proposed method is implemented on Silicon Lab MCU F930 with a power meter ADI ADE7736. The proposed system is compatible with Home Plug Command & Control (HPCC) 1.0 specification which is a power line communication (PLC) network standard.

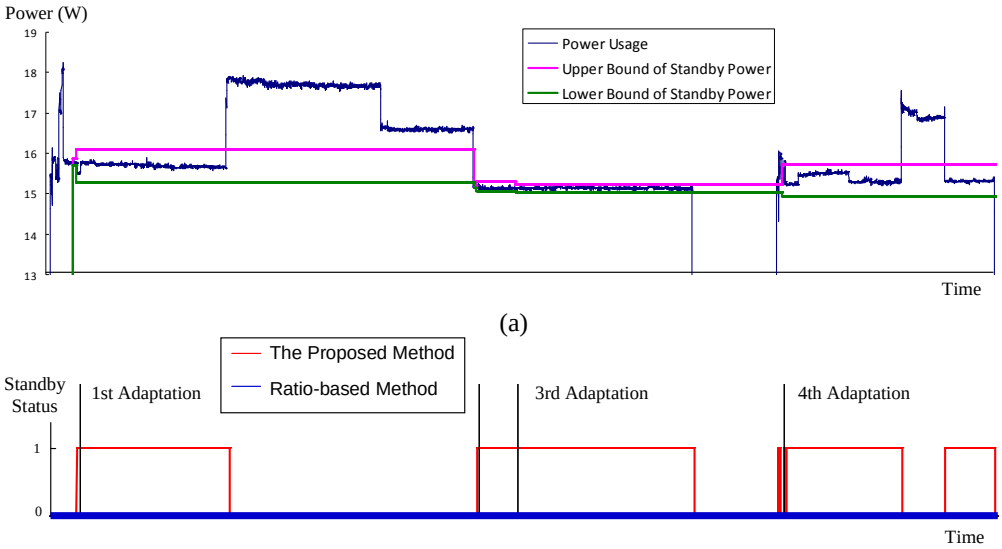


Fig. 6. An example of STB standby detection results: (a) detection range adaptation of standby power and (b) comparison of the proposed method and ratio-based method.

An example of STB standby detection results of the proposed system is shown in Fig. 6. With the proposed adaptive process, the detection range of standby power can be changed with time. Figure 6(a) shows the upper and lower bounds of standby power detection range have been performed four times adaptation as shown in Fig. 6(b). Each standby power detection range is adapted by real-time power consumption into different width. A comparison of the proposed method and ratio-based method is also shown in Fig. 6(b). The standby status is always 0, meant to be undetected, in the ratio-based method. Because of similar power consumption in standby mode and non-standby mode, the ratio-based method can not obtain any existing power value obtained in each load is less than a percentage of the maximum value of the power.

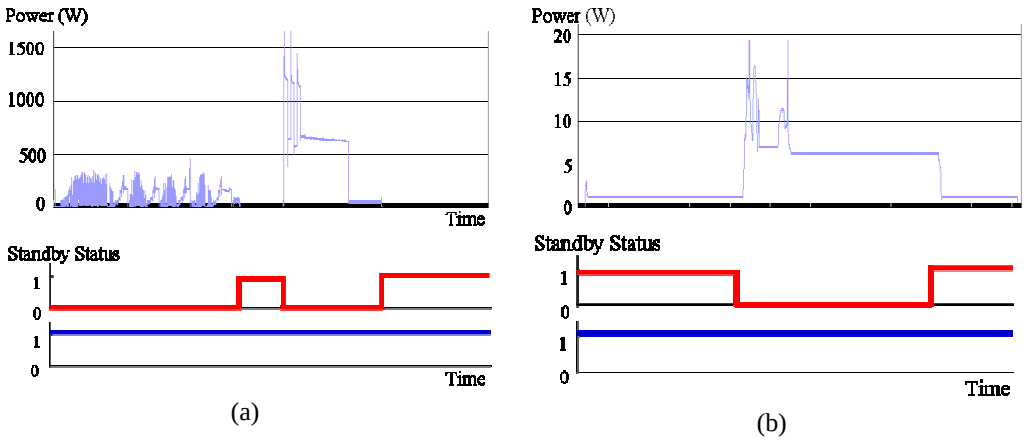


Fig. 7. The results of standby status detection of: (a) a combination all-in-one washer and dryer machine from washing to drying; (b) an inkjet printer while printing a document.

The other comparison results are shown in Fig. 7. We can found that the ratio-based method (the blue line) also can not correctly distinguish the standby status in Fig. 7(a) and (b). As previous mentioned, home appliances power consumptions would have high volatility depends on function provided. In this case, the ratio-based method can not filter high volatility power usage and a percentage from the obtained maximum value is still very high where the system can not apply to correctly detect the standby status of home appliances.

Conclusions

In this work, we design and implement a control system for standby power detection and saving. First, to enhance the detection of standby status, the proposed system presents a confidence-based adaptation method to apply detection in different power consumption status of a home appliance. Second, the proposed system also uses a confidence level verification to filter high volatility power usage. Finally, to detect similar power consumption in standby mode and non-standby mode, this work initializes standby power model training process to obtain the standby power range. The proposed method is implemented on Silicon Lab MCU F930 with a power meter ADI ADE7736 and compatible with HPCC 1.0 specification standard. The experiments show that the proposed method can be applied to different appliances and perform well on standby detection and power saving.

Acknowledgments. The authors would like to thank the Ministry of Economic Affairs (MOEA), Taiwan, R.O.C. for supporting the “Intelligent Control” science and technology project. We would also like to acknowledge the encouragement and support from the related leaders and all members of the N200 and N400 teams at Industrial Technology Research Institute South (ITRI).

References

[1] S. Siwamogsatham, P. Rattanawan, M. Kitjaroen, P. Songtung, P. Pongpaibool and K. Navanugraha: Smartly Saving Energy with a Zero Power Consumption Standby System, 11th Portland International Center for Management of Engineering and Technology (2011), p. 1-4

[2] M. Lee, Y. Uhm, Y. Kim, G. Kim and S. Park: Intelligent Power Management Device with Middleware based Living Pattern Learning for Power Reduction, IEEE Trans. Consum. Electron. Vol. 55(4) (2009), p. 2081-2089

[3] J. Heo, C.S. Hong, S.B. Kang and S.S. Jeon: Design and Implementation of Control Mechanism for Standby Power Reduction, IEEE Trans. Consum. Electron. Vol. 54(1) (2008), p. 179-185

[4] S.H. Lee and J.M. Yun: Design of Energy-Efficient Set-top Box, 15th IEEE International Symposium on Consumer Electronics (2011), p.75-78

[5] S.H. Ju, Y.H. Lim, M.S. Choi, J.M. Baek and S.Y. Lee: An Efficient Home Energy Management System based on Automatic Meter Reading, 15th IEEE International Symposium on Power Line Communications and Its Applications (2011), p.479-484

[6] Information on http://www.moeaboe.gov.tw/English/english_index.aspx

[7] J.B. Garcia, J.A.Z. Labarta, A.S. Andreu, J.G.B. Jane and M.T. Casas, U.S. Application 2010/0152916 A1. (2010)

Solution of Matrix Riccati Differential Equation of Optimal Fuzzy Controller Design for Nonlinear Singular System with Cross Term Using SIMULINK

M. Z. M. Kamali[1,a], N. Kumaresan[2,b] and Kuru Ratnavelu[2]

[1]Centre for Foundation Studies in Science, University of Malaya, Malaysia

[2]Institute of Mathematical Sciences, University of Malaya, Malaysia

[a]mzmk2000@yahoo.com, [b]drnk2008@gmail.com

Keywords: Simulink, Fuzzy differential equation, Takagi-Sugeno Fuzzy model, Matrix Riccati differential equation, cross term.

Abstract. Apart from other nontraditional approach to solve the matrix Riccati differential equation (MRDE), Simulink is applied to obtain the solution of MRDE for nonlinear singular system with cross term. Comparatively, similar and exact solution is possible to be achieved by using Simulink approach. Illustrative numerical example as well as tables are presented for comparison purpose.

Introduction

A fuzzy system comprises a form of multi-valued logic which is derived from fuzzy set theory. These fuzzy logic variables have a truth value which ranges between 0 and 1. The fuzzification block is responsible for transforming the crisp values into grades of membership for linguistic terms of fuzzy sets and then the inference mechanism where it simulates the human reasoning process by making fuzzy inference on the inputs and IF-THEN rules. Finally, the defuzzification engine where the fuzzy set obtained by the inference engine is transformed into a crisp value. The fuzzy system which consists of a singleton fuzzifier, product inference engine, center average defuzzifier and Gaussian membership function is called as the standard fuzzy system [1]. Implementing fuzzy systems in the control and modeling applications can help to obtain approximate solution [2, 3] from a system which have an incomplete information. Other than that it can be applied for system with tedious mathematical model to derive and it is very helpful for uncertain or approximate reasoning. Due to the nature of fuzzy controllers which are nonlinear, this is just perfect for it to be utilized in the control of nonlinear systems.

Singular systems have been widely used in the investigation of dynamical systems in electrical, chemical or mechanical engineering. The singular system is actually a mixture of algebraic and differential equations. Therefore, the algebraic equations represent the constraints to the solution of the differential part. The most difficult part of this system arises when there is a need for their control. The importance of solving the Riccati differential equations is vital in the optimal control theory. In 1998, Chen *et. al* [4] reported that the stochastic linear quadratic regulator problems can be well posed if solutions can be produced for the Riccati equation. Thus an optimal feedback control can be obtained. However, due to the presence of complicated nonlinear terms in the Riccati equations, this made the problems more difficult to be solved. The solution of this equation is difficult to obtain from two points of view. One is nonlinear and the other is in matrix form. Due to the development of efficient computational techniques, theorist have come out with new alternative methods to solve the MRDE, among them are the theoretical group from India led by

Balasubramaniam and co-workers [5-12]. They have solved the MRDE using the nontraditional methods such as the neural networks, genetic programming and ant colony programming. They reported that their non-traditional methods provide faster solutions compared to the traditional Runge-Kutta method.

The present paper implements the Simulink tool that runs as a companion to MATLAB software for solving the matrix Riccati fuzzy differential equation in order to get the optimal solution. This add-on package can be used to create a block of diagrams which can be translated into a system of ordinary differential equations. This paper is organized as follows: in the next section, the statement of the problem and the solution of the MRDE are presented. Later, the numerical example is discussed. Finally the conclusion section demonstrates the efficiency of the method.

Statement of the Problem

Given the singular non-linear system as

$$E\dot{x}(t) = A(x)x(t) + Bu(t), \quad x(0) = x_0 \ , \tag{1}$$

where the matrix E is a singular matrix, $x(t) \in R^n$ is a generalized state space vector and $u(t) \in R^m$ is a control variable. $A \in IR^{nxn}$, $B \in IR^{nxm}$ are known as coefficient matrices associated with $x(t)$ and $u(t)$ respectively, x_0 is given initial state vector and $m \leq n$. In order to derive the T-S fuzzy model from the nonlinear system, the first step is to determine the membership functions. This can be done by using the sector nonlinearity approach [13]. For simplicity, the matrix A(x) is taken as $A(x) = \begin{bmatrix} 0 & 1 \\ x_1(t) & x_2(t) \end{bmatrix}$ and the fuzzy variables, x_1 and x_2 are also denoted as z_1 and z_2, respectively. By calculating the maximum and minimum values of z_1 and z_2, the membership function can be obtained. Thus x_1 and x_2 can be represented for the membership functions M_1 , M_2 , N_1 and N_2 as follows

$$\begin{aligned} z_1(t) = x_1(t) = M_1(z_1(t)).max(z_1(t)) + M_2(z_1(t)).min(z_1(t)), \\ z_2(t) = x_2(t) = N_1(z_2(t)).max(z_2(t)) + N_2(z_2(t)).min(z_2(t)). \end{aligned} \tag{2}$$

Since M_1 , M_2 , N_1 and N_2 are fuzzy sets, their values can be calculated by using the following relations

$$\begin{aligned} M_1(z_1(t)) + M_2(z_1(t)) = 1, \\ N_1(z_2(t)) + N_2(z_2(t)) = 1. \end{aligned} \tag{3}$$

The membership function is named as "Small", "Big", "Positive", "Negative", respectively and from this membership functions, the nonlinear systems can be linearized into the i^{th} rule of continuous T-S fuzzy model of the the following forms. Given the singular non-linear system (Eq. 1) that can be expressed in the form of T-S fuzzy system: Model Rule i: If $z_1(t)$ is M_{i1} and $z_2(t)$ is M_{i2} ... $z_p(t)$ is M_{ip}, Then

$$E_i\dot{x}(t) = A_i(x)x(t) + B_iu(t), \quad x(0) = x_0, \text{ i} = 1, 2, 3, 4. \tag{4}$$

where M_{ij} indicates the fuzzy set rule of the fuzzy model, i is the number of model rules, $x(t) \in R^2$ is a generalized state space vector and $u(t) \in R^1$ is a control variable. $A_i \in IR^{2x2}$ and $B_i \in IR^{2x1}$ are known as coefficient matrices associated with $x(t)$ and $u(t)$ respectively, x_0 is given initial state

vector. Therefore the nonlinear system is modeled by the following fuzzy rules where the subsystems are defined as

$$A_1 = \begin{bmatrix} 0 & 1 \\ \max(z_1(t)) & \max(z_2(t)) \end{bmatrix}, A_2 = \begin{bmatrix} 0 & 1 \\ \max(z_1(t)) & \min(z_2(t)) \end{bmatrix},$$
$$A_3 = \begin{bmatrix} 0 & 1 \\ \min(z_1(t)) & \min(z_2(t)) \end{bmatrix}, A_4 = \begin{bmatrix} 0 & 1 \\ \min(z_1(t)) & \max(z_2(t)) \end{bmatrix}.$$

Now from this defuzzification process, the $E_i\dot{x}$ can be computed as as $E_i\dot{x} = \sum_{i=1}^{4} h_i(z(t))A_i x(t) + B_i u(t)$, where $h_i(z(t)) = \frac{\Pi_{j=1}^{2} M_j^i(z_j(t))}{\sum_{i=1}^{4}(\Pi_{j=1}^{2} M_j^i(z_j(t)))}$, for all *t*.

To minimize both state and control signals of the feedback control system, a quadratic performance index is minimized:

$$J = \frac{1}{2}\left(x^T(t)Qx(t) + u^T(t)Ru(t) + 2u^T(t)Hx(t)\right)dt \tag{5}$$

where the superscript *T* represents the transpose operator, $S \in lR^{2x2}$ and $Q \in lR^{2x2}$ are symmetric and positive definite (or semidefinite) weighting matrices for *x(t)*, $R \in lR^{1x1}$ is a symmetric and positive definite weighting matrix for *u(t)*. $H \in lR^{1x2}$ is a coefficient matrix.

Based on the standard procedure, *J* can be minimized by minimizing the Hamiltonian equation

$$\mathbb{H}(x(t),u(t),\lambda(t)) = \frac{1}{2}x^T Qx(t) + \frac{1}{2}u^T(t)Ru(t) + u^T(t)Hx(t) + \lambda^T(t)\left[A_i x(t) + B_i u(t)\right]$$

Using calculations of variations and Pontryagin's maximum principle, a linear state feedback control law

$$u(t) = -R^{-1}(B_i^T \lambda(t) + Hx(t))$$

can be obtained to the system in Eq. 4 and

$$\lambda(t) = K_i(t)E_i x(t),$$

where $K_i(t) \in lR^{2x2}$ is a symmetric matrix and it is the solution of the relative MRDE

$$E_i^T \dot{K}_i(t)E_i + E_i^T K_i(t)A_i + A_i^T K_i(t)E_i + Q - \left(H^T + E_i^T K_i(t)B_i\right)R^{-1}\left(H + B_i^T K_i(t)E_i\right) = 0 \tag{6}$$

for the singular system (Eq. 4).

Solution of MRDE

The MRDE is solved using Simulink in order to get the optimal solution for $K_i(t)$. In the solution matrix $K_i(t)$, since its symmetric and the system is singular, $k_{12}=k_{21}$ and k_{22} is free (let $k_{22} = 0$). After substituting the appropriate matrices in Eq. 6, the MRDE becomes the following system of nonlinear equations

$$\dot{k}_{11} = f_1(t, k_{11}, k_{12})$$
$$\dot{k}_{12} = f_2(t, k_{11}, k_{12}).$$

Procedure for simulink solution

Step 1. Choose or select the required graphical block diagrams from the simulink Library.
Step 2. Connect the appropriate blocks.
Step 3. Set up the simulation parameters and run the simulink model to obtain the solution.

Numerical Example.

Consider the optimal control problem, Eq. 5 is minimized, subject to the linear singular fuzzy system Eq. 4 where $\dot{x}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix}$,

$$A_i(x) = \begin{bmatrix} 0 & 1 \\ z_1(t) & z_2(t) \end{bmatrix}, S = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, E_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, R = 1, Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, H = \begin{bmatrix} 1 & 0 \end{bmatrix}, B_i = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

Here $z_1(t) = x_1(t)$ and $z_2(t) = x_2(t)$. For simplicity, let $x_1 \in [0.5, 3.5]$ and $x_2 \in [-1, 4]$. The minimum and maximum values of z_1 and z_2 can be calculated and by using Eq. 2 and Eq. 3, the membership functions can be obtained and shown in Fig. 1 and Fig. 2. Then, the nonlinear system is represented by the following fuzzy model.

Model Rule 1: IF $z_1(t)$ is Positive and $z_2(t)$ is Big, THEN $E\dot{x}(t) = A_1x(t) + Bu$,
Model Rule 2: IF $z_1(t)$ is Positive and $z_2(t)$ is Small, THEN $E\dot{x}(t) = A_2x(t) + Bu$,
Model Rule 3: IF $z_1(t)$ is Negative and $z_2(t)$ is Big, THEN $E\dot{x}(t) = A_3x(t) + Bu$,
Model Rule 4: IF $z_1(t)$ is Negative and $z_2(t)$ is Small, THEN $E\dot{x}(t) = A_4x(t) + Bu$,

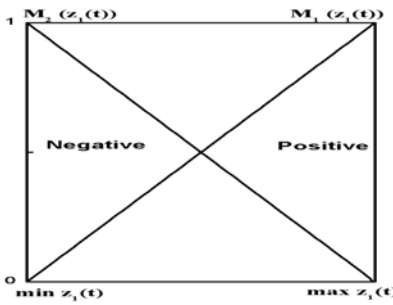


Figure 1. Membership function of $z_1(t)$

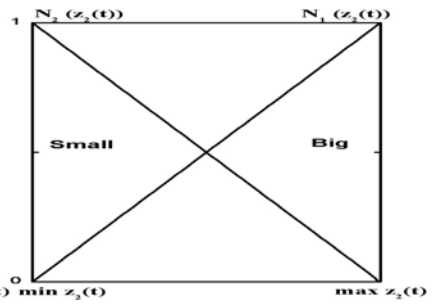


Figure 2. Membership function of $z_2(t)$

where $A_1 = \begin{bmatrix} 0 & 1 \\ 3.5 & 4 \end{bmatrix}$, $A_2 = \begin{bmatrix} 0 & 1 \\ 3.5 & -1 \end{bmatrix}$, $A_3 = \begin{bmatrix} 0 & 1 \\ 0.5 & 4 \end{bmatrix}$, $A_4 = \begin{bmatrix} 0 & 1 \\ 0.5 & -1 \end{bmatrix}$

If $z_1 = x_1 = 2.145$ and $z_2 = x_2 = 0.25$, the T-S fuzzy modelling implication can be derived as in Table 1. Using T-S fuzzy defuzzification process, the final values for $\dot{x}_1$ and $\dot{x}_2$ can be calculated as:

$$\dot{x}_1 = \frac{(0.2498 \cdot 0.25) + (0.5483 \cdot 0.25) + (0.2498 \cdot 0.25) + (0.4517 \cdot 0.25)}{(0.2498 + 0.5483 + 0.2498 + 0.4517)} = 0.25$$

$$\dot{x}_2 = \frac{(0.2498 \cdot 8.5075) + (0.5483 \cdot 7.2575) + (0.2498 \cdot 2.0725) + (0.4517 \cdot 0.8225)}{(0.2498 + 0.5483 + 0.2498 + 0.4517)} = 4.6639$$

Table 1. T-S Fuzzy Model Implication.

Implication	Premise	Consequence	Truth Value
Rule 1	$M_1(z_1)=0.5483$ $N_1(z_2)=0.2498$	$\dot{x}_1=0.25$ $\dot{x}_2=8.5075$	$0.5483 \wedge 0.2498=0.2498$
Rule 2	$M_1(z_1)=0.5483$ $N_2(z_2)=0.7502$	$\dot{x}_1=0.25$ $\dot{x}_2=7.2575$	$0.5483 \wedge 0.7502=0.5483$
Rule 3	$M_2(z_1)=0.4517$ $N_1(z_2)= 0.2498$	$\dot{x}_1=0.25$ $\dot{x}_2=2.0725$	$0.4517 \wedge 0.2498=0.2498$
Rule 4	$M_1(z_1)= 0.4517$ $N_1(z_2)= 0.7502$	$\dot{x}_1=0.25$ $\dot{x}_2=0.8225$	$0.4517 \wedge 0.7502=0.4517$

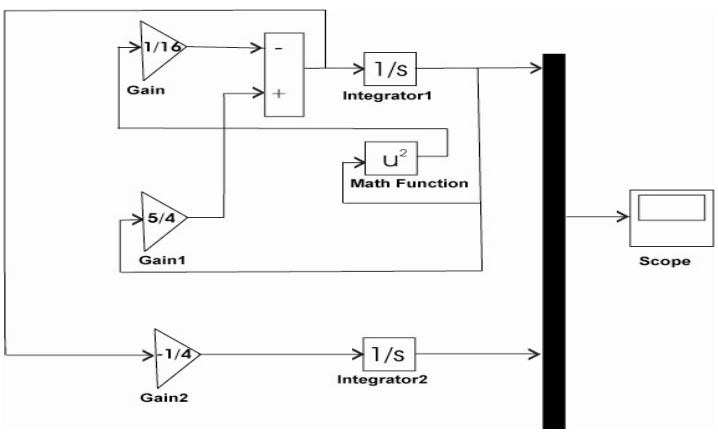


Figure 3. Simulink Model.

The results of $\dot{x}_1$ and $\dot{x}_2$ from the T-S fuzzy approximation are either close or similar to the original system in the Eq. 4. (i.e $\dot{x}_1$= 0.2490 and $\dot{x}_2$= 4.6630). The simulink model shown in Fig. 3 represents the systems of differential equations in k_{11} and k_{12}, with the terminal conditions $k_{11}(2)$ = 1.0 and $k_{12}(2)$ = 0.0 . These numerical solutions of MRDE are shown in Table 2. Similarly the solution of the above system with the matrix A_2, A_3 and A_4 can be solved using simulink.

Table 2. Solutions of MRDE.

	Exact		Simulink	
t	$k_{11}(t)$	$k_{12}(t)$	$k_{11}(t)$	$k_{12}(t)$
0.00	0.0860	0.2285	0.0860	0.2285
0.20	0.1103	0.2224	0.1103	0.2224
0.40	0.1415	0.2146	0.1415	0.2146
0.60	0.1813	0.2047	0.1813	0.2047
0.80	0.2322	0.1920	0.2321	0.1920
1.00	0.2971	0.1757	0.2971	0.1757
1.20	0.3799	0.1550	0.3799	0.1550
1.40	0.4852	0.1287	0.4852	0.1287
1.60	0.6187	0.0953	0.6187	0.0953
1.80	0.7875	0.0531	0.7875	0.0531
2.00	1.0000	0.0000	1.0000	0.0000

Conclusion

The solution of MRDE of optimal fuzzy controller design for nonlinear singular system with cross term has been obtained by using Simulink . The results are similar to the exact solution. A numerical example is given to illustrate the proposed method. The computational work of the optimal solutions are done in Matlab on PC, CPU 2.0GHz.

Acknowledgement

NK and KR would like to acknowledge the funding of this project by the UMRG grant (Account No: RG099/10AFR)

References

[1] L. X. Wang: IEEE Trans. Fuzzy Syst. Vol. 6 (1) (1998), p.137.

[2] L. A. Zadeh: Inform. Sciences, Part I: Vol. 8 (1975), p.199; Part II: Vol. 8 (1975), p. 301; Part III: Vol. 9 (1975), p.43.

[3] L. A. Zadeh: Inform. Sciences Vol. 178 (2008), p. 2751.

[4] S. P. Chen, X. J. Li and X. Y. Zho, SIAM J. Control Optim. Vol. 36 (5) (1998), p.1685.

[5] P. Balasubramaniam and A. Vincent Antony Kumar: Genet. Program. Evolvable Mach. Vol. 10 (2009), p. 71.

[6] P. Balasubramaniam, J. Abdul Samath, N. Kumaresan and A. Vincent Antony Kumar: Appl. Math. Comput. Vo.l 182 (2006), p. 1832.

[7] N. Kumaresan: Neural Comput. Appl. (2011),(in press).

[8] P. Balasubramaniam, J. Abdul Samath and N. Kumaresan: Appl. Math. Comput. Vol. 187 (2007), p. 1535.

[9] P. Balasubramaniam, J. A. Samath, N. Kumaresan and A. V. A. Kumar: Neural Parallel Sci. Comput. Vol. 15 (2007), p. 125.

[10] P. Balasubramaniam and N. Kumaresan: Appl. Math. Comput. Vol. 204 (2) (2008), p. 671.

[11] N. Kumaresan: Appl. Math. Model. Vol. 35 (2011), p. 3797.

[12] N. Kumaresan and P. Balasubramaniam, Int. J. Comput. Math. Vol. 87(14) (2010), p. 3311.

[13] S. Kawamoto, K. Tada, A. Ishigame and T. Taniguchi: IEEE Trans. Neural Netw. Vol. 4 (1993), p. 919.

Conflict Detection in Autonomic Systems Using Petri Networks

Siddhartha Moraes Amaral de Freitas[1, a], Catalin Meirosu[2,b] andDjamelFawziHadjSadok[1,c]

[1]Federal University of Pernambuco (UFPE)

Research Group of Computer Networks and Telecommunication (GPRT)

Recife, Pernambuco,Brazil

[2]Ericsson Research

Stockholm, Sweden

[a]sid@gprt.ufpe.br, [b] catalin.meirosu@ericsson.com, [c]jamel@gprt.ufpe.br

Keywords: Conflict detection, autonomic systems, Petri networks.

Abstract. The increase in the complexity in computational environments has promoted Autonomic Computing, where systems are able to perform tasks without human interference. But this technological innovation can lead to the emergence of problems related to actions taken by its elements. In this scenario we can perceive the need to identify and avoid conflicting actions that prejudice the entire system.

Introduction

The use of traffic engineering is essential due to the growth of the numbers of users and to the number of available services on the network. This scenario forces service providers to efficiently manage their resources with the goal to maintain a high Quality of Service (QoS). Furthermore, network systems are multi-technology and hardware is from multi-vendors; this makes the management task a complex one, in which success is totally dependent on administrator expertise.

With Autonomic Computing, possible solutions to management problems in these scenarios have been seen.The management of autonomic networks has four main features, called self-configuration, self-healing, self-optimization, and self-protection. With these characteristics, the system can perform its operations in an autonomic way, achieving the set of previous established goals created by administrators [1]. In this way, the self-management system tries to anticipate, diagnose, and isolate any network anomaly that can introduce the disruption of services.

Although Autonomic Computing brings several advantages, it is common that conflicts arise and cause distortions in the environment, making it sometimes inoperative. Since each entity behaves independently, the self-optimization of just one could cause a profitable effect in just a part of the system, instead of all. In this case, the identification of operations that can generate environment inconsistency should be realized and put into practice with the goal to avoid them. This paper describes the use of autonomic computing in an MPLS network and the potential conflicts that can appear due to this optimization, and presents a schema of conflict detection based on a Petri networks extension.

Autonomic Management

Autonomic Computing is a concept introduced by IBM [2] and is related to the use of computational elements with capabilities to perform tasks without human interference, in an optimized way, and

answering to external stimulus in an effective manner. These tasks are called self-x functionalities that comprehend: self-optimization, self-healing, self-protection, and self-configuration.

An autonomic system is based on an iterations loop, comprehending the environment monitoring, the data analyzing, the optimization possibilities planning, and the decision execution, as shown in Fig. 1.The Autonomic Manager (AM) is responsible for the management of autonomic entities.

A management network system is called autonomic if it implements the self-x functionalities. The network elements are entities that are always monitored and managed, and their behaviors are modified due to optimizations done by the AM. Therefore, a management system that is able to perform the self-configuration function should be also able to detect anomalies that can violate restrictions of the environment. To do this, the system should configure accordingly the active elements, adapting them to new network conditions without causing an impact on general performance. On the other hand, the self-optimization capability will guarantee that any change in the network status will be done with the goal of achieving some improvement in the current state, and preventing any non-beneficial changes from occurring.

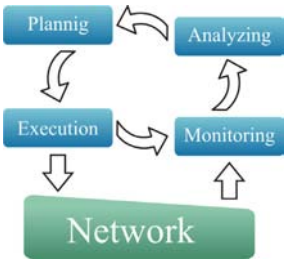


Fig. 1. Autonomic management cycle

Requirements of an autonomic management system. An autonomic system requires a knowledge base with information about the target monitored environment. This base is distinct in two domains [3]: one that has information related to knowledge of network elements, their capabilities, their types, their limitations, and the network topology; the other that has information related to acquired knowledge from last actions, that serves as a base for the decision making process in future operations done by AM.

The feeding of these knowledge bases can be done by an AM request, in a centralized model, in a timing way, or due to some modification in the network, such as a load increase. Furthermore, it can be done by the managed elements, in a distributed fashion, feeding the base due to some trigger occurrence, after some action accomplishment, or according to some time interval pre-established.

Conflicts. The organization of distinct entities which try to accomplish their self-optimization task without considering the behavior of their neighbors may lead to the occurrence of conflicts that will degrade the performance of the entire environment. This problem can reach up to a non-operational state.

The conflict occurrence happens when two or more entities try to make divergent optimization operations, each one acting for its own benefit. The severity of the occurrence of these conflicts is enhanced depending on the environment size, since the spreading of an erroneous action might cause a snowball effect that will influence all networkelements. In this way, the AMshould identify which operations can be harmful to other elements, allowing that only non-prejudicial ones are conducted. In this way, it is necessary to have a mechanism that is able to identify the conflicts'possibilities; this mechanism will proactively prevent wrongful actions from being taken by elements.

Conflicts identification in an autonomic management system using Petri networks

The Petri networks are graphic and can be used as mathematical modeling toolsto describe and

analyze systems. The idea was initially developed by Carl Adam Petri in 1962 [4] and allows that environment peculiarities of distributed systems, such as concurrence, synchronization, and non-deterministic tasks are easily represented [5].

Initially, Petri networks applications were focused on communication protocols [6]. However, due to theircharacteristics and functionalities, Petrinetworks are used in several kinds of systems, such as electronic commercial systems [7], ERP [8], supply chain [10], auto-organization systems (SON) [9], and so on.

Petri networks have been used as a powerful tool for detection and analysis of conflicts and deadlocks in distributed systems [10]. Beyond this, they have been used by many authors, such as Zeng et al. [11] who used Petri networks for conflict detection in automatically guided vehicles systems; and Jennifer Blackhurst et al. [10] who elaborated a methodology for conflict detection in a supply chain using Petri networks with matrix equations.

According to Jennifer Blackhurst, the conflicts existent in a supply chain are derived when individual entities, each optimized for their own iterations, are combined into a single system, such as in anautonomic environment, and several elements of a same area realize their self-optimization, leading the system to an inadequate state.

Methodology description. The modeling methodology used and described in this paper is based on work developed in [10]. Initially, possible actions realized by entities of the autonomic system are modeled as a Petri network. Due to the high level abstraction of this model, the creation of these models is not a complex task; however, at the same time, it is necessary to pay attention so that the representation model is reliable to the target environment.

After individual tasks modeling, the composition of these tasks is done. This composition should be realized as faithfully as possible to the real system model, since it is through the composition of individual models that the conflicts will be identified.

According to [10], each environment composition is represented by the following structure: (*P*, *T*, *I*, *O*, *M*), where each element is an element type, or a matrix of possible states or system characteristics.

The *P* element represents the system states, which are identified by the circle in the Petri network. The *T*element is the transition node and represents the realization of a possible action and seems like a rectangle in the graphic. The *I*element is the input matrix and represents the arcs' set from *P* to *T*. The *O*element is the output matrix and represents the arcs' set from *T* to *P*. Finally,the *M* element is the location and the number of tokens at the initial instant of the network. The Fig. 2 shows the composition with operator AND and its representation as (*P*, *T*, *I*, *O*, *M*) structure.

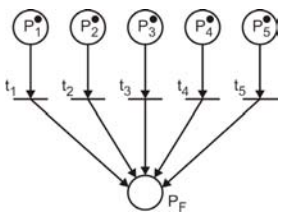


Fig 2. Operator OR usage(adapted from [10])

The matrices that represent the model are illustrated in Fig. 3:

$$I = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_F \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \quad O = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_F \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Fig 3. Matrixes *I* and *O*

The conflicts' identification check if the result generated by the modeling environment is according to an ideal result pre-established by the system creator.

Using the environment of Fig. 2 as example and supposing that the final result is M_{target} = (0,0,0,0,4), we apply the Eq. 1:

$$M_{target} = M' + x^{*}\mathrm{C}. \quad (1)$$

$$C = O - I \quad (2)$$

where*C* is derived from the matrix generated by Eq. 2, with the following result (Fig. 4):

$$C = \begin{array}{c} \\ T_1 \\ T_2 \\ T_3 \\ T_4 \\ T_5 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_F \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array}$$

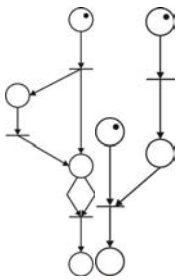
Fig4. Matrix C

The variable *x* is represented by an array of non-negative integers with lengths equal to the number of lines of matrix *C*, given by *x* = [1 1111]. Taking elements that composeEq. 1, we have the result (0,0,0,0,4), which is exactly the expected final value. With this, we can infer that the modeling environment is free of conflicts.

The occurrence of one conflict in this approach happens when the result of Eq. 1 is different from the expected final value, given by M_{target}. The solution of the conflict is done by remodeling or by evaluation of the operator used in the composition of individual elements. Depending on the represented model, the array *x*can have values relative to the weight given by a specific operation. In this way, the modification of these weights can also influence in the conflict resolution.

Using conflict detection methodology in an autonomic management network environment. As described previously, an autonomic management network system will collect environment information in an iteration loop. At each cycle, the collected data will be analyzed and if there are entities that reach some threshold, they will be used as input to the conflict management system, since it is feasible for themto stimulate some trigger.This conflict management system should create a composition of Petri networks models for each possible action and all models of actions in Petri networks are part of the knowledge base of the network.

The scenario used to prove the methodology comprehends just two tasks: the first is related to power saving, and the second is related to load balancing. When a node remains a long time without executing tasks, it should come in a non operational state, according to some policy previously defined to achieve a power saving goal. On the other hand, it is important to take into account the high utilization of an active network element, redistributing its load among other nodes to obtain load balancing. Fig.5 illustrates the model of these two actions.



(a) (b)

Fig 5. Models of two tasks: (a) energy saving and (b) load balancing

Fig.5(a) refers to energy saving action. After a node identifies its low utilization, it requests the redirection of its demands to other nodes and after this it executes the standby-mode action. Fig. 5(b) is related to load balancing action. A node requests the distribution of its load to the AM, which will identify which nodes can help it in its load redistribution.

With models, the conflict detection system will make the composition of both. In our case, the conflict will occur if the element selected by the load balancing algorithm is the same element that wants to make its energy saving.

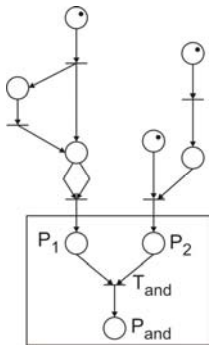


Fig 6. Composition of models

Fig.6 shows the composition generated by the union of two actions. The equation of conflict detection is applied in specific parts of the modeling. In the case of Fig. 6, it will be applied only at places *P1*, *P2*, and *P.and*, and at transaction *T.and*. In this way, the expected$M_{target}$of this model requires that, at minimum, the token at the last position is 0. Next, elements that compose the equation of conflict identification will be described as follows.

```
P = [P1, P2, P.and]
T = [T.and]
M0 = [1 4 0]
x = [1]
Mtarget = (x,x,0)
I = [1 1 0]
O = [0 0 1]
C = [-1 -1 1]
Mtarget = [1 4 0] + [1] * [-1 -1 1] = [0 -3 1]
```

According to the applied methodology, the conflict presence is identified. The solution is based on the remodeling, modifying operations or selecting one of them based on some priority criteria with the goal that the system doesn't reach an inoperative state.

The use of a Petri network extension with numeric elements instead of tokens makes the methodology more clear, due to the possibility of operators checking if a specific element can pass or not by a transition. In this way, tokens in the network can be represented by nodes that implement an operation. The occurrence of common elements in both networks should be analyzed by the *guard* operator that is responsible for testing whether an operation is valid. With this operator at the *T.and* transition, the element equality occurrence allows the token transition to *P.and*, confirming the conflict existence.

Conclusions and future works

This paper had as a goal to present the autonomic management system elements and their functions. Arguments that indicate the possibility of conflict existence in autonomic systems were presented, taking into consideration the application of conflict detection methodologies used in other areas. The technique applied used matrix equations with Petri networks, showing that some actions can be used

together, since they are not related to common elements. As the performance of this detection scheme operates proactively, theconflicting tasks' execution will be avoided, preventing the system from collapsing.

As a result of writing this paper, questions were raised, pointing to directions for the study of new techniques for detection of long-term conflict. In this way, we can highlight the need to create a tool that emulates the environment of an autonomic system. As a future research, we will apply the methodology presented and will be able to perform analysis for the identification of possible future conflicts.

Acknowledgements

This work has been supported by Ericsson Research in Brazil (EDB).

References

[1] Samaan, N. andKarmouch, A.: Towardsautonomicnetworkmanagement: an analysis of current and future research directions, vol. 11 (2009), p. 22

[2] J. Kephartand D. Chess.: The Vision ofAutonomic Computing, IEEE Computer, vol. 36 (2003), p. 41

[3] Mark A. Musen.:Automatedgenerationof model-basedknowledgeacquisitiontools. Morgan Kaufmann, San Francisco, (1989).

[4] Petri C. Kommunikation mit autimaten.:PhDdissertation, University of Bonn, West Germany, (1962).

[5] K. Jensen.:Coloured Petri Nets: Basic Concepts, Analysis MethodsandPracticalUse, Springer-Verlag, (1992)

[6] Zurawski R. Zhou M.: Petri Netsandindustrialapplications: a tutorial. IEEE Trasactions on Industrial Electronics, vol 41 (1994), p. 567

[7] Xu H. ShatzS.:Anagent-Based Petri Net modelwithapplicationtoseller/buyer design in electronic commerce, In: 50th International symposium on autonomousdecentralizedsystems(2001), p. 11

[8] Kwon O, Lee J.: A multi-agent intelligent systemforefficient ERP maintenance. Experts Systems withApplication, vol. 21 (2001),p. 191

[9] DianxiangXu, PritiBorse, Karl Altenburg, and Kendall Nygard.: A Petri netsimulatorforself-organizingsystems. 5th WSEAS International Conference on ArtificialIntelligence, Knowledge Engineering and Data Bases, USA (2006), p. 31

[10] Jennifer Blackhurst, Tong (Teresa) Wu, Christopher W. Craighead.: A systematicapproachforsupplychainconflictdetectionwith a hierarchical Petri Net extension, Omega, vol 36 (2008), p. 680

[11] Zeng L, Wang H, Jin S.: Conflictdetectionofautomatedguidedvehicles: a Petri netapproach. InternationalJournalofProduction Research, vol 29 (1991), p. 865

Personalized Predictive Model for Mobile Value Added Services

Meera Narvekar [1],Dr.S.S Mantha [2]

[1] DJ Sanghvi College of Engineering.Mumbai,INDIA

[2] SNDT University ,Mumbai , INDIA

narvekar.meera@gmail.com, ssmantha@vjti.org.in

Keywords - Mobile Value Added Services, Relative Ranking Algorithm, personalization, prioritization

Abstract: This paper provides an insight into personalized mobile value added services (MVAS). The paper describes a model to personalize user information related to various services offered to the user on a mobile phone. The model is self-learning model that is trained to anticipate the users need and preferences. A user profile for every subscriber is maintained which is derived from user query logs in the prioritized and summarized form. This user profile is updated at regular predefined intervals. The user will be able to receive alerts or information of his choice with respect to spatial and temporal behaviour exhibited.

Introduction

Mobile VAS has come a long way from simple SMS-based services to multimedia-rich content. With advances in handsets, cellular networks and Web technologies, users can access an almost infinite array of applications from Games, Ringtones, and Ecommerce tools to personalized user alerts. MVAS is fast becoming a major part of the revenue generation for telecom and media companies around the world. Mobile data is now typically between 25% and 40% of revenues in mature markets and by far the fastest growing service [16] This serves to emphasize that when the experience is right, mobile users are looking for ways to add value to their device – and are in many cases prepared to pay to get what they want.

In an era where network service providers of mobile phones go to ends to maintain their clientele, MVAS play a vital role in ensuring consumer loyalty. The model implemented today for VAS lacks personalization and fails to deliver required and appropriate information customized to the user's requirement. As well as driving revenues directly, value added services can enhance performance in other ways. Critically, in a competitive market, they help to build stronger relationships with customers, as customer data can be mined to create customized services that increase loyalty and stickiness. Experience from other markets demonstrates that VAS can help with:

- Revenue Growth
- Customer Acquisition
- Customer Retention

The paper proposes a model that strives to personalize MVAS for each subscribed consumer. The model is an evolutionary learning model that employs a rank algorithm to personalize and prioritize the services most likely to be useful for the consumer.

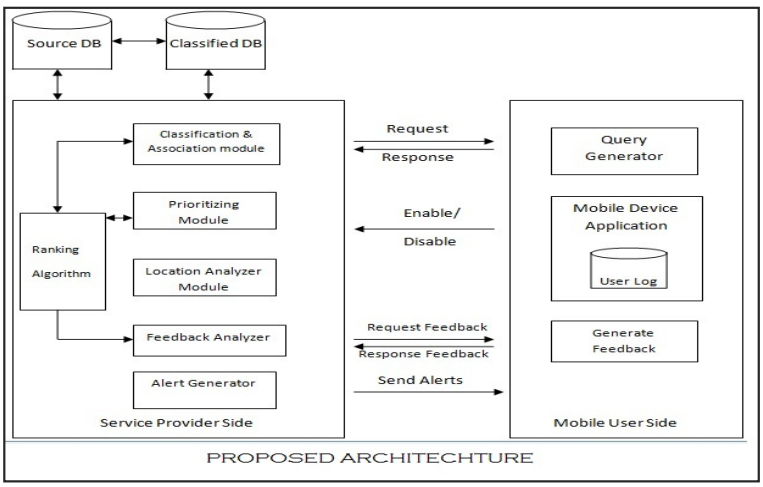
An initial survey of over 350 participants in Mumbai, India was conducted to create the initial training dataset. The survey acquires various user details like mobile brand, monthly income, age, gender etc to build a user profile. This data is then classified by Age, Gender and Socio-economic background using the Naïve-Bayes classifier algorithm to create an initial Learning model predictor. Based on this model, new network subscribers will be suggested about the likelihood of selecting a

given MVAS. The users in addition can subscribe to specific MVAS, thus preserving the individuality of users.

System Architecture

A detailed centralized source database holds the data on the subscribed users and their respective profiles. This database is maintained at the server end of the service provider. The user profiles are subjected to modules of prioritization, followed by summarization, thereby creating a summarized database. This summarized database containing the most likely useful information of a particular user will be sent to the user when he is either weakly connected or will be possibly disconnected. This summarized database can be located at the Visitor Location Register thereby enhancing the speed of requests/ alerts.

A Relative Ranking Algorithm is designed to predict and prioritize the MVAS most relevant to the subscriber. The algorithm 'learns' by assessing the user's location, logs, profile and feedback over a period of time, thereby trying to take the profile to be near accurate. An application at the user's side allows the users to launch queries for MVAS, which are logged as a part of the user's profile.



The model aims at personalizing preferential MVAS for each consumer subscribed to this service. The model uses a rank algorithm to rank various services with respect to the consumer. The initial training dataset is based on a survey conducted. The survey was local to the Mumbai city, India. The participants involved belonged to a wide range of age groups and socio-economic background.

Using this survey influential parameters for the generation of user profile were determined. These parameters were ranked according to their importance. This dataset forms the basis for classification and association rules of the model. It serves as a yardstick and also as an initial datasets for new subscribers.

The following steps give an idea as to how the system works:

Step 1: Once the training dataset is ready, it is stored as the source database. The datasets are classified from the source datasets. The Naïve-Bayesian classification technique is used to classify the source database. The classifications are based on age groups, gender, socio economic patterns and profession. Each service is assigned a specific priority based on the factor it is being classified. As and when new subscribers are added, this information serves as a guideline to the preferences of the consumer. This can be evaluated based on the basic consumer profile.

Naïve –Bayesian Classification which is based on Bayes rule of conditional probability assumes that contributions by all attributes are independent and that each attribute contributes equally to the classification problem, unlike ID3 algorithm. By analyzing the contribution of each 'independent'

attribute a conditional probability is determined. Our classification is made by combining the impact that different attribute such as gender, age, socio-economic background and profession have on the prediction to be made. Based on various condition probability values for each MVAS with respect to our 4 defining attributes (i.e. - gender, age, socio-economic background and profession) we can prioritize and then summarize our initial training data for further prediction. Advantages of Naïve Bayes technique are:

- It is very easy to use.
- Only one scan of the training data is required.
- It can easily handle missing values.

The logs of user downloads ,his explicit requests for services , parameters in his profile (i.e. fills up a registration form for subscribing to the services of a service provider) are used to create a baseline profile .Over a period of time the priority of favorite service is evaluated using a ranking algorithm. The algorithm by itself learns and evolves; thus making the system intelligent.

Step 2: The user who is now mobile ;on subscription of the link provided by the service provider will activate the link if wants alerts or needs to query or download the services.

Step 3: A regular feedback which is a unique feature of this model is taken from the user so as to gauge the system for accuracy and its efficacy. Based on the user's feedback various parameters of the system are revisited. The rank algorithm is a weight based algorithm. It gives more weight to user's response than to the data evaluated after classification and association. Thus the rank algorithm is exploited to consider the uniqueness of each individual's taste and preferences. The feedback alerts should be well timed and managed so as to be of no nuisance to the user.

Limitations and Assumptions:

The survey is limited to city of Mumbai which exhibits good network connectivity most of the time. We assume that the users are equipped with relatively sophisticated phones. This model is unique because it takes into account the feedback of the user. It also takes into account the temporal characteristics the user exhibits.

Conclusions

Mobile Value Added Services form the modern day solution for Service Providers to maintain their current clientele and also attract potential customers by enhanced customer satisfaction. Apart from this, they provide Service Providers a constant secondary income source. Since MVAS predictive model can be easily incorporated into the current Service provider's network infrastructure, it does not require substantial funds to be invested by the Service Providers to make this ideology a reality.

References

[1] Wei-Shinn Ku, Roger Zimmermann, Haixun Wang "Location Based Spatial query processing in wireless broadcast environments", IEEE Trans. Mobile Computing VOL. 7, NO. 6, JUNE 2008.

[2] Sergio Ilarri, Eduardo Mena, and Arantza Illarramendi, "Location dependent queries in mobile contexts: Distributing processing using mobile agents", IEEE Trans. Mobile Computing VOL. 5, NO. 8, AUGUST 2006.

[3] Weng Suxiang, Jin Yongsheng, "The Demonstration Research of Mobile Value-added Services in Undergraduate", CCDC 2009.

[4] Patrick Ngok, Zhiguo Gong, "Log Mining to Support Web Query Expansions", ICIA 2009.

[5] Kyriakos Mouratidis, Spiridon Bakiras, Dimitris Papadias,"Continuous Monitoring of Spatial Queries in Wireless Broadcast Environments.", IEEE Trans. Mobile Computing, VOL. 8, NO. 10, OCTOBER 2009.

[6] Darin Chan and John F. Roddick," Context-Sensitive Mobile Database Summarisation." ACSC '03 Proceedings of the 26th Australasian computer science conference - Volume 16

[7] Kahkashan Tabassum, Maniza Hijab, A. Damodaram, " Location dependent Query Processing – Issues, Challenges and Applications" ICCNT '10 Proceedings of the 2010 Second International Conference on Computer and Network Technology

[8] Panos K. Chrysanthis, Susan Lauzac,"Personalizing Information Gathering for Mobile Database Clients." In Proceedings of the 2002 ACM symposium on Applied computing.

[9] Jahangir Dadkhah Chimeh, "Mobile Services: Trends and Evolution", ICACT Feb 2009.

[10] S.Mitra, S.Das, " Query Processing in a Cellular Network - A Database Approach", VTC'01.

[11] Chon, H. D., Agrawal, D., and Abbadi, NAPA:Nearest available parking lot application, 18th IEEE International Conf. on Data Engineering (ICDE'02), USA, 496–497.

[12] G.R. Hjaltason and H. Samet, "Distance Browsing in Spatial Databases," ACM Trans. Database Systems, vol. 24, no. 2,pp. 265-318, 1999.

[13] Y. Cai, K.A. Hua, and G. Cao, Processing Range-Monitoring Queries on Heterogeneous Mobile Objects," Proc. IEEE Int'l Conf. Mobile Data Management (MDM '04), pp. 27-38, July 2004.

[14] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques,Morgan Kaufamn, 2001

[15] H. Hu, J. Xu, W. Sing Wong, B. Zheng, D. Lun Lee, and W.-C. Lee, "Proactive Caching for Spatial Queries in Mobile Environments", Proc. 21st IEEE Int'l Conf. Data Eng. (ICDE '05), pp. 403-414, 2005.

[16]http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture_Research_Mobile_Value_Add_India.pdf

[17] Data Mining, Introductory and Advanced Topics by Margaret H Dunham, Pearson education.

[18]Nilton Bila, Jin Cao, Robert Dinoff, Tin Kam Ho, Richard Hull, Bharat Kumar and Paulo Santos, "Mobile User Profile Acquisition Through Network"

A Dynamic Workflow Model Based on Petri Net and Instance Migration

Huifang Li [1, a], Ming Zhang[2, b]

School of Automation, Beijing Institute of Technology (BIT),

Haidian District, Beijing, 100081, P. R. China

[a]huifang@bit.edu.cn, [b]ysuzhm0313@163.com

Keywords: Workflow; Dynamic change; Petri net; Reachability; Migration;

Abstract. To improve the dynamics of workflow systems, this paper researches a dynamic workflow model and an instance migration method. Firstly, we use workflow net to describe the dynamic changes for workflow models, such as inserting or deleting activity nodes. Each dynamic change can be described by a set of operating primitives, and WFMS can work out what changes have happened to the process model by accessing these operating primitives. Secondly, this paper proposes a simple method for calculating incidence matrix to verify the reachability of corresponding workflow net. We also give an instance migrating model for the running instances transferring from old model to the new one, and preserving the usability of all work finished.

Introduction

Workflow is a computerized model of business process. With the technology widely used in practical work, it is noticed that the existed workflow management systems lack the support for business with dynamic changes. Once a business process is determined, all of the tasks will be executed step by step according to the predefined workflow model. But with the continuous changes in business environment, uncertainty and variability have become intrinsic features of modern business processes. To improve the workflow systems and make it suitable for the dynamic changes has become a hot topic in workflow field.

Van der Aalst [1] proposed an automatic calculation of change region to ensure that the instance outside of the change region can be transferred correctly. Casati et al. [2] presented a systematic approach to tackle the dynamic workflow evolution and instance migration. Van der Aalst [3] proposed an advanced inheritance concept to support flexibility, but it is only suitable for circumstances where the modified model is just a subclass of the old one.

It is necessary to check the correctness of a modified workflow model when some dynamic changes have been done. The approaches for dynamic change analysis are categorized into graph-based and trace-based approaches [4, 5]. Graph-based approaches primarily focus on comparing the workflow structure fore and after. Trace-based approaches primarily emphasize on execution of workflow instances.

The remainder of this paper is organized as follows. Section 2 introduces basic concepts and techniques. Dynamic changes of workflow process model such as adding and deleting nodes, structural change of process are described formally based on Petri net in Section 3, and a set of standardized operating primitives to describe the common dynamic changes based on WF-net are proposed. Section 4 presents a simplified method for calculating incidence matrix to verify the correctness of a modified model. An approach to determine how to transfer the running instance from old model to the new one is proposed in accordance with the change region of the model and the state of the instance in Section 5, while some conclusions and future researches of this work are summarized in Section 6.

978-1-61275-004-0/10/$25.00 ©2012 IERI

Basic Definition

Petri net is an effective graphical and mathematical modeling tool for a variety of systems. It provides several powerful means to describe and analyze complex systems with parallel, asynchronous, distributed and random characteristics. It is a directed graph including places, transitions and the directed arcs between them. A formal definition of Petri net is as follows:

Definition 1 [6] A Petri net is a 4-tuple $PM = (P, T, F, M_0)$, where

(1) $P = \{p_1, p_2, \cdots\}$ is a finite set of places;

(2) $T = \{t_1, t_2, \cdots\}$ is a finite set of transitions;

(3) $F \subseteq (P \times T) \bigcup (T \times P)$ is a set of arcs;

(4) $M_0 : P \to \{0, 1, 2, \cdots\}$ is the initial state.

Let $\bullet t$ denote the set of input places for a transition t. The notations $t\bullet$, $\bullet p$ and $p\bullet$ have similar meanings, e.g., $p\bullet$ is the set of transitions sharing p as an input place.

Workflow net (WF-net) was introduced by Van der Aalst [7]. Tasks are represented by transitions because they are the active elements of Petri nets, conditions are represented by places because they are the passive element of Petri net, and the arcs between places and transitions are used to specify the partial ordering of tasks.

Definition 2 A Petri net $PN = (P, T, F)$ is a Workflow net (WF-net) if and only if:

(1) There is one source place $i \in P$ such that $\bullet i = \varnothing$;

(2) There is one sink place $o \in P$ such that $o\bullet = \varnothing$;

(3) Every node $x \in P \bigcup T$ is on a path from i to o.

Dynamic Changes of a WF-net

In this paper, we emphasize on the influence of the control-flow resulting from changes. In addition, the changes can also lead to resources or data conflict, but this problem is not the scope of this paper.

Classification of Dynamic Changes. Workflow Management Coalition (WfMC) defined four basic routing structures: sequential, parallel, choice, iterative routing. They are expressed in Fig. 1 from (a) to (e).

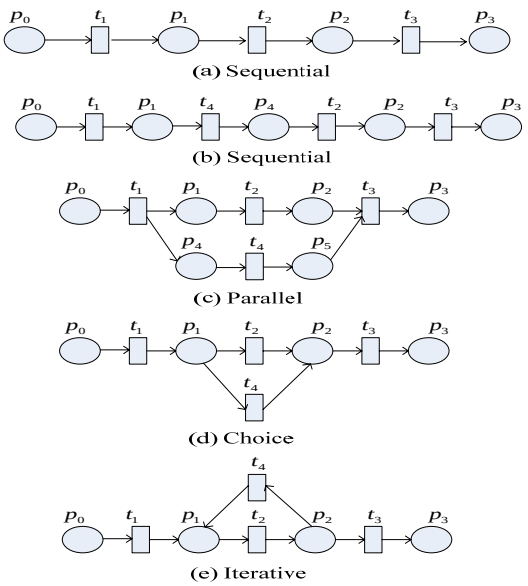


Fig. 1 Four routing structures of workflow based on WF-net

Any business process involved in enterprises can be expressed by a WF-net using a combination of these four routing structures. The actual business process changes reflected in the workflow process model are operations for the model nodes, which include:

(1) Add/delete nodes of sequential routing ((a) to (b), (b) to (a));

(2) Add/delete parallel branch ((a) to (c), (c) to (a));

(3) Add/delete choice branch ((a) to (d), (d) to (a));

(4) Add/delete iterative branch ((a) to (e), (e) to (a));

(5) Transform each other among the four routing structures.

For example, the change form (a) to (c) can be explained as adding a new task t_4 paralleled with t_2. Only when both t_2 and t_4 has completed, the task t_3 can start. Another change from (b) to (c) can be regarded as follows: making t_2 and t_4 executed concurrent to reduce total process execution time and improve process efficiency. These changes are common in actual workflow systems.

Standardized Operating Primitives for Dynamic Changes. Here, we define a set of standardized operating primitives to describe the common dynamic changes based on WF-net above-mentioned. The workflow management system can acquire what dynamic changes have happened when it accesses these primitives.

(1) Add nodes for sequential routing: $AddS(p_a, t_b, [t_n, p_n, \cdots, t_{n+m}, p_{n+m}, \cdots])$

This operating primitive can be explained as: t_n, p_n etc. nodes are added orderly between p_a and t_b. This operation is equivalent to the dynamic changes that adding several tasks $t_n, t_{n+1}, \cdots, t_{n+m}, \cdots$ before task t_n. 'S' is the first alphabet of sequential. For example, the dynamic change from model (a) to (b) in Fig. 1 can be described as $AddS(p_1, t_2, [t_4, p_4])$ using operating primitive above-mentioned.

(2) Add parallel branch (node): $AddP(t_a, t_b, [p_n, t_{n+1}, p_{n+1}, \cdots, t_{n+m}, p_{n+m}, \cdots])$

This operating primitive is to add a new branch parallel with the original branch from t_a to t_b. The new branch includes nodes $p_n, t_{n+1}, p_{n+1}, \cdots, t_{n+m}, p_{n+m}, \cdots$. 'P' is the first alphabet of parallel. For example, the dynamic change from model (a) to (c) in Fig. 1 can be described as $AddP(t_1, t_3, [p_4, t_4, p_5])$.

(3) Add choice branch (node): $AddC(p_a, p_b, [t_n, p_{n+1}, t_{n+1}, \cdots, p_{n+m}, t_{n+m} \cdots])$

This operating primitive is to add a new choice branch between p_a and p_b. 'C' is the first alphabet of choice. For example, the dynamic change from model (a) to (d) in Fig. 1 can be described as $addC(p_1, p_2, [t_4])$.

(4) Add iterative branch (node): $AddI(p_a, p_b, [t_n, p_{n+1}, t_{n+1}, \cdots, p_{n+m}, t_{n+m} \cdots])$

This operating primitive is to add an iterative branch after p_a and this branch ends at p_b. 'I' is the first alphabet of iterative. For example, the dynamic change from model (a) to (e) in Fig. 1 can be described as $AddI(p_2, p_1, [t_4])$.

Similarly, there are several operating primitives to describe deleting branch (node):

(1) $DeleteS(p_a, t_b, [t_n, p_n, \cdots, t_{n+m}, p_{n+m}, \cdots])$

(2) $DeleteP(t_a, t_b, [p_n, t_{n+1}, p_{n+1}, \cdots, t_{n+m}, p_{n+m}, \cdots])$

(3) $DeleteC(p_a, p_b, [t_n, p_{n+1}, t_{n+1}, \cdots, p_{n+m}, t_{n+m} \cdots])$

(4) $DeleteI(p_a, p_b, [t_n, p_{n+1}, t_{n+1}, \cdots, p_{n+m}, t_{n+m} \cdots])$

All of the dynamic changes can be described by combination of the 8 basic operating primitives mentioned above. For instance, the dynamic change from model (b) to (c) in Fig.1 can be described as $DeleteC(p_1, p_2, [t_4, p_4])$ and $AddP(t_1, t_3, [p_4, t_4, p_5])$.

Model Verification. In workflow management system, once a new model is generated resulting from process dynamic changes, the model correctness should be verified. It is troublesome to check

the correctness of a large and complex model. If dynamic changes occur frequently, it is important for a workflow management system to timely verify if the resulting model is correct or not.

From the preceding definitions and theorems, it is easy to see that each state of the new model must be reachable from its initial state if it is correct. In fact, reachability is also an important criterion for instance migration. If a new workflow model is unreachable, it will not be put to use, let alone the migration of any running instance under its old version. So, it is necessary to validate if the WF-net is reachable. We use incidence matrix to verify Petri net's reachalility. Firstly, how to calculate the incidence matrix is demonstrated.

Definition 3 [8] Let $PN=(P,T,F)$ be a net: $P=\{p_1,p_2\cdots,p_m\}$, $T=\{t_1,t_2\cdots,t_n\}$, the incidence matrix A of PN is given by $a_{ij}=a_{ij}^{+}-a_{ij}^{-}$, where

$$a_{ij}^{+}=\begin{cases}1 & (t_i,p_j)\in F\\ 0 & \text{otherwise}\end{cases}\qquad a_{ij}^{-}=\begin{cases}1 & (p_i,t_j)\in F\\ 0 & \text{otherwise}\end{cases}$$

For a complex model with frequent changes, it is troublesome and time-consuming to calculate the incidence matrix whenever changes occur. Based on the static change region concept presented by Van der Aalst [1], we propose a simplified method for calculating incidence matrix.

Definition 4 Let $PN^O=(P^O,T^O,F^O)$ and $P^N=(P^N,T^N,F^N)$ be two sound WF-net. The static change region in the context of a change from PN^O to PN^N is set $SC=\bigcup_{(x,y)\in X}\{x,y\}$ where $X=(F^O\setminus F^N)\cup(F^N\setminus F^O)$.

PN^O and PN^N in the definition are the WF-net of an old and new model respectively.

A set of nodes affected by dynamic changes in the old and new model respectively can be obtained according to the static change region. In the other hand, the static change region will be used to calculate the incidence matrix.

For example, in Fig. 1, the old model (d) is replaced by a new model (c) because of dynamics changes. So, the static change region is given by comparing the old and new models as follows: $SC=\{p_1,p_2,p_4,p_5,t_1,t_3,t_4\}$.

The affected nodes are (p_1,p_2,t_1,t_3,t_4) in model (d) and $(p_1,p_2,p_4,p_5,t_1,t_3,t_4)$ in model (c).

In addition, we also can get the affected nodes directly according to its corresponding operating primitives. The operating primitives of the dynamic change from model (d) to (c) in Fig. 1 are $DeleteC(p_1,p_2,[t_4])$ and $AddP(t_1,t_3,[p_4,t_4,p_5])$. Nodes included in the operating primitives are all affected nodes. So, $(p_1,p_2,p_4,p_5,t_1,t_3,t_4)$ and (p_1,p_2,t_1,t_3,t_4) are the affected nodes of model (d) and (c) respectively.

Incidence matrix of net (d) is given by:

$$A=\begin{array}{c}\begin{matrix}p_0 & p_1 & p_2 & p_3\end{matrix}\\ \begin{bmatrix}-1 & 1 & 0 & 0\\ 0 & -1 & 1 & 0\\ 0 & 0 & -1 & 1\\ 0 & -1 & 1 & 0\end{bmatrix}\end{array}\begin{matrix}t_1\\ t_2\\ t_3\\ t_4\end{matrix}\rightarrow\begin{array}{c}\begin{matrix}p_0 & p_3 & p_1 & p_2\end{matrix}\\ \begin{bmatrix}0 & 0 & -1 & 1\\ -1 & 0 & 1 & 0\\ 0 & 1 & 0 & -1\\ 0 & 0 & -1 & 1\end{bmatrix}\end{array}\begin{matrix}t_2\\ t_1\\ t_3\\ t_4\end{matrix}$$

Firstly, we place the affected nodes of model (d) at the right lower corner of matrix A through simple row and column transformation. So, we can get a new matrix CA just including the affected nodes, and CA is defined as:

$$CA=\begin{array}{c}\begin{matrix}p_1 & p_2\end{matrix}\\ \begin{bmatrix}1 & 0\\ 0 & -1\\ -1 & 1\end{bmatrix}\end{array}\begin{matrix}t_1\\ t_3\\ t_4\end{matrix}$$

Secondly, we can get a matrix CA' to express the relation of affected nodes of model (c).

$$CA' = \begin{array}{c} \begin{matrix} p_1 & p_2 & p_4 & p_5 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} t_1 \\ t_3 \\ t_4 \end{matrix}$$

Thirdly, we place CA' at the right lower corner of A and fill in the left blanks with value 0, and the other positions are kept their same value as before.

Finally, we get the incidence matrix A' of model (c):

$$A' = \begin{array}{c} \begin{matrix} p_0 & p_3 & p_1 & p_2 & p_4 & p_5 \end{matrix} \\ \begin{bmatrix} 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} t_2 \\ t_1 \\ t_3 \\ t_4 \end{matrix} \rightarrow \begin{array}{c} \begin{matrix} p_0 & p_1 & p_2 & p_3 & p_4 & p_5 \end{matrix} \\ \begin{bmatrix} -1 & 1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \end{matrix}$$

blank

CA'

To this end, we can get the affected nodes according to operating primitives of dynamic changes, but not to draw up a WF-net graph of the model. Additionally, we can quickly calculate the incidence matrix of a new model resulting from dynamic changes when the affected nodes are only a part of the original model.

Theorem 1 [8] A Petri net $PN = (P, T, F)$ is a unique reachable vector net if and only if the incidence matrix A of PN is a full rank matrix, i.e., $Rank(A) = |T|$. For a unique reachable vector net, if there is a non-negative integer n-dimensional vector X, where $M = M_0 + A^T X$, then M is reachable from M_0.

We use the incidence matrix to verify if all the state of the new model is reachable from its initial state according to Theorem 1. The new model is not sound if its corresponding WF-net is unreachable.

Approach of Instance Migration

Dynamic Workflow Model. Here, a dynamic workflow model is proposed to describe the workflow migration. We define a 5-tuple W=(OM, NM, PI, CR, CSN).

- OM represents the old workflow process model based WF-net;
- NM represents the new workflow process model based WF-net;
- PI is a set of instances which are running under the old model: $PI = \{PI_1, PI_2, \cdots\}$;
- CR is a set of change regions on the old model: $CR = \{CR_1, CR_2, \cdots\}$;
- CSN is a set of nodes where each one is the former node of the corresponding change region: $CSN = \{CSN_1, CSN_2, \cdots\}$. For example, CSN_i is the former node of CR_i.

Workflow Migration Method and Rules. According to the criterion of migration validity and the dynamic workflow model mentioned before, a workflow migration method based on WF-net is proposed as follows:

- Once the workflow system detects a new model, all the running instances under the old model will be suspended.

- The workflow system scans the old and new models to obtain the change regions (CR), unchanged regions and CSN within the old and new models.
- The mutual distribution relationship between the executed nodes of running instances and change regions will be determined.
- Different migration strategies will be implemented depending on the distribution relationship. Here, two rules are given to transfer the running instances:
  Rule 1: If the executed node is not belongs to CR, the instance should be transferred to the new model directly.
  Rule 2: If an active node belongs to CR_i, the workflow system must rollback its relevant instance to CSN_i corresponding to CR_i before transferring the instance into new model.
- Once the instance is transferred into new model, it will keep on executing to the end. The later instances will be started under the new model.

Summary

There are many kinds of dynamic changes in workflow systems. The changes existing in process model can be regarded as modifying nodes, such as transitions and places within workflow nets.

This paper provides a set of operating primitives to describe the dynamic changes. When adjusting a workflow model to adapt some dynamic changes, it is critical to verify the correctness of the resulting model. We propose a simplified method to calculate incidence matrix, which can be used to verify if the new model is reachable from its initial state. Actually, the instances should be executed according to the new model as soon as possible, and all finished work should be as useful as possible. Migration is the best solution for this problem and a method is also given for transferring all active instances into the new model.

The work in this paper starts a good direction for dynamic workflow management. It is considerably important for improve the adaptability of existing workflow management systems. In the near future, some extensions for our method can be further researched so as to make it have comprehensive applicability.

References

[1] W.M.P. van der Aalst: Exterminating the Dynamic Change Bug: A Concrete Approach to Support Workflow Change. Information System Frontiers. 3, 297-317 (2001).

[2] F. Casati, S. Ceri, B. Pernici and G. Pozzi: Workflow Evolution. Data and Knowledge Engineering. 24, 211-238 (1998).

[3] W.M.P. van der Aalst and T. Basten: Inheritance of Workflows: an Approach to Tackling Problems Related to Change. Theoretical Computer Science. 270, 125-203 (2002).

[4] S. Rinderle, M. Reichert, P. Dadam: Correctness Criteria for Dynamic Changes in Workflow Systems –a Survery. Data and Knowledge Engineering. 50, 9-34 (2004).

[5] S. Rinderle, M. Reichert and P. Dadam: Flexible Support of Team Processes by Adaptive Workflow System. Distributed and Parallel Databases. 16, 91-116 (2004).

[6] P. Sun and C.J. Jiang: Analysis of Workflow Dynamic Changes Based on Petri Net. Information and Software Technology. 51, 284-292 (2009).

[7] W.M.P. van der Aalst: The Application of Petri Nets to Workflow Management. The Journal of Circuit, System and Computers. 8, 21-66 (1998).

[8] C.Y. Yuan: Theory and Application of Petri Net. Publishing House of Electronics Industry, Beijing (2005).

Safety Distance by Simulation and Collision Avoidance on a Road's Danger Zones

SCHREIBER Peter[1,a], MORAVCIK Oliver[1,b],TANUSKA Pavol[1,c]
VAZAN Pavol[1,d],VRABEL Robert[1,e],BARTUNEK Marian[2,f], HUSAR Peter[3,g]

[1]Slovak University of Technology, Pavlinska 1, 917 24 Trnava, Slovakia

[2]Delphi Slovensko s.r.o., Cacovska cesta 1447/1, 905 01 Senica, Slovakia

[3]University of Technology Ilmenau, Helmholtzplatz 5, 98693 Ilmenau, Germany

[a]peter.schreiber@stuba.sk, [b]oliver.moravcik@stuba.sk, [c]pavol.tanuska@stuba.sk, [d]pavol.vazan@stuba.sk, [e]robert.vrabel@stuba.sk, [f]marian.bartunek@delphi.com, [g]peter.husar@tu-ilmenau.de

Keywords: Safety, braking distance, braking force, adhesion, simulation.

Abstract. The intension of this paper is to describe a simulation for finding the safety distance between two cars on different surfaces which can be used by Pre-crash Braking-control of ACC system. Simulation is provided not only for different surfaces but also for changes between them, changes of a surface gradient and the speed of the cars. The results of the simulation is centered around the cars safety distance.

Introduction

The European Commission requested a decrease in the number of road fatalities of 50% by 2010 which made a challenge for the automotive industry. The main areas of car safety have been defined in the European Action Program: In the area of technical equipment the contemporary power of computers enables the simulation of various processes of car safety assurance.

Simulation is necessary for investigation, adjustment and testing of various assistance systems like ACC, ESP and the Pre-crash Braking-control system, etc. [1], [2]. There are more programs available on the market for the simulation of the car's behavior, for example, CarSim, Adams/Car or Matlab [2], [3].

This paper contains a proposal of road danger zones marking and an assistance system to eliminate serious accidents caused by excessive speed. During the next research, there has to be a set of situations during which the speed on the road danger zone can not be decreased. The assistance system program procedure has to assure the elimination of collisions caused by speed decreasing.

Model of the car

The ACC system, with its pre-crash braking-control system, needs to know as accurately as possible the safety distance between vehicles corresponding to the actual situation. Safety distance s_s is the difference between braking distances of a rear car s_{br} and front car s_{bf} according to (Fig. 1) – this means a difference of distances if cars brake fully from their initial speed. For a front car we have to consider the rear in a logical and more practical way because we measure distance by radar sensor from the rear.

For the simulation it is necessary to know the braking distances of both cars. The breaking distance of a front car is calculated with an estimation of its maximum deceleration a_{max}. A better solution is to obtain a breaking distance of the frontal car to rear car through a communication channel. There are more studies and articles about communication between cars [8] [8].

The safety distance of a rear car to front car (Fig. 1) is computed according to (1):

$$s_s = \frac{v_2^{\,2}}{2a_{2\max}} - \frac{v_1^{\,2}}{2a_{1\max}} + t_r.v_2 + s_0 \tag{1}$$

where s_s is the safety distance, v_1 and v_2 are car speeds, a_{1max} and a_{2max} are car decelerations, t_r is reaction time [9] and s_0 is reserve.

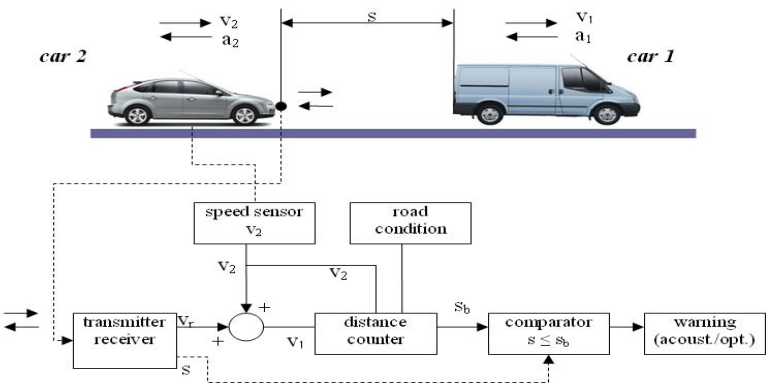


Figure 1: The principle estimating a safety distance [7]

Model for determining of the safety distance in MATLAB/Simulink

The MATLAB/Simulink safety distance model is determined by the simulation that is given in Fig. 2. The model neglects side forces and is not concerned with side stability, it only simulates the braking distances of two cars for longitudinal movement. The model reflects different road surfaces and slopes. It is possible to:

- alternate the road surface during braking,
- change the speed value of cars by braking their start,
- change the adhesion rate for different surfaces according to the speed,
- switch on/off the ABS,
- control the ABS with a PID-controller,
- deliberate the cars inclination in longitudinal direction during the braking,
- simulate a changing road inclination,
- deliberate the air resistance,
- simulate the hydraulic delay of a braking system.

There are programmed two curves of friction coefficient in dependence on wheel slip which are switched by Simulink "Switch" module in regard to the traversed distance. Stopping trajectory is given by integration of the speed v_x. The tangential forces of the left and right wheels are added up and used as inputs for the module "Longitudinal Vehicle Dynamics". The simulation ends by $v_2 = 0$ m/s.

The ABS is simulated with a PID controller. The values of the controller's components are obtained in the Matlab/Simulink® Response Optimization module.

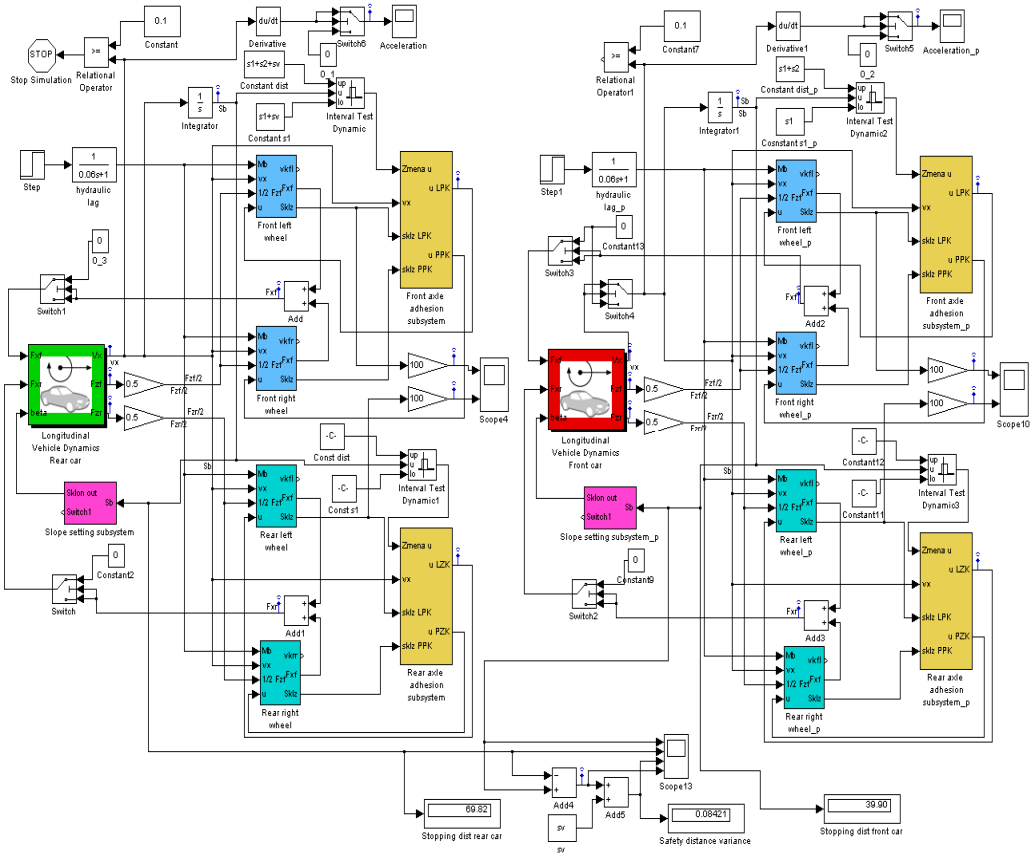
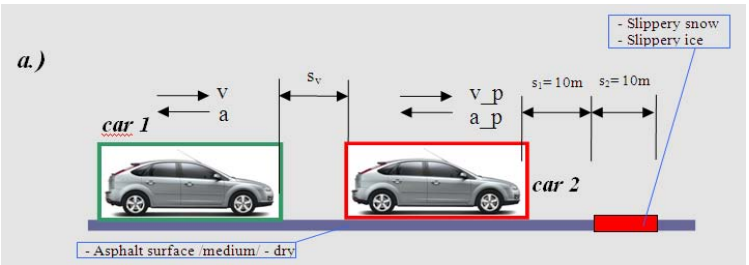


Figure 2: The two cars model of longitudinal movement for the determination of a safety distance

The model provides us with the ability to simulate situations with different conditions which positively or negatively affect the breaking trajectory position. The non influence of the absent ABS system on the safety distance is checked by model simulation, via switched off feedback from the wheel slip regulator.

Typical problems for the determination of safety distance are showed in Fig.3 To scan and identify road surface with adhesion on tenths of a meter is problematic – especially if such part of the road is in front of the front car (Fig. 4a). A case which requires an increase of the safety distance is if the front car starts to brake just behind a worse adhesion surface (Fig. 4b, 4c).



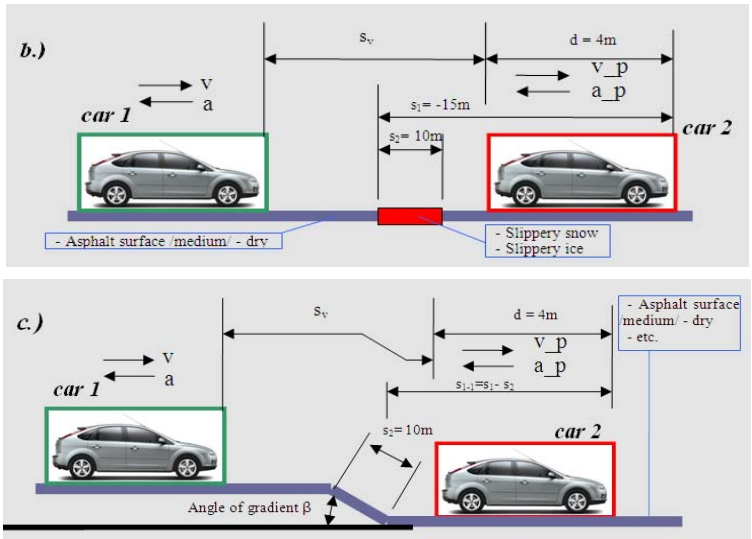
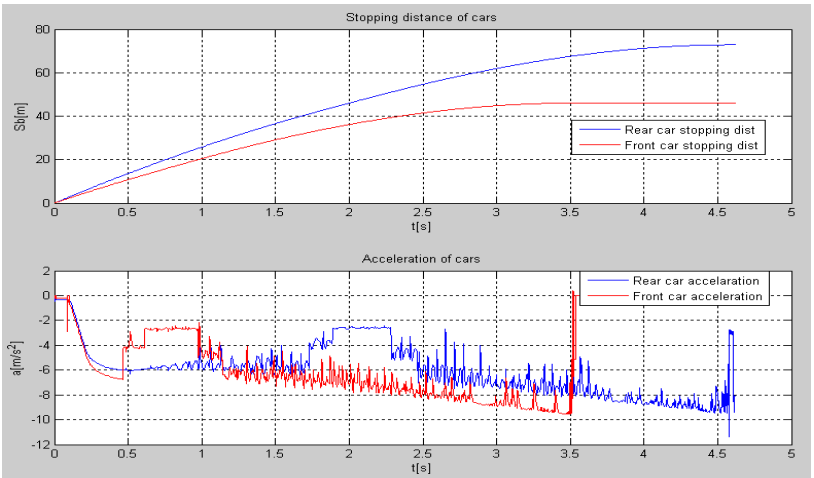


Figure 3: Situations with problematic determination of safety distance

Fig. 4 shows a breaking trajectory and friction coefficient curves which provide us with information about breaking flow. For front car has a set speed of 80km/h and the rear of 100km/h. Charts in Fig. 4a are related to the situation in Fig. 3a, Fig. 4b to Fig. 3b and Fig. 4c to Fig. 3c. On the first friction coefficient chart from Fig. 5, it can be seen that the front car gets through a 10 meters long surface with a worse adhesive condition after 10 meters of trajectory. We can also see movement of the front and rear axle through this surface. During the decreasing of speed, the friction coefficient increases and what can be noticed in the chart is the likely increase of deceleration. In this case the safety distance takes 27 meters.

Moving on a surface with a worse adhesive condition behind the front car (Fig. 3b, 4b) requires an increased safety distance to 33.5 meters. If there is a 10 m slope 9%/5.1˚ (Fig. 3c, 4c) behind the front car then an increase of the needed safety distance takes 3 meters (the needed safety distance is 30 m). The model gives us possibilities that can also be combined with various situations.



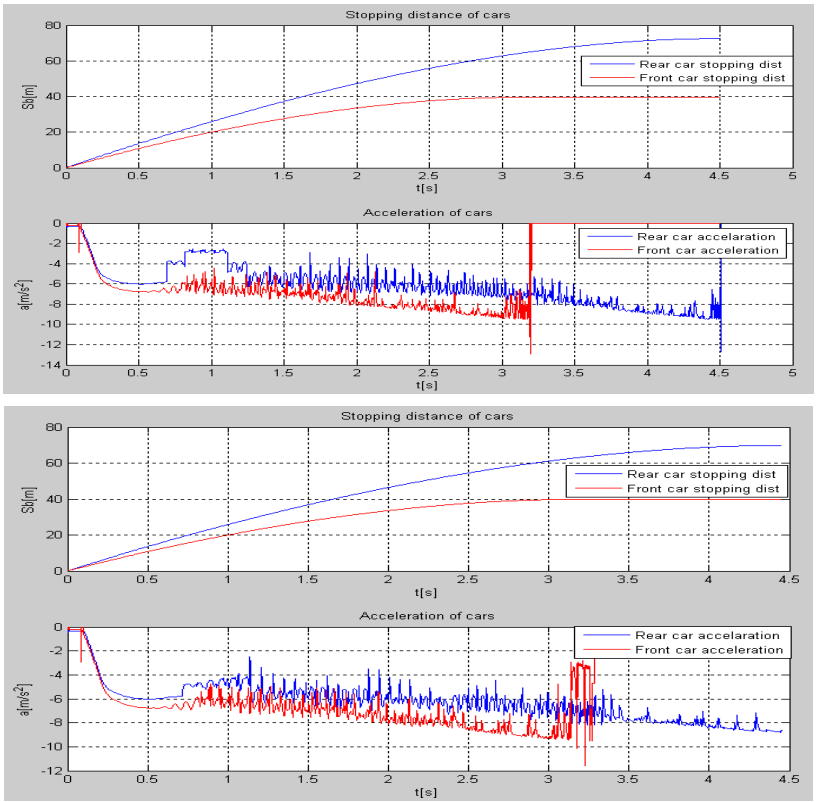


Figure 4: Breaking trajectory and friction coefficient curves for situations from Fig. 3

Beside the curves above, we can export different curves from a model, e.g., front and rear axle loads, wheels slip, braking moments, friction coefficient, etc.

Friction coefficient

The friction coefficient needs to be known for a correct simulation. This is dependant on the tire and road contact surface. A method for detecting the friction coefficient of the road surface includes a step of detecting the operating angle of the vehicle's steering wheel, the vehicle's speed and the operating pressure of the hydraulic power steering unit for the front wheels, and a calculation step of calculating the friction coefficient of the road surface on turning of the vehicle[11,12,13].

Results from determining the safety distance by simulation

- defined maximum deceleration a_{max} for ACC is not suitable for safety distance calculations on roads where a surface with good adhesion can change to a surface with weak adhesion,
- a relatively short range of road with weak adhesion (in the simulation defined on 10 meters) passed by both breaking cars that had a negligible safety distance influence (Fig. 3a),
- relatively short range of road with weak adhesion (in simulation defined on 10 meters) passed just by rear car during breaking (Fig. 3b) affected the safety distance significantly and can not be neglected,
- during the safety distance calculation the road's slope needs to be considered (Fig. 3c). A needed extension of the safety distance on a descending slope increases by increasing the cars' speed difference (highest speed of rear car).

Conclusion

Development in the field of process simulation is accelerated by the growing power of computers. Simulations and their graphical interfaces simplify the process of analyzing and provide a relatively simple optimization and calculation of process characteristics. A correct mathematical model is the basis for simulation. The proposed simulation model can determine a safety distance in various conditions such as, various speeds, road surfaces (varying during braking), road inclination etc. This can also be used for estimating the braking distance between a car and a static obstacle [14],[15]. The next important aspect is the economical modesty of virtual simulations whereby physical testing would be expensive.

References

[1] YOSHIDA, T., KURODA, H., NISHIGAITO, T.: Adaptive Driver-assistance Systems. Hitashi review, 2004.

[2] *Active Safety : Delphi Adaptive Cruise Control* [Online], Information on <http://delphi.com/manufacturers/auto/safety/active/adaptive-cruise-control/>

[3] CarSim. [online], Information on <http://www.carsim.com/products/carsim/>

[4] MATLAB/Simulink. [online], Information on <http://www.mathworks.com/products/simulink/> [cit. 2011-06-27]

[5] *The mission and the objectives of the CAR 2 CAR Communication Consortium*. [Online], Dostupné na internete:< http://car-to-car.org/> *Accessed:* 2011-09-15; 10:45 CET

[6] *OPEL-EYE*. [Online], Information on < http://www.opel.com/microsite/astra-gtc/innovation.html#/innovation_opel_eye> *Accessed:* 2011-09-6; 12:45 CET

[7] VLK, F. Automobilová elektronika 1 Asistenční a informační systémy (Car Electronics 1, Assistance and Information Systems), Vol. 1., Brno, Prof. Ing. František Vlk, DrSc. publishing, 2006. 269 pp. ISBN 80-239-6462-3

[8] FOSTER, H., MUKHIJA, A., ROSENBLUM, D. S., UCHITEL, S.: A model-driven approach to dynamic and adaptive service brokering using modes. In Service-Oriented Computing ICSOC 2008, volume 5364 of Lecture Notes in Computer Science, s. 558–564. Springer Berlin, Heidelberg 2008.

[9] *Reakčná a brzdná dráha vozidla* [Online], Information on <www.bcp.sk>

[10] VLK, F. Dynamika motorových vozidel (Dynamics of motor-cars), Vol 2., Brno, Prof. Ing. František Vlk, DrSc. publishing, 2003. 432 pp. ISBN 80-239-0024-2

[11] MOMOSE, N., YOSHIDA, H. Method and apparatus for detecting friction coefficient of road surface, and method and system for four-wheel steering of vehicles using the detectedfriction coefficientof road surface, United States Patent Number: 5,365,439

[12] FUKAMIZU, H. et al. Road surface condition detection system, United States Patent Number: 4,690,553

[13] HALASKOVA, J., VOJTESEK, A.: Stanovení brzdné dráhy vozidel za různých podmínek. (Braking distance estimation in different conditions), Brno: VUT v Brně.

[14] VAZAN, P: The Application of Simulation Methods in Manufacturing System Control. - 1 st ed. - Köthen : Hochschule Anhalt, 2009. - 102 s. - ISBN 978-3-86011-025-6

[15] TANUSKA, P., KUNIK, S., KOPCEK, M.: Exothermic CSTR: Modeling, control & simulation. In: Annals of DAAAM and Proceedings of DAAAM Symposium. - ISSN 1726-9679. - Vol. 20, No. 1 Annals of DAAAM 2009 & Proceedings of the 20th international DAAAM symposium "Intelligent manufacturing & automation, November 2009, Vienna, Austria, ISBN 978-3-901509-70-4, s. 0203-0204

Solving a Four-Point Boundary Value Problem Fordynamical Systems with High-Speed Feedback with MATLAB

Robert Vrabel[1,a], Peter Schreiber[1,b], Oliver Moravcik[1,c], Ingrida Mankova[1,d]

[1]Institute of Applied Informatics, Automation and Mathematics, Faculty of Materials Science and Technology, Hajdoczyho 1, 917 01 Trnava, Slovakia

[a]robert.vrabel@stuba.sk,[b] peter.schreiber@stuba.sk, [c] oliver.moravcik@stuba.sk, [d]ingrida.mankova@stuba.sk

Keywords: Dynamicalsystemswith high-speed feedback, four-point boundary value problem, MATLAB,

Abstract. This submission deals with the four-point boundary value problem for the nonlinear second-order systems with high-speed feedback. Especially, we focus on the simulation techniques of the solutions on given finite length interval from an appearance of the so-called boundary layers point of view. Boundary layers correspond to the rapid region of transition in the exact solution. Singular Perturbation Theory is the mathematical framework that yields the tools to explore the complicated dynamical behavior of these systems.

Introduction

The study of singularly perturbed control systems as a mathematical framework for description of dynamical systems with high-speed feedback is motivated by many problems coming from chemistry, physics and engineering [1, 2, 3, 4]. Singular perturbations and time-scale techniques have become common tools for the modeling, analysis, and design of control systems. One of the typical behaviors of singularly perturbed systems is the boundary layer phenomenon: the solutions vary rapidly within very thin layer regions near the boundary. Boundary layers are formed due to the nonuniform convergence of the exact solution $y_\epsilon$of the mathematical model to the solution η of degenerated problem in the neighborhood of the ends x_i and x_f of the considered interval $[x_i, x_f]$.

We use the singular perturbation techniques to approximate exact solutionby the linear combination of solution of degenerated problem and exponentiallysmall layer functions.

In this submission, we will consider a singularly perturbed system with quadratic nonlinearity of the form

$$\epsilon y'' + ky = y^2 + u(x),\ k < 0, u \in C^2([x_i, x_f]), 0 < \varepsilon << 1, \tag{1}$$

subject to the nonlocal boundary conditions

$$y_\epsilon(x_{m1}) - y_\epsilon(x_i) = 0,\ y_\epsilon(x_f) - y_\epsilon(x_{m2}) = 0,\ x_i < x_{m1} < x_{m2} < x_f. \tag{2}$$

Singular perturbation method is applied to obtain an approximate solution of singularly perturbed systems (1), (2) composed of a solution η of a degenerate problem $ky = y^2 + u(x)$, small constant and two boundary layer functions to recover the lost nonlocal boundary conditions in the degeneration process.

Using the software package MATLAB, we show that the solutions of (1), (2), in general, start with fast transient $\left(\left|y_{\epsilon}^{'}(x_i) \to \infty\right|\right)$ and after decay of this transient they remain close to the solution $\eta(x)$ of a degenerate system with an arising new fast transient of $y_{\epsilon}(x)$ from $u(x)$ to $y_{\epsilon}(x_f)\left(\left|y_{\epsilon}^{'}(x_f) \to \infty\right|\right)$, which is the so-called boundary layer phenomenon [5, 6].

Boundary value problems (1), (2) can arise in the study of the steady-states of a heated bar with a thermostat described by scalar partial differential equation

$$\frac{\partial y}{\partial t} = \epsilon \frac{\partial^2 y}{\partial x^2} + ky - y^2 - u(x)$$

with stationary condition $\partial y/\partial t = 0$, where the controllers at $x = x_i$ and $x = x_f$ maintain a temperature according to the temperature registered by the sensors at $x = x_{m1}$ and $x = x_{m2}$, respectively. In this case, we consider a uniform bar of length $x_f - x_i$ with non-uniform temperature lying on the x-axis from $x = x_i$ to $x = x_f$. The parameter ϵ represents the thermal diffusivity. Thus, the singular perturbation problems are of common occurrence in modeling the heat-transport problems with large Peclet number [7, 8].

The following assumptions will be made throughout the paper.

A1. For a degenerated problem (putting $\epsilon = 0$in (1)) $ky = y^2 + u(x)$ there exists C^2 function η such that $k\eta(x) = \eta^2(x) + u(x)$ on $\left[x_i, x_f\right]$.

Denote $H(\eta) = \left\{(x, y); x_i \le x \le x_f, |y - \eta(x)| < d(x)\right\}$, where $d(x)$ is the positive continuous function on $\left[x_i, x_f\right]$ such that

$$d(x) = \begin{cases} |\eta(x_{m1}) - \eta(x_i)| + \delta & \text{for } x_i \le x \le x_i + \dfrac{\delta}{2} \\ \delta & \text{for } x_i + \delta \le x \le x_f - \delta \\ \left|\eta(x_f) - \eta(x_{m2})\right| + \delta & \text{for } x_f - \dfrac{\delta}{2} \le x \le x_f, \end{cases}$$

$\delta$is a small positive constant.

A2. The function $f(x, y) = y^2 + u(x)$satisfies the condition

$$\left|\frac{\partial f(x, y)}{\partial y}\right| \le w < -k \text{ for evere } (\mathrm{x}, \mathrm{y}) \in \mathrm{H}(\eta).$$

Boundary layer functions

The boundary layer functions are defined as follows (for detail see [9])

$$\zeta_{\epsilon}(x) = \frac{\eta(x_{m1}) - \eta(x_i)}{D} \cdot \left(e^{\sqrt{\frac{m}{\epsilon}}(x_f - x)} - e^{\sqrt{\frac{m}{\epsilon}}(x - x_f)} + e^{\sqrt{\frac{m}{\epsilon}}(x - x_{m2})} - e^{\sqrt{\frac{m}{\epsilon}}(x_{m2} - x)} \right),$$

$$\hat{\zeta}_{\epsilon}(x) = \frac{\left|\eta(x_f) - \eta(x_{m2})\right|}{D} \cdot \left(e^{\sqrt{\frac{m}{\epsilon}}(x - x_i)} - e^{\sqrt{\frac{m}{\epsilon}}(x_i - x)} + e^{\sqrt{\frac{m}{\epsilon}}(x_{m1} - x)} - e^{\sqrt{\frac{m}{\epsilon}}(x - x_{m1})} \right),$$

$$D = \left(e^{\sqrt{\frac{m}{\epsilon}}(x_f - x_i)} + e^{\sqrt{\frac{m}{\epsilon}}(x_{m2} - x_{m1})} + e^{\sqrt{\frac{m}{\epsilon}}(x_{m1} - x_f)} + e^{\sqrt{\frac{m}{\epsilon}}(x_1 - x_{m2})} \right)$$

$$-\left(e^{\sqrt{\frac{m}{\epsilon}}(x_i-x_f)}+e^{\sqrt{\frac{m}{\epsilon}}(x_{m1}-x_{m2})}+e^{\sqrt{\frac{m}{\epsilon}}(x_f-x_{m1})}+e^{\sqrt{\frac{m}{\epsilon}}(x_{m2}-x_i)}\right),$$

$$m=-k-w.$$

The assumptions of the main theorem of the paper [9] (Theorem 2.1 – the assumptions A1 and A2) determining the sufficient conditions for existence of solutions of (1), (2) are satisfied if and only if there exists $w>0$ such that

$$\tfrac{1}{4}(k^2-(w-k)^2)<u(x)<\tfrac{1}{4}(k^2-(w+k)^2) \quad \text{on} \quad [x_i,x_f] \tag{3}$$

$$|u(x_{m1})-u(x_i)|<\tfrac{1}{8}\big(w-k-\lambda(x_i)\big)\big(\lambda(x_i)+\lambda(x_{mi})\big) \tag{4}$$

$$\big|u(x_f)-u(x_{m2})\big|<\tfrac{1}{8}\Big(w-k-\lambda\big(x_f\big)\Big)\Big(\lambda\big(x_f\big)+\lambda(x_{mi})\Big) \tag{5}$$

$$|u(x_{m1})-u(x_i)|<\tfrac{1}{8}\big(w+k+\lambda(x_i)\big)\big(\lambda(x_i)+\lambda(x_{mi})\big) \tag{6}$$

$$\big|u(x_f)-u(x_{m2})\big|<\tfrac{1}{8}\Big(w+k+\lambda\big(x_f\big)\Big)\Big(\lambda\big(x_f\big)+\lambda(x_{mi})\Big), \tag{7}$$

where$\lambda(x)=\sqrt{k^2-4u(x)}$.

For an illustrative example we consider the problem (1), (2) with $k=-2$,

$$u(x)=x, x_i=0, x_f=\frac{1}{2}, x_{m1}=\frac{1}{6\text{and } x_{m2}}=\frac{2}{6},$$

i. e.

$$\epsilon y^{''}-2y=y^2+x, \tag{8}$$

$$y_\epsilon(1/2)-y_\epsilon(0)=0, \quad y_\epsilon(1/2)-y_\epsilon(2/6)=0. \tag{9}$$

It is not difficult to verify that the solution

$$\eta(x)=-1+\sqrt{1-x}$$

of degenerated problem (putting $\epsilon=0$ in (8)) satisfies the conditions $(3)-(7)$ for every

$$w\in\left(\frac{4}{3}\big[\lambda\big(x_f\big)+\lambda(x_{m1})\big]^{-1}+2-\lambda\big(x_f\big),2\right)\doteq(0.9957,2).$$

Thus, on the basis of theorem mentioned above, there exists $\epsilon_0=\epsilon_0(\omega)$ such that for every $\epsilon\in(0,\epsilon_0]$ the problem (8), (9) has the unique solution in H.
For $\eta\big(x_f\big)-\eta(x_{m2})\le 0$ we define the approximate realization $\tilde{y}_\epsilon(x)$ of singularly perturbed system (1), (2) by

$$\tilde{y}_\epsilon(x)=\eta(x)+\zeta_\epsilon(x)+\hat{\zeta}_\epsilon(x)+C\epsilon \tag{10}$$

and analogously, for $\eta\big(x_f\big)-\eta(x_{m2})\ge 0$ we define

$$\tilde{y}_\epsilon(x)=\eta(x)+\zeta_\epsilon(x)-\hat{\zeta}_\epsilon(x)-C\epsilon \tag{11}$$

where the ϵ −independent constant $C=\frac{1}{m}max\{|\eta^{''}(x)|;x\in[x_i,x_f]\}$ with $m=-k-w$.

The unique solution of (8), (9) is $O(\epsilon)$ close to the approximate solution (10)on $[x_i, x_f]$, that is, to the function (Fig. 1)

$$\tilde{y}_\epsilon(x) = -1 + \sqrt{1-x} + \zeta_\epsilon(x) + \hat{\zeta}_\epsilon(x) + \epsilon\left[(2-w)\sqrt{2}\right]^{-1}.$$

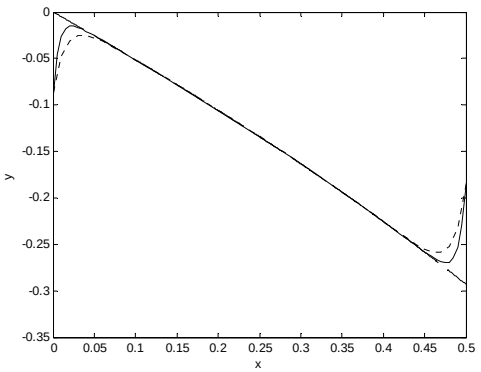


Fig.1. Solution of$\epsilon y'' - 2y = y^2 + x, y_\epsilon\ (1/6) - y_\epsilon(0) = 0, y_\epsilon\ (1/2) - \ y_\epsilon(2/6) = 0$,(thesolid line) for $\epsilon = 0.0001$. The dotted and dashed lines represent the approximate solution$\tilde{y}_\epsilon\ (x)$(for $w\ =\ 1.5$) and solution of degenerated problem, the function $\eta(x) = -1 + \sqrt{1-x}$,respectively.

Simulation using MATLAB

The following is MATLAB7 code that solves the example (8), (9):

```
function threebvp3(solver)
ifnargin< 1    %Number of function input arguments
solver = 'bvp4c';
end
bvpsolver = fcnchk(solver);
% Initial mesh - duplicate the interface point xc .
xc1=0.1666; xc2=0.3333
xinit = [0, 0.03, 0.05, 0.07, 0.10, 0.15, xc1, xc1,
0.20, 0.23, 0.25, 0.28, 0.30, xc2, xc2,
0.35, 0.40, 0.42, 0.45, 0.47, 0.50];
% Constant initial guess for the solution
yinit = [-0.13; 1];
% The initial profile
sol = bvpinit(xinit,yinit);
sol = bvpsolver(@f,@bc,sol);
figure
plot(sol.x,sol.y(1,:),'k')
title(' ')
xlabel('t')
ylabel('y')
functiondydx = f(x,y,region)
dydx = zeros(2,1);
dydx(1)=y(2);
switch region
case 1
dydx(2)=(2*y(1)+y(1)^2+x)/0.0001; %\epsilon=0.0001
```

```
case 2
dydx(2)=(2*y(1)+y(1)^2+x)/0.0001;
case 3
dydx(2)=(2*y(1)+y(1)^2+x)/0.0001;
otherwise
error('MATLAB:threebvp3:BadRegionIndex','
Incorrect region index:\%d',region);
end
end
% ------------------------------------------
% Boundary (and internal) conditions
function res=bc(YL,YR)
res=[YL(1,1)-YR(1,1) % y0(1)-y1/6(1)=0
YR(1,1)-YL(1,2) % con of y(1) at xc1=1/6
YR(2,1)-YL(2,2) % con of y(2) at xc1=1/6
YR(1,2)-YL(1,3) % con of y(1) at xc2=2/6
YR(2,2)-YL(2,3) % con of y(2) at xc2=2/6
YL(1,3)-YR(1,3) ]; % y2/6(1)-y1/2(1)=0
end
% ------------------------------------------
end % threebvp3
```

References

[1] R.Vrabel and M. Abas: *Frequency control of singularly perturbed forced Duffing's oscillator*. J. of Dynamical and Control Systems 17, No. 3, pp. 451-467 (2011)

[2] P.Tanuska, S.Kunik and M.Kopcek:*Exothermic CSTR: Modeling, Control &Simulation.* In: Annals of DAAAM for 2009&proceeding of the 20th International DAAAM Symposium. Book Series: Annals of DAAAM and proceedings 20, pp. 203-204 (2009)

[3]D.Mudroncik, P.Tanuska andM.Galik: *Surge Control of Natural Gas Centrifugal Compressor*. In: Second International Conference on Computer and Electrical Engineering, Vol.1, Proceedings. Book Series.International Conference on Computer and Electrical Engineering ICCEE, pp. 110-113. DOI: 10.1109/ICCEE.2009.112 (2009)

[4]R.Vrabel: *Singularly pertubedanharmonic quartic potential oscillator problem*. ZeitschriftfürangewandteMathematik und Physik ZAMP 55, pp. 720-724 (2004)

[5]M.Gopal: Modern Control System Theory. New Age International, New Delhi (1993)

[6] P.Kokotovic,H.K.Khali and J. O'Reilly: Singular Perturbation Methods in Control, Analysis and Design. Academic Press, London (1986)

[7]E.Burman, J.Guzman and D.Leykekhman: *Weighted error estimates of the continuous interior penalty method for singularly perturbed problems.* IMA Journal of Numerical Analysis 29, No. 2, 284-314 (2009)

[8] A. Khan, I. Khan, T. Aziz and M. Stojanovic: *A variable-mesh approximation method for singularly perturbed boundary-value problems using cubic spline in tension.* International Journal of Computer Mathematics81, No. 12, 1513-1518 (2004)

[9] R.Vrabel: *Nonlocal Four-Point Boundary Value Problem for the SingularlyPerturbed Semilinear Differential Equations.* Boundary value problems,Article ID 570493, 70493-70493 (2011)

A New Representation of Emotion in Affective Computing

Leonid Ivonin[a], Huang-Ming Chang[b], Wei Chen[c], and Matthias Rauterberg[d]

Designed Intelligence Group, Department of Industrial Design,

Eindhoven University of Technology

Den Dolech 2, 5612 AZ, Eindhoven, The Netherlands

[a] l.ivonin@tue.nl, [b] h.m.chang@tue.nl, [c] w.chen@tue.nl, [d] g.w.m.rauterberg@tue.nl

Keywords: Emotion, Affective Space, Polar Coordinate System

Abstract. In the recent years, increasing attention has been paid to the area of affective computing, which deals with the complex phenomenon of human emotion. Therefore, a model for describing, structuring, and categorizing emotional states of users is required. The dimensional emotion theory is one of widely used theoretical foundations for categorization of emotions. According to the dimensional theory, emotional states are projected to the affective space, which has two dimensions: valence and arousal. In order to navigate in the affective space, Cartesian coordinate system is used, where emotion quality is defined by combination of valence and arousal. In this paper, we propose another representation of the affective space with polar coordinate system. The key advantages of such a representation include (1) capability to account not only for emotion quality, but also for emotion intensity, (2) reasonable explanation of the location of neutral emotion in the affective space, and (3) straightforward interpretation of the meaning of an emotional state (quality defined by angle and intensity defined by distance from the origin). Although in our experiment most of the induced emotions can be differentiated with polar coordinate system, further investigation is still needed to find out either Cartesian or polar coordinates system represents affective space better in practice.

Introduction

The dimensional emotion theory is based on the idea of a reduction of complex multidimensional phenomenon to more simple representation [1], which involves a low number of meaningful dimensions. The most common variation of the dimensional theory involves the dimensions of arousal and valence, and, therefore, creates a two-dimensional affective space [2]. The dimensional emotion theory is popular among researchers and is used in many applications, such as [3].

However, relying exclusively on the valence and arousal dimensions to describe emotional state seems insufficient to represent an important aspect of emotion, namely the intensity. Traditionally the fact that emotion can vary in intensity received surprisingly little reflection in theories of emotion. Frijda et al. pointed out in [4] that the intensity is one of the most salient features of emotion and one cannot talk about emotion without talking about emotion intensity. These considerations trigger a question how the intensity of emotion can be reflected in the affective space. According to Russell [5], the circular ordering of emotions in the affective space can complement the dimensional representation and the distance from an emotional state to the origin of the space can be interpreted as the intensity of emotion. Therefore, it might be reasonable to use polar coordinate system to navigate in the affective space.

Aims of the Present Study

Although polar coordinate system was already applied earlier [6], the majority of researches preferred to use Cartesian coordinate system to determine the positions of emotions within the affective space. The horizontal axis (X) was commonly used to represent valence and the vertical one (Y) was used for arousal. The intersection of the axes was considered to be the point of origin and represent a kind of neutral emotional state. However, such an approach is questionable for several reasons. First, it does not offer a convenient way to account for emotion intensity. Second, according to the dimensional theory the origin represents the neutral emotional state. However, based on the experimental data from the previous research [7], the neutral emotional state is not precisely located in the origin of the affective space. These observations challenge the statement that the origin represents neutral state. Moreover, they question where in the affective space the origin should be located and what the meaning of the neutral emotional state is. In our study, we investigate whether an introduction of a new way to represent the affective space might solve the issues mentioned above.

Approach

There are two widely used coordinate systems for two-dimensional spaces: Cartesian and polar. Each of them allows an unambiguous identification of a point in 2D space, but there are tasks, which can be easier solved in Cartesian coordinates system than in polar, and vice versa. A good example of such a task is the equation of a circle centered at the origin that is more simple and elegant in polar rather than in Cartesian coordinate system.

Our hypothesis, which we want to put forward in the present paper, is that although the affective space is usually described with Cartesian coordinate system, it might be more appropriate and advantageous to navigate the affective space with polar coordinate system. To the best of our knowledge it was first proposed by Russell [5] that emotional states in the affective space can be distinguished with angle in polar coordinate system. Moreover, he suggested that neutral emotional states would fall near the origin of the affective space, while the states with strong intensity would be located further from the origin. Therefore, the distance between the origin and an affective state is interpreted as emotion intensity. Similar ideas can be found in the work of Reisenzein [8], who argued that dimensional theory should account for emotion quality and emotion intensity because otherwise a theory cannot be regarded as an adequate theory of the structure of emotional experience. Reisenzein used Cartesian coordinate system and proposed that emotion quality is defined by the proportion of valence and arousal and emotion intensity is defined by absolute values of valence and arousal [8]. In polar coordinate system this would mean that quality of emotion is defined by angle, and emotion intensity is defined by radius.

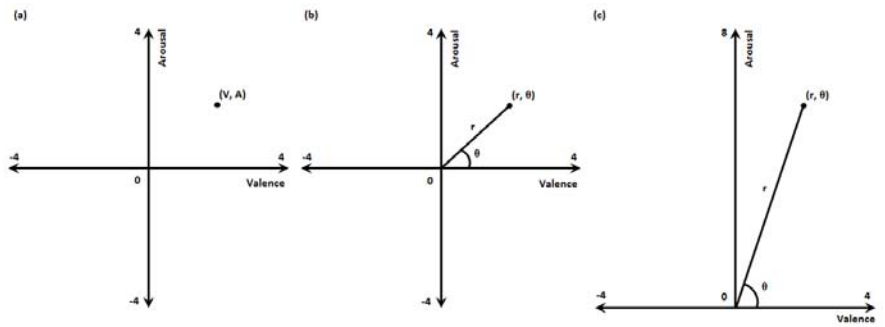


Fig. 1 Representations of the affective space: Cartesian (a), polar (b), and modified polar coordinate systems (c). A dot in the affective space corresponds to a particular emotional state.

However, there is a difficulty associated with the representation of the emotional states in polar coordinate system, which is determined by the fact that rules of linear statistics do not apply for circular data. Consider the following two angles as an example: 2º and 358º. If we operate in a linear

space, then calculating the mean of these angles would result in an angle of 180º. However, it is obvious that the result is wrong and the correct mean angle is 0º. This example illustrates the fundamental difference between linear and circular statistics [9]. Unfortunately, only few statistical software packages support analysis of circular data, and it might be an obstacle for adopting the representation of the affective space with polar coordinate system. Furthermore, in polar coordinate system we faced the problem of interpretation of the emotional states that are located close to the origin of the coordinate system. It was unclear what quality (defined by angle) of emotion with zero intensity (defined by distance) is, because in polar coordinate system a point with zero distance from the origin can have arbitrary angle.

Taking the above-mentioned issues into account, we transformed the dimension of arousal by adding 4 to the original values. The modified coordinate system has now the origin located at the (-4) end of the old dimension of arousal (as shown on Figure 1(c)). The obvious benefit of this modification is that it avoids the difficulty with the statistical analysis of circular data, because in modified polar coordinate system angle can vary only between 0º and 180º, and, for this reason, linear statistics can be used. Moreover, the representation of the affective space with modified polar coordinate system solves the problem of interpretation of emotional states with zero intensity, because they can be assumed to have quality (defined by angle) of 'neutral' emotion. The validity of the modified polar system needs an investigation. In order to address the questions highlighted above we designed an experiment with the data from International Affective Picture System (IAPS) [10]. The results will be mapped into the affective space as shown on Figure 1 (a, c) and the representations will be evaluated with the aim to identify which of them is supported by empirical data as the most suitable representation of emotion.

Experiment

According to the previous study that applied IAPS [10], the stimuli in these databases are generally clustered into four categories: Positive-Arousing, Positive-Relaxing, Neutral, and Negative. IAPS contains sufficient amount of visual stimuli, including 1194 pictures. For our experiment we selected six stimuli for each category (24 pictures in total) from the original pool of stimuli that are contained in this database. Our study followed the method used by IAPS, utilizing the Self-Assessment Manikin (SAM) [11] as a measuring tool for participants to consciously report their emotion. SAM captures two dimensions of an emotional state: valence and arousal. The selected stimuli were presented to participants on a computer screen and their self-assessment ratings for every stimulus were collected via a web-based interface.

For our experiment, 37 healthy participants, including 17 males and 20 females, were recruited with payment of 9 euros. The participants had diverse nationalities: 17 from Asia, nine from Europe, eight from Middle East, and three from South America. Participants in this study had a mean age of 26 years and nine months, ranging from 18 to 50 years (one under 20 years old, two above 40 years old).

Table 1. An overview of the stimuli used in the experiment, including five categories: Positive-Arousing (PA), Positive-Relaxing (PR), Neutral (NT), and Negative (NG).

Category	Stimuli (Code Number)
PA	Erotic (4652), Erotic (4668), Cupcakes (7405), Sailing (8080), Bungee (8179), RollerCoaster (8490)
PR	Butterfly (1605), Rabit (1610), Baby (2060), NeutBaby (2260), Nature (5760), Clouds (5891)
NT	Mushroom (5530), RollinPin (7000), HairDrier (7050), Book (7090), Lamp (7175), Cabinet (7705)
NG	BurnVictim (3053), BabyTumor (3170), AimedGun (6230), Attack (6350), Vomit (9321), Dead (9412)

The experiment was set up with a web-based system for presentation of stimuli and the experimental data were stored in a database for further analysis. The experiment followed

within-subjects design. Every participant went through the same set of stimuli (see Table 1). Before the start of the experiment, each participant completed a tutorial to get familiar with the controls and the interface. Once the experiment session began, the screen started to display pictures one at a time in a random order. Each picture was exposed to a participant for six seconds. Then the interface paused for five seconds. The SAM scaled were shown after the pause. Participants had unlimited time to report their emotional feelings. Another 5-second pause appeared after the self-report, which was meant to let participants calm down and recover from the previously induced emotion. Then the next picture was shown. All of the 37 participants ran through the whole procedure individually.

Results

Multivariate analysis of variance for repeated measurements, which we conducted with two coordinate systems, demonstrated that there is a significant main effect of the category of stimuli on the self-assessment ratings provided by participants (Cartesian: $F(6,31) = 98.742$, $p<0.001$; modified polar: $F(6,31) = 105.595$, $p<0.001$). Moreover, inference tests of within-subject contrasts among all of the four categories were performed in two coordinate systems using univariate analysis of variance. The mean values of the self-assessment ratings for every category are plotted in the affective space with both Cartesian and modified polar coordinate systems at Figure 2.

Table 2. Inferential statistics of the self-assesment ratings in the affective space using catesian and modified polar coordinate systems for every category (Positive-Relaxing (PR), Positive-Arousing (PA), Neutral (NT), and Negative (NG)) is shown. ΔM indicates the difference in the mean values of two categories (category 1 (C1) minuses category 2 (C2)); p value shows the results of the tests of within-subject contrasts on ratings among each category in two different representations (Cartesian and modified polar).

Attributes		**Cartesian**				**Modified polar**		
	C1 \ C2	**PA**	**NT**	**NG**		**PA**	**NT**	**NG**
Valence	**PR**	ΔM=0.77 p<0.001 ***	ΔM=1.648 p<0.001 ***	ΔM=5.045 p<0.001 ***	**Distance**	ΔM=-1.548 p<0.001 ***	ΔM=0.635 p=0.016 *	ΔM=-2.266 p<0.001 ***
	PA	-	ΔM=0.878 p<0.001 ***	ΔM=4.275 p<0.001 ***		-	ΔM=2.183 p<0.001 ***	ΔM=-0.718 p=0.001 ***
	NT	-	-	ΔM=3.397 p<0.001 ***		-	-	ΔM=-2.901 p<0.001 ***
Arousal	**PR**	ΔM=-2.356 p<0.001 ***	ΔM=-0.262 p=0.381	ΔM=-2.744 p<0.001 ***	**Angle**	ΔM= -24.923 p<0.001 ***	ΔM= -29.493 p<0.001 ***	ΔM=-67.376 p<0.001 ***
	PA	-	ΔM=2.094 p<0.001 ***	ΔM=-0.388 p=0.103		-	ΔM= -4.57 p=0.154	ΔM=-42.453 p<0.001 ***
	NT	-	-	ΔM=-2.482 p<0.001 ***		-	-	ΔM=-37.883 p<0.001 ***

(* represents p value < 0.05, which shows significance; ** represents p value <= 0.01, which shows high significance; *** represents p value <= 0.001, which shows very high significance.)

According to the experimental data presented in Table 2, most of the categories of stimuli can be differentiated by the valence ratings using Cartesian coordinate system. Only the arousal ratings of Positive-Relaxing and Neutral as well as of Positive-Arousing and Negative categories are not significantly different in Cartesian coordinate system. In modified polar coordinate system the angles of Positive-Arousing and Neutral categories were not significantly different. Other categories could be distinguished by angle in modified polar coordinate system. The distance from the origin in modified polar coordinate system successfully allowed differentiation between all the categories.

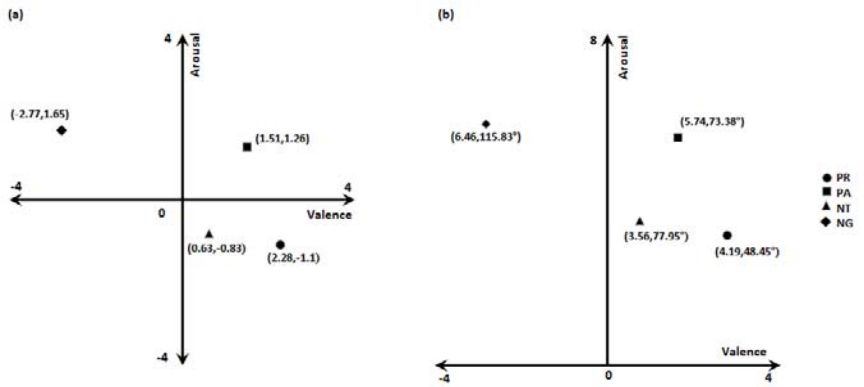


Fig. 2. Four categories (Positive-Relaxing (PR), Positive-Arousing (PA), Neutral (NT), and Negative (NG)) of the stimuli are plotted in the affective space with Cartesian (a) and modified polar (b) coordinate systems.

Discussion

Based on the analysis of the data presented in Cartesian coordinate system, the differences between the four categories of stimuli are significant and their positions in the affective space are consistent with the previous research [10, 11]. Therefore, a conclusion can be drawn that the experimental materials and design are valid. In order to answer the research questions, we compared the representations of the affective space with Cartesian and polar coordinate systems. As it can be seen from Table 2, modified polar coordinate system has one non-significant effect among all categories, whereas Cartesian coordinate system has two. Although this data implies that modified polar coordinate system better describes the affective space, this advantage is not as clear as we expected. Nevertheless, the representation of the affective space with polar coordinate system (modified or non-modified) provides an additional benefit of the capability to define emotion quality and emotion intensity in a straightforward manner. However, in the previous sections a modified configuration of polar coordinate system with the transformed dimension of arousal was introduced for the following reasons.

First, we encountered the above-mentioned problem of arbitrary values of an angle that corresponds to emotional states with zero intensity. For instance, it is unclear what emotion quality should have emotion with zero intensity. Should it have quality of the corresponding emotion or neutral quality? If the first assumption is correct, then it is necessary to know the angle, which defines the emotion quality; however, the angle cannot be computed, because the emotional state is located in the origin. On the other hand, if the second assumption is correct, there is a contradiction between locations of the neutral emotional state (see Figure 2(a)) and the emotional states with zero intensity. According to the empirical data, neutral emotions are not located in the origin of the affective space and have certain emotion quality and intensity. For this reason it is not very plausible to treat emotional states with zero intensity as ‘neutral’ emotion.

Second, intuitive considerations seem to challenge the concept of negative arousal. Indeed, it seem to be plausible that there are emotional states with high, medium or low arousal and theoretically with zero arousal as well, but it is not clear how arousal can be negative and what is the meaning of negative arousal. Moreover, the literature in this field also suggests that arousal does not have negative values [12]. Therefore, from our point of view, the configuration of the affective space should only contain non-negative values of arousal.

Analysis of the experimental data presented in modified polar coordinate systems revealed that all four categories of stimuli, except the pair of Positive-Arousing and Neutral, can be distinguished one

from another by angle. It is not clear why these two categories have the same emotion quality and this question should be further investigated. Overall, the modified polar coordinate system enabled us to distinguish every category of stimuli by angle and thus justified the usefulness of the idea to transform the dimension of arousal.

Conclusion

Affective computing applications require a model for structuring and categorization of emotional states. As it was discussed above, the dimensional emotion theory is so far one of the most widely used theoretical foundations for description of emotion. In the dimensional theory emotional states are mapped to the affective space, which has the two dimensions of valence and arousal, using Cartesian coordinate system. According to our experimental data, polar coordinate system is also capable of representing the affective space and, in fact, is more suitable for this purpose than Cartesian coordinate system, because emotion quality and emotion intensity can be naturally expressed with angle and distance from the origin. However, in order to have meaningful and convenient for statistical analysis representation of emotion quality and emotion intensity, the origin of polar coordinate system should be redefined via the dimension of arousal with only non-negative values. Despite of the first promising results it is still unclear whether modified polar coordinate system is superior to Cartesian coordinate system, and, therefore, further research with larger sets of emotional stimuli is required to validate it.

Acknowledgments This work was supported by the Erasmus Mundus Joint Doctorate (EMJD) in Interactive and Cognitive Environments (ICE), which is funded by Erasmus Mundus of the European Commission under EMJD ICE FPA n 2010-2012.

References

[1] Rauterberg, M.: Emotions: The Voice of The Unconscious. Entertainment Computing - ICEC 2010. pp. 205–215 (2010).

[2] Lang, P.J.: Cognition in emotion: Concept and action. In: Izard, C.E., Kagan, J., and Zajonc, R. (eds.) Emotions cognition and behavior. pp. 192-226. Cambridge University Press, New York (1984).

[3] Mandryk, R., Atkins, M.: A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. International Journal of Human-Computer Studies. 65, 329-347 (2007).

[4] Frijda, N.H., Ortony, A., Sonnemans, J., Clore, G.L.: The Complexity of Intensity: Issues Concerning the Structure of Emotion Intensity. Emotion. 13, 60-89 (1992).

[5] Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology. 39, 1161-1178 (1980).

[6] Rafaeli, E., Revelle, W.: A premature consensus: are happiness and sadness truly opposite affects? Motivation and Emotion. 30, 1-12 (2006).

[7] Chang, H.-M., Ivonin, L., Chen, W., Rauterberg, M.: Multimodal Symbolism in Affective Computing: People, Emotions, and Archetypal Contents. Informatik-Spektrum. In press.

[8] Reisenzein, R.: Pleasure-arousal theory and the intensity of emotions. Journal of Personality and Social Psychology. 67, 525-539 (1994).

[9] Fisher, N.I.: Statistical Analysis of Circular Data. Cambridge University Press (1996).

[10] Ribeiro, R.L., Teixeira-Silva, F., Pompéia, S., Bueno, O.F.A.: IAPS includes photographs that elicit low-arousal physiological responses in healthy volunteers. Physiology & Behavior. 91, 671-675 (2007).

[11] Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Affective ratings of pictures and instruction manual. NIMH, Center for the Study of Emotion & Attention (2005).

[12] Goldin, P.R., Hutcherson, C.A.C., Ochsner, K.N., Glover, G.H., Gabrieli, J.D.E., Gross, J.J.: The neural bases of amusement and sadness: a comparison of block contrast and subject-specific emotion intensity regression approaches. NeuroImage. 27, 26-36 (2005).

Improved Huffman Algorithm in Multi-channel Synchronous Data Acquisition and Compression System

Ma Xian-Min[1,a]，Zhou Gui-Yu[2,b]
[1]Xi'an University of Science & Technology，Shanxi Xi'an 710054, China
[2]Xi'an University of Science & Technology，Shanxi Xi'an 710054, China
[a]e_mail: maxm@xust.edu.cn ，[b]shambhalary@163. com

Keywords： Huffman algorithm; Data acquisition; Data compression; LZW algorithm; 120 emergency treatment ambulance command system

Abstract. An improved Huffman algorithm is proposed in order to take up as litle as storage space during the process of transmiting data remotely and transmitting more useful information in the limited channel capacity in 120 emergency treatment ambulance terminal system. The multi-channel simultaneous data acquisition and compression system based on DSP and FPGA technology are designed ,which are composed of multi-channel data acquisition module, DSP data processing module, ambulance interface module. The research result indicates that the presented Huffman algorithm can reduce the requirement power and bandwidth in the compressing data technology and raise communication efficiency compared with LZW algorithm. Therefore the multi-channel data acquisition and transmission rate are improved though the DSP embedded data compression algorithm, and the system performance is safety and reliable for 120 first aid dispatch and command system.

Introduction

With the development of science and technology and the universal applications of data acquisition technology, many data acquisition system's technical specifications have been demanded increasingly such as sampling rate, resolution, memory depth, the rate of disposing digital signal, interference and so on. In many applications, ultra-high-speed data acquisition system is required to complete some work which a number of low-speed data acquisition system could not be done[1]. The system has designed multi-channel data acquisition and compression system to be designed to address these requirements, which multi-channel simultaneous data acquisition can be implemented impersonation signal band-pass filtering, amplification, conditioning function, the related signal synchronization collection, after analysis obtained the requirements of relevant information, the DSP processing module can meet the data processing, getting data compression, reducing the power and bandwidth requirements, improving communication efficiency, bus interface module enables to making the treated data through the serial bus communication with various external interrupt This design of multi-channel synchronous data acquisition and compression system can collect data with good synchronization, occupy transmission time and storage space as little as possible and transmit more useful information in the limited channel capacity with the modularity of functions, standardization of interfaces and it' s also with the flexible features according to the functional requirements.

System Design

FPGA and DSP-based multi-channel synchronous data acquisition and compression system is mainly assembled of data acquisition module、data processing module and the data interface module. Firstly, the input analog signal via the sensor realizes condition and amplification through signal conditioning circuit and converts to input signal which is suitable for A / D converter. A / D converter converts the analog signal to digital signal which is controlled by Field Programmable Gate Arrays1 (FPGA1). FPGA1 collects the processed data according to the set sampling rate and puts the collected data writing to FIFO1 via the bus. When FIFO1 is half-full, a half full signal will be sent to trigger the DSP interrupt, if DSP receives the interrupt response, it will make a data read into data processing module from the FIFO1, at the same time, DSP compresses the collected data writing into the interface module and sends to the monitoring system for further processing through the Field Programmable Gate Arrays2 (FPGA2) via RS-422 bus [2]. The flow chart of design ideas is shown in Fig. 1.

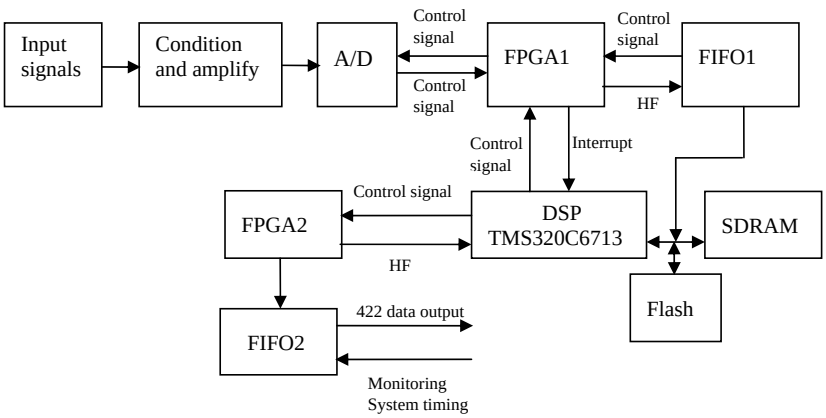


Fig. 1 The flow chart of system overall

Firstly, the system initializes the CSL library of CSL, PLL, GPIO and interrupt-related registers after powering on and waits for the arrival of the interrupt signal. FPGA in acquisition module controls AD converter to make the processed signal write into FIFO, when FIFO data has half full signal, HF signal generates an interrupt signal on the DSP TMS320C6713B. When DSP enters the interrupt, it makes the length of frame data read into memory (SDRAM). After processing the interruption, DSP compresses the data at the high-speed, and then writes the processed data into the soft FIFO [3]. The workflow chart of TMS320C6713 processing module is shown in Fig. 2.

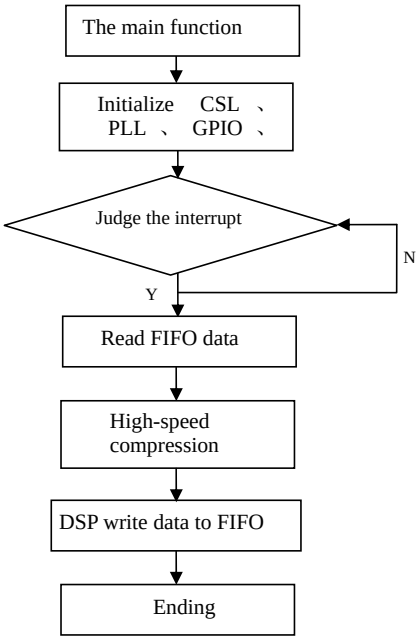


Fig. 2 The processing module workflow chart of TMS320C6713

Multi-channel Data Acquisition module

The program starts AD analog to digital conversion by controlling HOLD of AD and controls the AD data output mode through controlling AD Address. The sampling frequency of single-channel could be up to 3000 kHz.

This design has taken the methods to make FPGA1 read data from AD, continued writing into high-speed FIFO1. DSP determines whether the FIFO1 has half full signal, if it's true, DSP will read FIFO1 data volumely for dealing with the mismatching between high-speed data flow requirements and the specified sampling rate which is the lower speed data stream. Program-control uses Verlog HDL language which is the hardware description language for logical design and has become the IEEE standard [3,4]. The ModelSim simulation timing diagram of the program main signal is shown in Fig. 3.

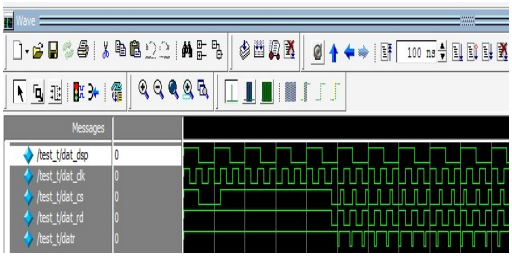


Fig. 3 ModelSim simulation timing diagram of Program main signal

Data Compression Algorithm

The lossless compression is that compression coding source information can be recovered fully when extracted. Shannon-Fano coding, Huffman coding, arithmetic coding, dictionary coding are all

lossless compression methods which are used frequently [5]. This design mainly compares the LZW algorithm with Huffman algorithm.

Data compression's basic theory is information theory, entropy can be used to measure the size of the amount of information. The average information entropy of Source is given by Eq. 1 [5]:

$$H(X) = -\sum_{i=1}^{N} P_i * \log(P_i) \quad (1)$$

Where its unit is bpp, it is a measure of the uncertainty of random source. P is the probability of each corresponding symbol. Shannon said that the best compression performance of the lossless compression algorithm is that the source coding average bits is equal to the entropy of the source for the discrete sources which meets the conditions of definition of the entropy, there is no lossless compression algorithm which the coding's average bits is less than the source entropy. Therefore, the entropy is the lower limit of no memory discrete source and lossless compression. For continuous random variables X's differential entropy is given by Eq. 2 [5]:

$$H(X) = -\int_{-\infty}^{+\infty} P_X(x) \log^{p(x)} dx \quad (2)$$

The amount of information refers to the needed information which can choose one from many equal possible events. Assumely that the probability of being selected a number X from numbers is P(X), if the probability of any number of being selected is equal, the amount of information is given by Eq. 3 [5]:

$$I(x) = -\log_2^N = -\log_2^{\frac{1}{N}} = -\log_2^{p(x)} = I[P(X)] \quad (3)$$

The method of measuring mounts of information in information theory is given by Eq. 4 [5]:

$$I(x_j) = -\log_a^{p(x_j)} \quad (4)$$

Where P(Xj) is Xj's prior probability from information source X. I(Xj) is called the information quantity after the occurrence of Xj.

In general, different compression algorithms have different strengths and weaknesses, different complexity of the algorithm on the space requirements and compression ratio is also different. This is not only dependent on the compression method, but also be related to the characteristics of the compressed data.

The performance of Huffman compression algorithm compared and LZW compression algorithm is shown in Table1.

Table1. The Comparison Between The LZW Algorithm And Huffman Algorithmin Performance

Algorithm	Real-time	Complexity	Storage area	Compression rate	Applicable occasions
LZW	better	general	generally smaller	better	Any data
Huffman	need pre-treating	general	larger, need restoring data	very good	Any data, especially suitable for text data

LZW is a compression algorithm with a simple logic, fast speed, cheap hardware implementation, which is widely used in compressing and packing to documentary. The applicable scope of LZW algorithm is that the original data string is better to have a lot of strings appeared repeatedly and the more them repeat, the better their compressive effects are. Otherwise, the effect of compression is worse. The flow chart of LZW algorithm is shown in Fig. 4.

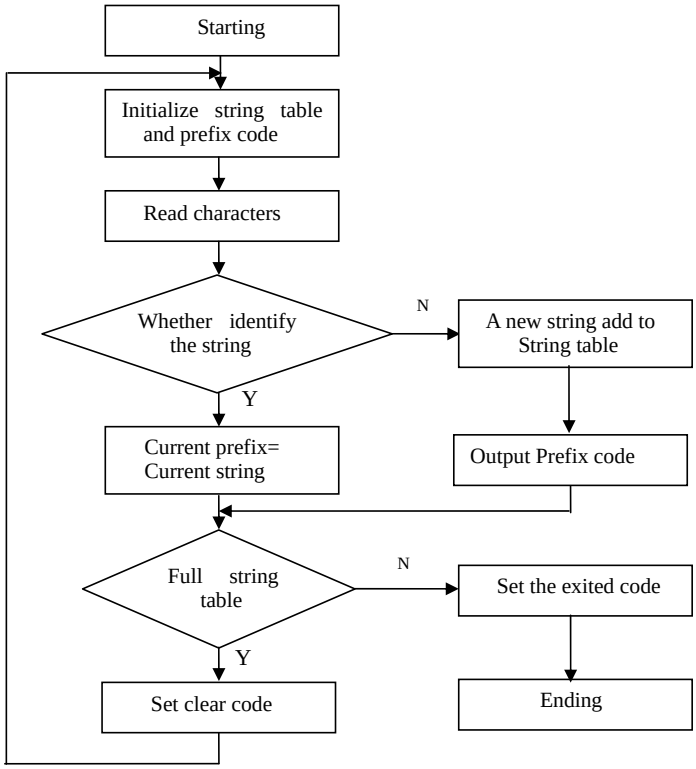


Fig. 4 The flow chart of LZW algorithm

The basic principle of Huffman is based on the probability of the source characters being appeared to construct the code. The output code of the encoder is words which the word-length is unequal in coding of variable word length. It is assigned different word length to output code words according to the probability of the input information symbols. As to the source character which owns larger probability appearence can get a shorter code length, while, it will gets a longer code length, finally makes the average of coding code words shortest. It can be proved that, using this method can be made the output code shortest to the average code length, closest to the source entropy, best to the coding method. Two issues need noting when using Huffman encoding .One is that Hoffman coding has no function with protecting error, the code can be translated correctly one by one if no error in code strings in the process decoding, otherwise, the wrong codes can not be corrected by compute. On the other hand, it is difficult to find or call the contents among the compressed files freely to decode because Huffman codes are variable-length code, so it needs to be more considered before storing code. Because of Huffman codes' absolute superiority in the file compression, Hoffman code is widely used despite of these Shortcomings [6]. The flow chart of Huffman algorithm is shown in Fig. 5.

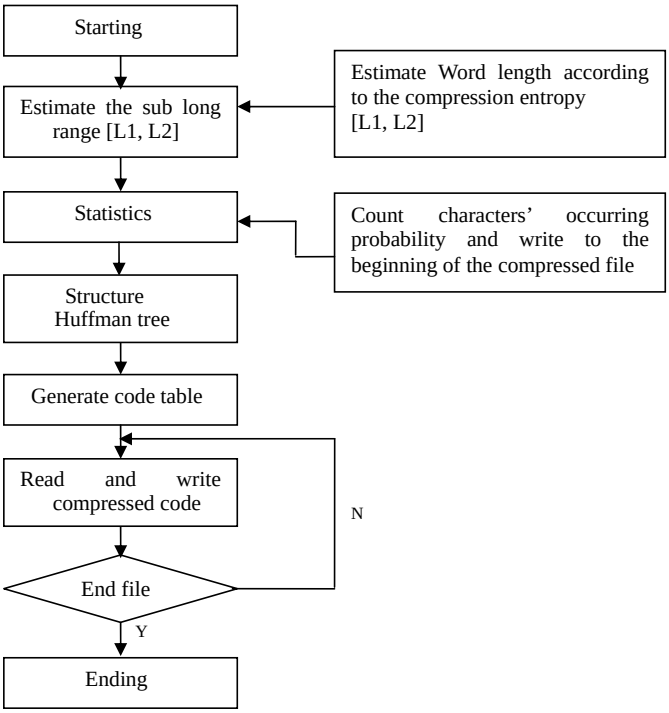


Fig. 5 The flow chart of Huffman algorithm

The selection of algorithm is mainly compared the compression speed and compression removal rate [7].From Table2, Huffman compression algorithm has a higher compression rate and is applicable to compress of any data (especially suitable for text data).Since this system's subjects are basically the terminal GPS data's compression, the collected car's basic information and the patients' information, in this condition, using text files to storage is feasible. In the experiment, the lower machine sends regular data to be compressed, and then these compressed data are conducted to be unpackaged and decompressed through the PC. If the wrong structures of frame are detected in the process of decompressing, the decompression system will prompt error because abnormalities structures of the frame can not be extracted normally. If the frame is correct, then the decompression system calls the subfunctions which have been prepared in dynamic link libraries. The result of comparing the restored data to the original data is consistent, proving that the system is safe and reliable. Through the experiment, it proves that Huffman compression algorithm has a higher efficiency in compressing text data. Comparing the two algorithms, we choose Huffman coding to compress data. The comparison of the rate of compressing data between the LZW algorithm and Huffman algorithm are shown in Table2.

Table2. The comparison of the rate of compressing data between the LZW algorithm and Huffman algorithm

LZW algorithm			
size of source files	size after pretreating data	the size after coding	compressed rate
427.56KB	212.33KB	77.56KB	81.89%
1.25M	512.55KB	215.05KB	82.21%
25.78M	12.34M	4.51M	82.53%
45.67M	22.74M	7.96M	82.59%
112.36M	55.94M	19.58M	81.97%

Huffman algorithm			
size of source files	size after pretreating data	the size after coding	compressed rate
427.56KB	299.24KB	281.68KB	34.12%
1.25M	875.34KB	826.75	33.86%
25.78M	18.86M	16.95M	34.25%
45.67M	32.48M	30.13M	34.33%
112.36M	80.17M	74.12M	33.95%

Summary

This article presents the improved Huffman algorithm in the data compression process of the multi-channel simultaneous data acquisition and compression system for the 120 emergency treatment ambulance terminal system, which collects a lot of date simultaneously using the method of combining field programmable gates array Virtex-5 FPGA with digital signal processor (DSP) to reduce the power consumption. A lot of experiments confirm that the improved Huffman algorithm compared with LZW compression algorithm is feasible, and all the technique specification meets the system requirements.

References

[1] SongFeng, SunWei. The design of data acquisition system based on FPGA [J]. Information Technology, 2009,10(3):23-25.

[2] Zhu Zhen-yong, and Weng Mu-yun. Design and Application of FPGA.Xidian University Press, 2002

[3] J. Mcallister, R. Woods, R Walk, D. Reilly. Multidimensional DSP Core Synthesis for FPGA [J]. The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 2006, 43(2-3).

[4] A. Yu. Bulgakov,V. N. V'yukhin,Yu. A. Popov. A 24-Bit Data Acquisition System [J]. Instruments and Experimental Techniques, 2001, 44(2).

[5] P G Howard, J S Vitter. Analysis of arithmetic coding for data compression [J]. Information Processing and Management,1992,28(6):13-15

[6] I H Witten, R M Neal, J G Cleary. Arithmetic coding for data compression [J]. Comm. ACM, 1987, 30(6):11-13

[7] Takuya K, Tetsuya M, Yusuke S,et al. Collage system: a unifying framework for compressed pattern matching [J] .Theoretical Computer Science, 2003, 298:253-272.

Affective Smart City: A First Step for Automatic Governance

Francesco Rago[1, a], Stefano G. Rago[2,b] and Alberto Panico[2,c]

[1]Megatris Comp. LLC, Newark, DE,US

[2] Università degli Studi di Napoli Federico II, Italy

[a]francesco.rago@megatris.com, [b]stefano.rago@megatris.com, [c]panico87@gmail.com

Keywords: affective computing, transition systems, weak bisimulation, sentiments' network, smart city.

Abstract. Affective Smart City approach can be enlarged to bigger systems where the affective status of the system is larger but always related to human support. A city can be considered as a human aggregate that has feelings as a result of feelings and behaviors of all inhabitants. Having an emotional management, an affective city could activate sets of resources, specialized to recognize some condition. Thus, Anger and Fear permit a focus on defense and protection, while Hunger and Thirst evolved for nutrition. We start constructing a symbolic or finite abstraction of event driven real-world in the form of a finite transition system. We define an associated affective transition system with weak bisimulation. We can refine the discrete affective supervisor into a hybrid controller that will enforce the desired behavior depending on the prevalent sentiment. The prevalent affective status formation is based on a specific "sentiment leader" derived by a sentiments' network composed of elements with non-identical but adaptive dynamics. Such leader indicates "what to do" because it is associated to a control model.

Introduction

Affective computing was always considered as an approach to support single human needs to understand sentiments. This approach can be enlarged to bigger systems where the affective status of the system is larger but always related to human support. A city can be considered as a human aggregate that has feelings as a result of the feelings and behaviors of all inhabitants.

A computer system can be used to collect all the data generated by economic and social events. We can associate affective computing statuses to specific aspects of social life. Transitions from real-world statuses to a computer network affective status permit a higher level view of collective behaviors that can be regulated using the triggering of control models. We have used results from different fields to demonstrate that similar systems do not need a new conceptual formalization but just a careful integration of concepts and theoretical tools.

One definition of affective computing. The Affective Computing is a type of system that has the technical capability of emotional intelligence. It is able to use emotions in a logical and intelligent way. It directs action as a function of emotional state. When we talk about emotions related to a machine mean something very different from human consciousness.

A computer is aware of the perception of the body in different ways because the network of sensors and actuators is different from human's net. Psychologists and neurologists know that emotions are both cognitive and physical. They involve thoughts and body sensations. Computer's emotions are a set of mechanisms that make computer aware of its conditions. In this case, a computer may take an

autonomous behavior and change its programming according to the preferences or the needs of a person ora group. It changes internal rules to achieve the assigned goals.
So, if a group does something that is not acceptable, because of group's disappointment, computer should be able to adjust its behavior accordingly. In order to do this it must be able to understand, or to have models of, effects of social distress. This can be made by associating real-world measured states to affective internal states.

Basic Emotions. Following Picard [1], the most common four emotions are: fear, anger, sadness, and joy. Plutchik [2] distinguished among eight basic emotions: fear, anger, sorrow, joy, disgust, acceptance, anticipation and surprise. Although the precise names vary, the two most common categories are "arousal" (calm/excited), and "valence" (negative/positive). Another category [3] can called "control" or "attention" addressing the internal or external source of the emotion. For affective computing, the recognition and modeling problems are simplified by the assumption of a small set of discrete emotions, or small number of categories, even if it is not necessary such feelings are equivalent of human sentiment. We think it is necessary to avoid an anthropocentric approach.

Smart City and affective computing. Smart cities can be identified along six main axes or dimensions [4] [5][6] where the more meaningful is the smart governance. The concept of smart city has its main focus not only on the role of ICT infrastructure, but also it has been carried out on the role of human capital/education, social and relational capital and environmental interest as important drivers of urban growth. To describe a smart city and its characteristics it is necessary to develop a set of characteristics defined by a number of factors. Furthermore each factor is described by a number of indicators. Having in mind the automation of governance, it is possible a ranking if targets are defined well and a transparent structure is used. The objective of this ranking is to compare characteristics and to identify strengths and weaknesses of a city at run time. Therefore it is not useful to focus solely on the performance of one aspect of city development, but to focus on the performance of a broad range of integrated characteristics.

A forward-looking development approach should consider issues as awareness, flexibility, transformability and synergy. Awareness seems especially important for a smart city, certain potentials can be mobilized if inhabitants, companies or the administration are aware of the city's status. To support this knowledge status and to guide a flexible and rapid active response, the city management system should possess self-awareness. If we look at the definition of awareness [7]: "Awareness is the state or ability to perceive, to feel, or to be conscious of events, objects or sensory patterns. In this level of consciousness, sense data can be confirmed by an observer without necessarily implying understanding. More broadly, it is the state or quality of being aware of something. In biological psychology, awareness is defined as a human or an animal perception and cognitive reaction to a condition or event." Often emotions, not thoughts, motivate us. Without awareness of feelings, it's impossible to fully understand our own behavior, appropriately manage emotions and actions, and accurately "read" the wants and needs of others [8]. In biological sentient beings like humans, emotions determine motivation, not thoughts. Without awareness of what they are feeling, it's impossible to fully understand their behavior, appropriately manage our actions. Emotions are the glue that gives meaning to life and connects them to other people. They are the foundation of ability to understand themselves and relate to others.

Having an emotional management an affective city can activate sets of resources [9] "while turning certain others off". Minsky calls such resources "Critics", each of which is specialized to recognize a specific condition and then to activate a specific collection of other resources. Thus, Anger and Fear permit a focus on defense and protection, while Hunger and Thirst evolved for nutrition improving the parameters related to distribution nets. We can devise a new step in systems paradigm: the *Affective Smart City*.

The recognition of city's status can generate affective reactions. The measurement of sentiment modulation can be valued by general trend of economic and social indicators and valuing the statistical behavior of city's citizen. The proposal is not related to the general status of citizens but rather to measure observable clusters of such states.

Events as consequences of collective behavior. To have an idea of persons' behavior and statistical emotions, we are interested in detecting certain states of the real world and for this reason the notion of events is relevant to our problem, since events can be considered as state transitions triggers. Giving an initial real-world state, certain sequences of events may lead to a target state that has to be detected. Hence, the detection of interesting real-world states is closely related to the detection of event patterns. Event notifications can be considered as software representations or real-world events. An event notification typically consists of a type specification and of an arbitrary number of parameters. Each parameter has a name, a type and a value. The occurrence of a physical event and the resulting delivery of an event notification triggers computation to evaluate or to react to the change in the real world. The detection of real-world states with sensor networks can be supported by assuming a simple model where each sensor node can assume states depending on the output value of the sensors attached to it. Using these states, a transition system can be constructed to specify state transitions.

Let us assume that the state transitions are mapped to event notifications that contain a node identifier as a parameter. Whenever the sensor reading on a node happens, a corresponding event notification is emitted. For the specification of complex states involving multiple events, a different abstraction can be used where complex real-world states (involving multiple sensor nodes) can be conveniently compared to traditional approaches for composite-event detection.

The recognition of complex events can be associated to affective status. Various predicates allow the specification of constraints on the actual states of the involved sensor nodes. Each state sentiment specification is accompanied by a set of actions or a control model that is executed whenever the actual configuration of sentiment matches the state specification. A sentiment is related to a different view or interpretation of world and this is associated to a specific control model.

Action specifications can take a number of different forms, such as a control model in a specific programming language, or the specification of an event notification that should be generated when a match occurs.

If we suppose that sentiments are organized as a network composed of peers, the phase of the collective behavior is hard to predict, since it depends on the initial conditions of all the coupled elements. To let the whole network converge to a specific trajectory "behavior", a "leader" sentiment can be identified. The leader is an element whose dynamics is dominant in a time frame and thus primary respect to the other sentiments until next transition. This transitory leader specifies "what to do" and which specific control model to apply during a time frame.

A formalization of affective Smart City. We start constructing a symbolic or finite abstraction of the event driven system in the form of a finite transition system. When the provided abstraction is correctly constructed, we can refine the discrete affective supervisor into a hybrid controller that will enforce the desired behavior on the original control system.

The heart of the approach lies in the construction of the finite affective abstraction, which is the central theme of this paper. The paper investigates the properties of affective status formation that is based on a specific "sentiment leader" [11].

Synthesis of control software is regarded as the synthesis of a real-world supervisor using a transition systems.

Transition Systems. We can emulate the behavior of real-world using transition systems [11]. Each recognized event generates a transition in the system. The last is associated to another transition system which represents the point of view of feelings. Sentiments are connected in a network and the last transition defines the sentiment with a label indicating the control model associated to such status.

Definition: A transition system T is a quintuple (Q,L,→,O,H) consisting of:

- a set of states Q ;
- a set of labels L;
- a transition relation →⊆QxLxQ;
- an output set O;
- an output function H:Q→O.

A metric transition system is a transition system in which the output set is equipped with a metric d:OxO→R+.

Definition: A run of a transition system T is a string r ∈ Q* for which there exists l ∈ L* satisfying $r(i) \xrightarrow{l(i)} r(i+1)$ for i=1, …, |r|-1. A string l∈ O* is said to be an output run of if there exists a run r of T such that H=s. The language of T, denoted by L(T) , is the set of all output runs of T. The mapping from one transition system to another can be defined using the concept of equivalence. Simulation and bisimulation relations are standard mechanisms to relate the properties of transition systems [12]. Intuitively, a simulation relation from a transition system T_1 to a transition system T_2 is a relation between the corresponding state sets explaining how a run of r of T_1 can be transformed into a run of T_2. Generally the concept of bisimulation equivalence is used in process algebra, but we need a weaker notion: the weak bisimulation.

Definition: A binary relationship R⊆Q1xQ2 is a weak bisimulation if (q_1, q_2), $(q'_1, q'_2) \in R$ and $q_1 \xrightarrow{1} q'_1$ and imply the existence of $q_2 \xrightarrow{2} q'_2$.

We can introduce metrics to permit software evaluation and we shall relax definition by requiring to simply being close where closeness is measured with a metric on the output set.

Definition: Let T_1=(Q1,L1,$\xrightarrow{1}$,O1,H1) and T_2=(Q2,L2,$\xrightarrow{2}$, O2,H2) be metric transition systems and let $\varepsilon, \vartheta \in R_0^+$. A relation R⊆Q1xQ2 is said to be a let (ε, ϑ) – qualitative approximate simulation relation from T_1 to T_2 if:

1) $\varepsilon, \vartheta \in R$ implies $d(H_1(q1),H_2(q2)) \leq \varepsilon$;

2) $d(H1(q_1),H2(q_2)) \leq \vartheta$ implies $(q1,q2) \in R$;

3) (q_1, q_2), $(q'_1, q'_2) \in R$ and $q_1 \xrightarrow{1} q'_1$ and imply the existence of $q_2 \xrightarrow{2} q'_2$;

where d is a function $Q_1xQ_2 \rightarrow R$ meaning the distance between sentiment (meta) status and events status.

The adequacy of the notion of approximate simulation relation has its characterization in terms of known stability concepts and its essential role in the symbolic control methodology [10][13].

We can assume that T_2 is a sentiment transition system, it describes the status on the system itself. The status focuses on the behavior to apply depending on the evolution of events in time.

The transition from one status to another cannot be linear, but depends from the global values. The transition for each status delivers a temporal "winner" that apply a control model for the appropriate interaction with the real world.

The transition at sentiment level is considered as a "mood change" depending on all other status. During a specific amount of time the system will be focused on a specific sentiment. It can be considered as result of the dynamics of a coupled network containing one "winner leader" and n followers [11].

A sentiments network is composed of elements with non-identical but adaptive dynamics. A leader may be located in any position inside a network. Its dynamics are fixed or slowly changing in a specific time interval, while those of the followers are adapted to the leader. Such leader indicates

"what to do" in that specific time frame because it is associated to a control model. Calculation is make again in the following time interval and leader may change.
Synchronization or group agreement can still be achieved in such a network with only local interactions.
Consider a coupled network containing n elements

$$x_i(t+1) = f(a_i, x_i(t)) + \sum_{j \in N_i} K_{ij}(x_j - x_i) \quad (1)$$

$x_i$is the state of the ith sentiment. We assume that the uncoupled dynamics f (x_i,a_i,t) is continuous, smooth and identical to each element except the value of a parameter set a_i, which is different for each participant. For notational simplicity, the coupling forces are set to be diffusive, where all coupling gains are defined symmetric positive, and the couplings between the followers are bidirectional with $K_{ij} = K_{ji}$ if both $i, j \neq 0$. N_i denotes the set of peer-neighbors of element i, which for instance could be defined as the set of the followers within a certain distance around element i. Thus N_i can be defined arbitrarily.
Denote Ω as the set of the sentiments, whose adaptation laws are based on local interactions dynamics, which is assumed to be identical for each element.

$$a_i(t+1) = P_i W^T(x_i, t) \sum_{j \in N_i} K_{ij}(x_j - x_i) \forall i \in \Omega \quad (2)$$

Where $W\ (x_i, t)$ is defined as:

$$W\ (x_i, t) = f(a_i, x_i(t+1)) - f(a_i, x_i(t)) \quad (3)$$

The states of all the elements will converge together asymptotically if appropriate conditions hold [12]. The convergence guarantees the consistence of the process even if the convergence will not be reached for time reasons. The behavioral consequence is the choice of each control model defined by an association between each sentiment status and a specific model. The choice at time t will be determined by:max $\{x_i(t)\}$. The associated control model can be defined as follow:

Definition: Let T_1=(Q1,L1,$\overset{1}{\rightarrow}$,O1,H1) the real-world transiction system and T_2=(Q2,L2,$\overset{2}{\rightarrow}$, O2,H2) the affective transition systems with weak bisimulation, a control system $\sum = (Q, U, \mathcal{U}, f)$ associated withΣis defined by:

- Q =Q1xQ2 ;
- U is a compact subset of Q ;
- $\mathcal{U}$ is a set of measurable functions from intervals of the form $]a, b[\subseteq R$ to U;
- f: R^nxU → $R^n$is a continuous map satisfying the Lipschitz assumption.

A C_1 curve x: $]a, b[\rightarrow R^n$ is a trajectory of Σ if $\exists u \in U$ such that x(t+1)= f(x(t),u(t)).
We assume that the control system is uniformly globally stable [14]. The adequacy of the introduced notions will be justified with characterization in terms of known stability concepts and its essential role in the proposed symbolic control methodology.

Theorem: Let be a control system satisfying *Stabilizability Assumption*, then, exists ε, ϑ-approximate simulation relation from T_1 to T_2.
The demonstration is included in [14].

Experimental results. Using the models management of Scorpius System Bus [15] we have created the following simulative situation. By using a set of events and a set of indicators about Pollution, People mobile communication and Viability it was possible to simulate the affective systems response. An affective error is measured in the choice of appropriate controls models and their

parameters. We have classified the events that SSB can collect: about 500 different simple and complex events with a class directed to collect citizens mobile communications to the systems. A T_s transition system was built to simulate real-world environments. Some smart cities indicators [6] where used as states and calculated at each time frame. The affective transition system was defined using Hidden Markov Chain during a learning previous phase: the transition systems is based on thirty-two machine sentiments statuses. Each transition calculation happens with the same time frame. Weak bisimulation relation between the transitions systems was built with a neural net during a learning previous phase.

According to results, is clear that a machine is more error prone than an optimization solution, even if we suppose possible citizens could benefit from "errors". New research is necessary to value the response on larger number of events to confront classical controls with affective smart city.

Table 1. Experimental results:

Events Type	Events instances	T_s transitions	T_A transition	Decisions Error
Pollution	400	235	21	3%
People mobiles	1,000	126	25	2%
Traffic	456	230	31	1,2%

Conclusion

If we provide an affective mechanism to answer to external stimuli in a more effective way we will help big systems like cities to be more focused on citizen needs. The choice of a control model related to a "city mood" status permits to be more confident about the quality of service, where and how long, a city demands. The behavior is not reactive, even if it can be, but is more oriented to a balance of sentiment provided by a synchronization and group cooperation algorithm.

References

[1] Picard, R. W.: Affective Computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321,Perceptual Computing; 20 Ames St., Cambridge,1995.

[2] R. Plutchik: *A general psychoevolutionary theory ofemotion*, In: Emotion Theory, Research, and Experience (R. Plutchik and H. Kellerman, eds.), vol. 1, Theories ofEmotion, Academic Press, 1980.

[3] N. L. Stein, K. Oatley: Basic Emotions. Hove,Lawrence Erlbaum Associates, 1992,double issue of the journal Cognition and Emotion, Vol. 6, No. 3 & 4, 1992.

[4] R.L. Stein, F. Fertner: Smart cities – Ranking of European medium-sized cities. http://www.smart-cities.eu/. Vienna: Centre of Regional Science.

[5] N. Komninos: Intelligent cities: innovation, knowledge systems and digital spaces. London, Spon Press.

[6] Smart cities final report: http://www.smartcities.eu/download/smart_cities_final_report.pdf

[7] Information on http://en.wikipedia.org/wiki/Awareness.

[8] R. L. Segal, F. Fertner: http://helpguide.org/toolkit/developing_emotional_awareness.htm

[9] Minsky, M.: The Emotion Machine, Simon & Shuster Ed., 2006.

[10] W. Wei, J.-J.E. Slotine: *A theoretical study of different leader roles in networks*, IEEE Transactions on Automatic , July 2006 , Volume: 51 Issue:7.

[11] P. Tabuada: *An Approximate Simulation Approach to Symbolic*, IEEE Transactions on Automatic , July 2008 , Volume: 53Issue:6.

[12] E. M. Clarke, O. Grumberg: Model Checking, Cambridge, MA, MIT Press, 1999.

[13] M.Gelfond, V. Lifschitz: *Action Languages*, Linköping Electronic Articles in Computer and Information Science, vol 3, nr 16.

[14] P. J. Antsaklis,V. Lifschitz: Linear Systems. New York: Mc-Graw-Hill, 1997.

[15] F. Rago: SSB, http://www.megatris.com/?q=system/files/SSBArchitectureV1.0.pdf,2011.

Solving the Airlines Recovery Problem Considering Aircraft Rerouting and Passengers

Meilong Le[1, a], Chenxu Zhan[2,b] and Congcong Wu[3,c]

[1]The Scientific Research Academy, Shanghai Maritime University, Shanghai, P. R. China

[2] The Scientific Research Academy, Shanghai Maritime University, Shanghai, P. R. China

[3] The Scientific Research Academy, Shanghai Maritime University, Shanghai, P. R. China

[a]meilongle@hotmail.com, [b]zhanchenxu@hotmail.com, [c]wcc131208@hotmail.com

Keywords: aircraft recovery; passenger recovery; airlines operation; integrated recovery

Abstract. Severe weather conditions, air traffic control and mechanical failures often disturb a flight schedule. To deal with these disruptions, airlines must modify their schedules. Modifications include aircraft, crew and passengers recovery. We present a new method of modeling. The method transforms the airlines recovery problem into a Vehicle Routing Problem with a time window. And we propose an optimization model considering the aircrafts and passengers. By means of model solving, we get new aircraft rotations. Finally, the data from a domestic airline was used to test our model.

Introduction

The execution of flight schedule is important to airlines, as well as to passengers. During the daily of airlines, disruptions caused by severe weather, air traffic control and mechanical failures often prevent airlines from executing their schedules as planned. If the disruptions don't be dealt effectively, it may results in additional cost and loss. The airlines must handle the disruptions by rescheduling flights within a time window and determining which flights should be delayed or canceled. In general, airlines recovery divides into three stages, aircraft recovery, crew recovery and passenger recovery.

The recovery of passenger must ensure the feasibility of the aircraft and crew schedules. Compared with the cost of aircraft or crew, that associate cost with passengers is low. Hence, airlines concerned more on the recovery of aircraft and crew, and they always consider the passenger recovery in the integrated recovery problem. In the past three decades, many papers are presented for airline recovery. Some of them focus on single recovery problem such as passenger recovery, aircrafts recovery or crew recovery. But more and more scholars devote themselves into the study of integrated recovery problem in recent years.

Lettovsky (1997) [1] presents a Passenger Flow Model (PFM) for the passenger recovery problem(PRP)which is a sub-model of an integrated recovery model. The objective function of PFM is maximizing the passenger revenue. The author divides the model into three stages. In the first stage, the author put the passengers, whose itinerary was same, into a set. Determining possible flight path of the network is performed in the second stage. Finally, based on the previous two stages, he applies an optimization model to determine the optimal allocation of passenger seats.

Bratu and Barnhart (2006) [2] propose two models for PRP which are Disrupted Passenger Metric model (DPM) and Passenger Delay Metric model (PDM). Zhang and Hansen (2008) [3] focus on the hub-spoke network, and present a model for solving the PRP. They introduce ground

transportation as the alternative mean of passengers' recovery. The model determines the flights which have to be cancelled or replaced. The authors emphasize that the model is merely appropriate for the condition that the demand keeps saturated in a certain period. Bisaillon et al. (2010) [4] consider the PRP in aircraft recovery problem and present a space-time network converting from the flight schedule. The target in their paper is minimizing the total cost containing of the operation cost and the impact to passengers. Eggenberg et al. (2010) [5] provide a constraint-specific recovery network model which could apply in aircraft, crew and passenger recovery problem respectively. They present a passenger-specific network for PRP. The model is introduced, and based on an assumption that the passengers should finish their itineraries within a maximum delay. The objective function is transporting all passengers to their destinations while minimizing the delay cost and cancellation cost.

Jafari et al. (2010b) [6] extends the model from Jafari et al. (2010a) [7] for solving the aircraft recovery problem considering passengers. The aircraft rotation is the object of their study and it reduces the size of the problem. Meanwhile, the authors introduce maximum delay (for aircrafts), minimum connecting time (for passengers), and minimum turn-around time (for aircrafts). The function minimizes the cost contained operational recovery cost, flight cancellation cost and delay cost of disrupted passenger.

In this paper, we focus on aircraft and passenger recovery problem. We propose a model named Passenger-Aircraft Recovery Model (PARM) and using a Heuristic Algorithm to solve it. In section 2, we describe the airline recovery problem. In section 3, the method and model is presented. In section 4, we give some instances and present computational results from solving our model; the analysis is given, as well. The conclusion and further research are described in section 5.

2 Method and modeling

2.1 Modeling Method

We view the recovery problem as a complicated vehicle routing problem (VRP) with a time window and modeled according to the requirement of VRP. Aircrafts are vehicles while airports are nodes, and passengers are commodities. For each ACR, we view it as a route. The lower bound of time window is the scheduled arrival time, and the upper is the scheduled arrival time plus a maximal delay.

2.2 Mathematical Model

Our target is to minimize the total delay cost. To deliver the passengers, the aircraft must arrive the airports as planned. According to the requirement of our method, an aircraft is forbidden to access the same airport twice except the starting airport. We define the parameters and variables as follows.

AC	set of airports;
HUB	set of hub airports;
A	set of regional airports, $AC = H \bigcup A$;
K	set of aircrafts;
T_{ij}	the flying time between airport i to airport j, $i, j \in A$;
D_i	the number of passengers who depart from the hub and ended in airport i;
CAP_k	the capacity of aircraft k;
T_{limit}	the limit of longest flying time;
a_{ki}	the actual time of aircraft k arrives at airport i;
e_{ki}	the scheduled time of aircraft k arrives at airport i;
C_i^{delay}	the delay cost of airport i;
N	the number of aircraft;

M infinity;

x_{ijk} 1, if the aircraft k flying from airport i to airport j, otherwise, 0;

Mathematical formulation

$$Min \sum_{k\in K}\sum_{i\in AC} C_i^{delay}(a_{ik}-e_{ik}) \quad (1)$$

$$\text{s.t.} \sum_{j\in A}\sum_{k\in K} x_{ijk} \le N, \forall i \in HUB \quad (2)$$

$$\sum_{j\in A} x_{ijk} \le 1, \forall i \in HUB, \forall k \in K \quad (3)$$

$$\sum_{i\in AC} x_{iik} = 0, \forall j \in A, k \in K \quad (4)$$

$$\sum_{j\in A} x_{ijk} = \sum_{j\in A} x_{jik}, \forall i \in HUB, k \in K \quad (5)$$

$$\sum_{j\in AC} x_{ijk} = \sum_{j\in AC} x_{jik}, \forall k \in K, i \in A \quad (6)$$

$$\sum_{j\in AC}\sum_{k\in K} x_{ijk} = 1, \forall i \in A \quad (7)$$

$$\sum_{i\in AC}\sum_{k\in K} x_{ijk} = 1, \forall j \in A \quad (8)$$

$$\sum_{i\in A}\sum_{j\in AC} D_i x_{ijk} \le CAP_k, \forall k \in K \quad (9)$$

$$\sum_{i\in AC}\sum_{j\in AC} T_{ij} x_{ijk} \le T_{limit}(k), \forall k \in K \quad (10)$$

$$a_{kj} \ge a_{ki} + T_{ij} - M(1-x_{ijk}), \forall j \in A, i \in AC, k \in K \quad (11)$$

$$a_{ki} \ge e_{ki}, i \in AC, k \in K \quad (12)$$

The objective function (1) minimizes the total delay cost. Constraints (2) enforce the number of aircraft on duty. Constraints (3) ensure that each regional airport should be accessed only once and the aircrafts should return back to the starting airport. Constraints (4) - (6) ensures the network flow conservation, constraints (4) impose that each aircraft must start off at the starting airport, i.e. the hub airport. Constraints (7) and (8) indicate that all passengers must be delivered to their destinations. The capacity of aircraft, the limitation of flying time and the arrival time are imposed in constraints (9) – (12).

3 Heuristic Algorithm

Heuristic algorithms, for example GA can be used in solving the mentioned problem. A chromosome of the population can be indicated as ($0, i_{11}, i_{12}, \cdots, i_{1s}, 0, i_{21}, \cdots, i_{2t}, 0, \cdots, 0, i_{m1}, \cdots, i_{mv}, 0$), i_{kj} presents the airport number and the length of chromosome is N+m+1. 0 presents the hub airport and it cuts the chromosome into m segments which indicates m aircraft routes.

The object function can be transformed as $Min\sum_{k\in K}\sum_{i\in AC} C_i^{delay}(a_{ik}-e_{ik}) + M\sum_{k=1}^{N} max\left(D_{ij}x_{ijk} - CAP_k, 0\right)$, $M\sum_{k=1}^{N} max\left(D_{ij}x_{ijk} - CAP_k, 0\right)$ indicates the capacity constraint violation penalty of the solution. In order to strictly meet the capacity constraints, we set the M as a number large enough and $f_i = 1/Z_i$, which f_i is the fitness value of chromosome i and Z_i is the formula (14).

As to the crossover and mutation, this paper applies the method provided in Li Jun et al.[9,10].

We set the crossover rate p_c=0.85, mutation probability p_m=0.3, population size popSize =30 and the maximum number of evolutionary generation maxGen=200.

4 Applications and Analysis

4.1 Disruption Scenarios

In this paper, we apply the instance of a domestic airline in China for testing our method and model. Original ACRs are shown as Table 4.1. We design two disruption scenarios, airport closure and unavailable air route. Due to the weather, the hub airport PVG and HGH is closed until 09:00. We set the delay cost of each passenger as ￥0.87 per minute.

Table 4.1 Original ACR of each Aircraft

AC	DTAT	DTT	ARAT	ART	ARPax	AC	DTAT	DTT	ARAT	ART	ARPax
5341	PVG	815	SZX	1035	45	5330	HGH	825	CAN	1030	90
	SZX	1120	XMN	1220	45		CAN	1105	XIY	1340	55
	XMN	1305	KWL	1430	40		XIY	1410	WUX	1550	10
	KWL	1510	KWE	1555	35		WUX	1610	CSX	1750	65
	KWE	1630	HSN	1720	35		CSX	1820	HGH	2010	59
	HSN	1755	WUH	1915	25	5353	HGH	900	PEK	1115	91
	WUH	1940	PVG	2115	60		PEK	1325	CGQ	1510	60
2632	PVG	800	SJW	1000	96		CGQ	1600	TNA	1800	58
	SJW	1145	NKG	1320	12		TNA	1840	KOW	2040	12
	NKG	1400	WNZ	1500	50		KOW	2120	HGH	2250	30
	WNZ	1630	PVG	1745	30	5393	HGH	800	FOC	920	40
2631	PVG	940	KMG	1300	83		FOC	1040	KHN	1150	75
	KMG	1405	CTU	1520	56		KHN	1220	CKG	1420	56
	CTU	1625	HAK	1835	48		CKG	1500	LHW	1610	53
	HAK	1930	PVG	2200	60		LHW	1700	HGH	1900	78

4.1.1 Airport Temporary Closure

Solving the model by Lingo 9.0, the new ACRs are obtained. As it indicates, ACR of aircraft 5353 and 2631 is unchanged and others are shown as Table 4.2. The delay cost of new ACRs is ￥28757.85. Using GA, we obtain the new ACRs of Hub Airport Temporary Closure shown as Table 4.3. The ACRs of 5353 and 2631 is unchanged and the cost of new ACRs is ￥30711.00.

Table 4.2 New ACRs of Hub Airport Closure in GA

AC	DTAT	DTT	ARAT	ART	
2632	PVG	900	SJW	1100	
	SJW	1245	CKG	1435	
	CKG	1515	LHW	1625	
	LHW	1715	KWE	1855	
	KWE	1930	HSN	2020	
	HSN	2055	PVG	2145	

Table 4.3 New ACRs of Hub Airport Closure

AC	DTAT	DTT	ARAT	ART
2632	PVG	900	SJW	1100
	SJW	1245	CKG	1435
	CKG	1515	NKG	1705
	NKG	1720	KWE	1920
	KWE	1940	HSN	2030
	HSN	2055	PVG	2145

5341	PVG	900	SZX	1120		5341	PVG	900	SZX	1120
	SZX	1205	XMN	1305			SZX	1205	XMN	1305
	XMN	1350	WNZ	1500			XMN	1350	WNZ	1500
	WNZ	1630	WUH	1915			WNZ	1630	WUH	1915
	WUH	1940	PVG	2120			WUH	1940	PVG	2120
5393	HGH	900	FOC	1020		5393	HGH	900	FOC	1020
	FOC	1140	KHN	1250			FOC	1140	KHN	1250
	KHN	1320	KWL	1430			KHN	1320	KWL	1430
	KWL	1510	NKG	1700			KWL	1510	LHW	1700
	NKG	1740	HGH	1835			LHW	1740	HGH	1940
5330	HGH	900	CAN	1105		5330	HGH	900	CAN	1105
	CAN	1140	XIY	1415			CAN	1140	XIY	1415
	XIY	1445	WUX	1625			XIY	1445	WUX	1625
	WUX	1645	CSX	1825			WUX	1645	CSX	1825
	CSX	1855	HGH	2045						

4.1.2 Unavailable Air Route

Assumping the air route between PVG and SZX becomes unavailable. The new ACRs is shown in Table 4.4. The ACR of 5353 and 2631 is unchanged and the cost of new ACRs is ￥14589.90. Using GA as show in Table 4.5. The cost is ￥19218.3 and the ACRs of 5353 and 2631 are unchanged.

Table 4.4 New ACRs for Unavailable_Air route

AC	DATA	DTT	ARAT	ART	
5341	PVG	1030	KHN	1150	
	KHN	1220	XIY	1345	
	XIY	1415	WUX	1555	
	WUX	1615	CSX	1755	
	CSX	1825	PVG	2015	
5393	HGH	800	FOC	920	
	FOC	1040	SZX	1200	
	SZX	1245	WNZ	1500	
	WNZ	1630	WUH	1915	
	WUH	1940	HGH	2115	
5330	HGH	825	CAN	1030	
	CAN	1105	XMN	1220	
	XMN	1305	KWL	1430	
	KWL	1510	KWE	1555	
	KWE	1630	HSN	1720	
	HSN	1755	HGH	1845	
2632	PVG	800	SJW	1000	
	SJW	1145	NKG	1320	
	NKG	1400	CKG	1550	
	CKG	1630	LHW	1740	
	LHW	1830	PVG	2030	

Table 4.5 New ACRs of Unavailable Route in GA

AC	DATA	DTT	ARAT	ART
5341	PVG	1030	KHN	1150
	KHN	1220	XIY	1345
	XIY	1415	WUX	1555
	WUX	1615	CSX	1755
	CSX	1825	PVG	2015
5393	HGH	800	FOC	920
	FOC	1040	NKG	1320
	NKG	1400	WNZ	1500
	WNZ	1630	WUH	1915
	WUH	1940	HGH	2115
5330	HGH	825	CAN	1030
	CAN	1105	XMN	1220
	XMN	1305	KWL	1430
	KWL	1510	KWE	1555
	KWE	1630	HSN	1720
	HSN	1755	HGH	1845
2632	PVG	800	SJW	1000
	SJW	1040	SZX	1310
	SZX	1350	CKG	1550
	CKG	1630	LHW	1740
	LHW	1830	PVG	2030

4.2 Analysis

Through our model, the feasible ACRs are obtained for disruption scenarios. In the above scenario, the model works well. New ACRs of two scenarios make 19.15% reduction. The new ACRs obtained not only satisfy the target of delivering passengers, but also reduce the cost. Obviously, it is better than simply postponing the affected flights in the practice in China. The costs obtained by different methods are compared as Table 4.6.

Table 4.6 Cost Comparison

	Hub Airport Closure	Unavailable Air Route
Trival solution(no ptimization)	35569.95	----
Exact Solution	28757.85	14589.90
Approximate Solution	30711.00	19218.30

5. Conclusion and Future Research

The paper focuses on ACRs of airlines, we enforce that each passenger must be delivered to his or her destination. We present a new modeling method for airline recovery problem, especially the aircraft and passenger recovery problem.

Obviously, it is not enough just focusing on delay. For airlines, they might mobilize more resources to solve delays, such as ferrying in aircrafts and crew, but it may raise the total cost. So, the integrated recovery should be studied in the feature research. Meanwhile, we use Lingo 9.0 on a Pentium-3 PC and it takes 45 minutes averagely for solving the scenarios. The GA takes 52s, 66s and 98s in solving the problem. Therefore, we will focus on heuristic algorithm in the future study.

Acknowledgment

The work was partially supported by research grants from Shanghai Municipal Natural Science Foundation [No. 10190502500]. Shanghai Maritime University Start-up Funds. Shanghai Science & Technology Commission Projects [No. 09DZ2250400] and Shanghai Education Commission Project [No. J50604].

References

[1] Lettovsky L (1997), Airline operations recovery: an optimization approach, PhD thesis, Georgia Institute of Technology, Atlanta, USA.

[2] Stephane Bratu, Cynthia Barnhart (2006), Flight operations recovery: New approaches considering passenger recovery, Journal of Scheduling 9(3): 279-298.

[3] Zhang Y, Hansen M (2008), Real-Time Intermodal Substitution: Strategy for Airline Recovery from Schedule Perturbation and for Mitigation of Airport Congestion, Transportation Research Record: Journal of the Transportation Research Board 2052:90-99.

[4] Serge Bisaillon, Jean-François Cordeau, Gilbert Laporte and Federico Pasin (2009), A large neighbourhood search heuristic for theaircraft and passenger recovery problem, 4OR 9(2):139-157.

[5] Niklaus Eggenberg , Matteo Salani, Michel Bierlaire (2010), Constraint-specific recovery network for solving airline recovery problems, Computers & Operations Research 37:1014–1026.

[6] N. Jafari, S. Hessameddin Zegordi (2011), Simultaneous recovery model for aircraft and passengers, Journal of the Franklin Institute 348(7): 1638-1655.

[7] Jafari, Niloofar and Zegordi, Seyed Hessameddin (2010), The airline perturbation problem: considering disrupted passengers, Transportation Planning and Technology 33(2): 203-220.

[8] Li Jun, Xie Binglei, Yao-Huang Guo (2000), Genetic Algorithm for Vehicle Scheduling Problem with Non-Full Load, Systems Enging-Theory Methodology Application 20(3): 235- 239.

[9] Li Jun, Yao-Huang Guo (2001), Logistics vehicle scheduling theory and method, China Material Press, Beijing.

Performance of Improved Short-Length Raptor Coded Frequency- Hop Communication in Partial-Band Interference

ZENG Xianfeng[1, 2, a], GAO Fei[1, b] and BU Xiangyuan[1, c]
[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2] Dalian Air Force Communications NCO Academy, Dalian, Liaoning 116600, China
[a]20905247@bit.edu.cn, [b]gaofei@bit.edu.cn, [c]bxy@bit.edu.cn

Keywords: Anti-interference; Frequency Hopping; Fountain Code; Raptor code

Abstract. Fountain code and Frequency hopping (*FH*) communication are two independent research fields. To performs well in partial-band interference, aiming at the characteristics of short code length of FH, this work proposed an improved short-length fountain code to apply in *FH* communication system as forward error correction (*FEC*) code. After the Matlab platform modeling and simulating, we analyzed the anti-interference performance of the fountain code in *FH* system. The experiment results show that, the fountain code can improve the anti-interference performance of *FH* communication, which opened up a new application area for fountain codes.

1. Introduction

As its outstanding performance in anti-interference communication，Frequency Hopping (*FH*) communication has become an important communication means in modern warfare. In order to improve the anti-interference performance, many scholars have conducted the research in *FH* pattern and *FH* system, such as basing on chaotic map to construct *FH* sequences[1]，adaptive *FH* system[2], bad channel removal strategy, differential *FH* technology and so on[3]. In channel coding of *FH*, Reed-Solomon (*RS*) code is widely utilized，but RS code generally requires introducing interweaver, which brings high complexity. In recent years the research on Turbo code is very popular, Turbo code performs well if we can reduce the decoding delay and the high complexity. Low density parity check (*LDPC*) code is one of capacity-approaching code[4]，But its coding parameters are based on certain channel hypothesis, if the actual work environment is complex, its performance will be affected.

Fountain code, which was proposed for the binary erasure channel (*BEC*) by Luby and Byers in 1998[5], is a kind of new sparse code with many desirable features. Whereas traditional erasure codes like Reed-Solomon codes have a fixed code rate that must be chosen before the encoding begins, Fountain codes are rateless as the encoder can generate on the fly as many encoded symbols as needed. This is an advantage when the channel conditions are unknown or time-varying because the use of a fixed channel code rate would lead to either band-width waste if the erasure rate is overestimated or to poor performance if it is underestimated. Compared to Reed-Solomon codes, Fountain codes have lower encoding and decoding complexity, but require a few more encoded symbols at the receiver for successful decoding. Luby Transform (*LT*) code is the first kind of practical fountain code [6] and *Raptor code* is improved from *LT* code by Shokrollahi. By utilizing certain block code, such as *LDPC* code or *Hamming* code, cascading *LT* code, Raptor code realized linear time encoding and decoding [7]. Because of the desirable feature and capacity-approaching performance, Raptor code has been adopted as standard coding scheme of the 3GPP-MBMS, DVB-H and so on. Scholars are also exploring the application of fountain code in distributed network storage，cooperative communication, relay communication in wireless fading channel environment[8]. There are also papers studying about the capacity of short length fountain codes [9,

10, 11]. At the present time, the fountain code and *FH* communication are two independent study fields. This paper will base on the characteristics of *FH* communication, and design a short length Raptor code, in order to improve the anti-interference performance of *FH* communication system.

2. System Modeling

Interferences of *FH* communication classify mainly three kinds：tracking interference, broadband interference, and partial-band interference. Tracking interference can be overcome by increasing the jump speed and reasonable constructing network. Partial-band interference, in what the jammer focuses all disturbing energy to launch in partial band of *FH* communication, is a great threat to *FH* communication [12]. If the disturbing energy distributed in all band of the *FH* communication, it is called broadband interference. Because in the broadband interference the energy is distributed to all communication frequency spot, to form effective interference, compared to partial-band interference, the power require much greater. So in practical application, partial band interference is more effect.

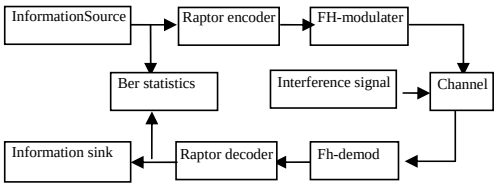


Fig.1 Model of Raptor code for anti-interference in frequency hopping communication system

In order to perform well in anti-interference, short length Raptor code is designed to apply in the FH system as *FEC* code, system model is shown in Fig.1. The information source data are coded in the Raptor encoder firstly, and then modulated by *FH* modulator. When transmitted in Additive White Gaussian Noise (*AWGN*) channel, interference signal is added into this channel. On the receiving terminal, the data are demodulated by the *FH* demodulator at first, and then through Raptor decoding return to information sink. For the purpose of testing the influence of Raptor on the system performance, a bit error rate statistical unit is added in the model.

3. Construction of Short Length Raptor code

Let C be a linear code of block length n and dimension k, and let $\Omega(x)$ be a degree distribution. A *Raptor code* with parameters $(k, C, \Omega(x))$ is an LT-code with distribution $\Omega(x)$ on n symbols, which are the coordinates of code words in C. The code C is called the *pre-code* of the Raptor code. The input symbols of a Raptor code are the symbols used to construct the codeword in consisting of *intermediate symbols*. The output symbols are the symbols generated by the LT-code from the intermediate symbols[7].

As an example, a design for Raptor code is given in [7]. In this paper the number of input symbols k is 65536. In the case, they first encode the k input symbols using a systematic LDPC code of high rate, and the degree distribution of output symbol $\Omega(x)$ is shown in Eq.1.

$$\Omega(x) = 0.007969x + 0.493570x^2 + 0.166220x^3 + 0.072646x^4 + 0.082558x^5 + 0.056058x^8 + 0.037229x^9 + 0.055590x^{19} + 0.025023x^{65} + 0.003135x^{66}. \quad (1)$$

Eq.1 is an optimized degree distribution for k equaling to 65536, in *FH* communication, the number of symbol is usually smaller. For example, a classic military short wave *FH* radio, CHESS system, whose hop speed is high up to 5000Hops/s, and the highest transmission rate is 19.2kbps, so each hop (per codeword) can send 4bit. Assuming 30 hops to compose a frame, then the k is 120bit; SINCGARS, ultra-short wave radio set, whose hop speed is 100Hops/s-300Hops/s, and the

transmission rate is 75bps-16kbps, each hop can send 160bit, Assuming 30 jumps to compose a frame as before, then the maximum k is 4800bit. Apparently the degree distribution shown in Eq.1 isn't the optimal scheme for *FH* communication.

According to the definition of Raptor code, considering the characteristic of *FH* communication, a short length Raptor code is constructed, of which the number of input symbol k is 950. Specific construction methods are as follows:

(1) Construct a systematic LDPC code, whose length is 1000, rate is 0.95(950bits for information words and 50bits for check words); the column weight of check matrix is 4.

(2) Divide the original message into k equal packets (the last packet can fill 0 to be equal), and then encode the data packet with LDPC code;

(3)Encode the output symbol of LDPC code with LT code, the encode algorithm is as follow:

1) Select a random degree d_i (i=1,2,3…) according to the degree distribution $\Omega(d)$ (d=1,2,3……), and choose d_i input symbol uniformly as the neighbor node of node i from the 1000 outputs of LDPC encoding

2) XOR the information bits in positions chose by 1)

3) Repeat step1) and step2) until receiving an ACK feedback signal from the receiver or sent default quantity of encoded symbols

The most decoding algorithm of Raptor code is based on belief propagation（*BP*）algorithm. There are many mature algorithms and the improved schemes of it[13,14,15], such as global iterative *BP* decoding algorithm, local iterative *BP* decoding algorithm, soft decision decoding algorithm and so on. In this paper, the decoding algorithm is global iterative soft decision *BP* decoding algorithm.

4. Improved degree distributing scheme

The purpose of the degree distribution is to generate just enough different degree bits to XOR so that a decoder can decode all information. Lower degree encoded bits are required for initiating the decoding process. The degree distribution is designed in such a way that，probabilistically, most of the information bits that are in higher degrees can be decoded from the lower degree bits.

The decoding process of Raptor code is generally based on *BP* algorithm where degree 1 encoded bits are immediately recovered in Tanner graph. The decoding of *BP* algorithm has low complexity and excellent performance when the number of input symbol reaches 10^{-4} [16]. To ensure the high success rate in decoding, when the length code is very short, the overhead of *BP* algorithm will increase，which will lead to the increasing of decoding delay，and reducing the instantaneity of *FH* system. Analyzing the influence of degree distribution on *BP* algorithm，the increasing of the selection probability of small degree, especially the probability of degree 1, would make the decoding more quickly and smoothly, and at the same time reduce the decoding delay, decoding cost and average degree weight. But it also reduces the rate of larger degree, which will cause a few data not to be recovered during the decoding process and increase the failure probability of decoding.

The key to solve this contradiction is choosing a suitable *pre-code* to enhance the ability of error correction. Basing on the analyzing mentioned before, we design a systematic LDPC code, whose code rate is 0.95, that is to say, it can still recover the original data as long as the unsuccessful rate of decoding less than 5% in the stage of LT decoding. And a modified degree distribution function for the Raptor code is shown in Eq.2. The average weight of the degree in this scheme, compared to the degree distribution shown in eq.1 is decreased by 29%, which will reduce the complexity and cost of system.

$$\Omega(x)=0.03299x+0.49357x^2+0.16722x^3+0.07265x^4+0.08256x^5 +0.05606x^8+0.03723x^9+0.05559x^{19}+0.003135x^{33}. \quad (2)$$

5. Simulation and analysis

In order to verifying the influence of improved scheme on system performance, according to the degree distribution scheme shown in Eq.1 and Eq.2, two Raptor codes are structured, scheme 1 and scheme 2, basing on the part 3, whose length is 950. The Raptor codes are applied as FEC code before FH modulation, and simulated in the process of FH communication through MATLAB platform. In the simulation, the decoding algorithm is global iterative soft decision *BP* decoding algorithm [13], and the number of iteration times is 40.

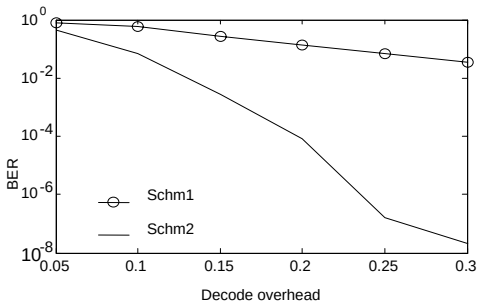


Fig.2 BER Performance comparing in No-interference environment

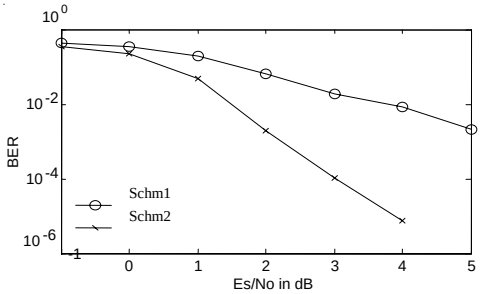


Fig.3 BER Performance comparing in partial band interference environment

Decoding performance in the absence of interference environment is shown in Fig.1. Experiment times are 100000, the horizontal axis is the decoding overhead, and the vertical axis is the bit error rate (*BER*). As is shown in Fig.2, scheme 2 compared to scheme 1, as the decoding overhead increasing, the bit error rate curve decreases more rapidly. When the decoding overhead is 0.3, the bit error rate of Scheme 2 is only 2.1 x 10^{-8}, and bit error rate of scheme 1 is $3.4x10^{-2}$, the performance improvement of scheme 2 achieves 62.01dB. Therefore, the scheme 2, a short length Raptor code based on the improved degree distribution shown in Eq.2, greatly improves the performance of *FH* communication system.

In the partial-band interference environment, the Raptor code scheme 1 and scheme 2 are applied in *FH* communication system for anti-interference, which is shown in Fig.3. The horizontal axis is single symbol *SNR*, and the vertical axis is *BER*. The interference factor ρ is 0.2, and the experiment times are 10000. The overhead of decoding is 0.4. It can be seen from the figure, as the increase of single symbol *SNR*, the *BER* curves both show a downward trend, but the curve of scheme 2 is more quickly. When the single symbol *SNR* is 5dB, the *BER* of scheme 2 is down to zero, and when the single symbol *SNR* is 4dB, the *BER* of scheme 2 is 7.37×10^{-06}, compared to the *BER* of scheme 1, which is 8.55×10^{-03}, the improvements of system performance reach to 30.65dB.

6. Conclusion

In this paper, we've constructed a short length Raptor code applying in *FH* communication system, which combined two independent study fields together. An improved degree distribution scheme is proposed for short length Raptor codes, whose average degree weight is decreased by 29%，and through modeling on Matlab platform, lots of experiments has done to analyze the performance of the improved Raptor code. The result shows that, without interference, the *BER* of improved Raptor code is lower by 62.01dB than the original scheme, and in the partial-band interference environment, where the *SNR* is 4 dB, the interference factor ρ is 0.2, the *BER* of improved Raptor code is lower by 30.65dB. The improved schemes provide a new *FEC* encoding measure for *FH* anti-interference communication and also open up a new area for the application of fountain codes.

References

[1] Y.W.Wang, L.P.Wang, J.H.W: Technique of Signal Spectrum Analysis and Frequency-Hopping Masking Encrypt Based on Chaotic Sequence. ICEEE. 7-9 Nov. 2010. Henan, China.

[2] M.Roy, H.S Jamadagni: Comparative Study of Adaptive frequency hopping with Power Control to Avoid WLAN Interference in WPAN Systems like Bluetooth. Consumer Communications and Networking Conference (CCNC), 7th IEEE. 2010.

[3] Z.Chen, S.Q.Li, B.H.Dong: Multi-user performance analysis of differential frequency hopping system over Rayleigh-fading channel. High Technology Letters 2008.02: 147-153.

[4] J.T.Zhao, M.Zhao, H.B.Yang: High Performance LDPC Decoder on cell before WiMAX System. Proceedings of the 2011 Third International Conference on Communications and Mobile Computing (CMC 2011), p 278-81, 2011.

[5] L.Yuan, J.P.An: Design of UEP-Raptor codes over BEC. European Transaction on Telecommunications, 2010.

[6] M. Luby: LT codes. Proceedings of the 43rd Annual IEEE Symposium on the Foundations of Computer Science (STOC), Vancouver, Canada, November, 2002.

[7] A.Shokrollahi: Raptor code. IEEE Transaction on Information Theory, 2006.

[8] X.M.Li. Research on Theory and key Technology of Novel Network Fountain Coding on http://isis.nsfc.gov.cn/portal/Proj_List.asp. 2009.

[9] W.Z.Huang, H.L.Li, J.Dill: Fountain codes with message passing and maximum likelihood decoding over erasure channels. (2011 Wireless Telecommunications Symposium (WTS 2011), New York City, USA.2011).

[10] E. A. Bodine, M. K. Cheng: Characterization of Luby Transform Codes with Small Message Size for Low-Latency Decoding. ICC '08. IEEE International Conference on Communications, 2008.

[11] S.Shamai, I.E.Telatar, S. Verdu: Fountain Capacity. IEEE Transactions on Information Theory, vol. 53, pp. 4372-4376, 2007.

[12] C.L.Liu, X.Y.Bu, H.J.Wang: Nonbinary LDPC coded frequency hopping systems over partial-band jamming channels. WiCOM 09.5th, International Conference on 24-26 Sept. 2009, Piscataway, NJ, USA, 2009.

[13] H.Jenkac, T.Mayer, T.Stockhammer: Soft Decoding of LT-Codes for Wireless Broadcast. Proc. IST Mobile Summit 2005.

[14] G.H.Yu，Y.H.Yang，Y.J.Wei: An Improved Algorithm for Decoding of Raptor Codes. Communications Technology. Vol.43, No.08, 2010.

[15] H.J.Zhu,Y.K.Pei,J.H.Lu: Algorithm improving the decoding performance of fountain codes. Journal of Tsinghua University (Science and Technology),Vol.50,No.4.pp.609-612 ,2010.

[16] L.Yuan: Research on encoding and decoding technologies for fountain codes in wireless multimedia transmission[D]. Beijing: Beijing Institute of Technology, 2011.

Analysis of Highway Rear-end Accidents based on FTA Method

Yun Jiang[1, a]

[1]School of Management, Wuhan University of Technology, China

[a]ppjy1990@163.com

Key words: Highway; Rear-end accident; FTA method

Abstract. In order to reduce highway rear-end accidents, the article used the fault tree analysis (FTA) method to analyze Highway Rear-end accident, and built up general universal model of FTA of rear-end according to the influence of various factors and the logical relationship. the January 8, 2011 Great Highway rear-end accident of five cars made use of FTA method, and analyzed qualitatively to find out minimal cut sets and structure importance. Then the main factors leading to accidents were identified. This method can reflect the fact systemically and objectively, find out the direct reasons for the accidents, has a good effect on forecasting and preventing Rear-end accidents.

Introduction

With the accelerated pace of life, the development of auto industry and the construction of the highway, rear-end accidents are also increasing. According to statistics, more than 80% of the accident was caused due to the driver reaction not timely, more than 65% of vehicle crashes are rear-end collisions [1]. In the highway which are fully closed, all entrances and exits controlled and centralized management , ruled out the lateral interference of pedestrians and vehicles from the road, accident rate only comprises 1/3~1/4 of general traffic accident, however, due to highway speed quickly, once there is a traffic accident, the severity degree will corresponding increase. According to statistics, in highway accident, each vehicle collision accidents account for about 33.4%, and mostly rear-end collision accident [2]. Therefore, the study of the main factors of freeway rear-end collision accident is of great practical significance for the control and prevention of highway rear-end accidents.

At present, the study of highway traffic accidents remains in the qualitative analysis and the related research review less, mostly qualitative analysis. Such as Li (2000) the highway rear-end accident cause analysis, the article analyzed highway rear-end accidents cause from five aspects which are driver's visual characteristics while high-speed driving, characteristics of driving field in different models, characteristics of the natural vehicle acceleration and deceleration, the handling characteristics of the driver and compliance of the vehicle; Zhang (2010) the study of objective factors and countermeasures of Xi'an highway accidents, the article analyzed the objective factors of highway accidents based on the traffic characteristics in time and spatial distribution; Dai (2010) causes analysis management measures of highway traffic in the fog, the article analyzed the causes and characteristics of traffic accidents in foggy weather, and provide ideas for highway traffic management in foggy weather.

To try to eliminate hidden dangers of expressway traffic safety, firstly, to use scientific and reasonable methods to find out the multiple source of highway rear-end collision accidents. Through review of relevant literature, we find the application of FTA (Fault Tree Analysis) is relatively mature, Such as Zhang (1996) cause of the fire investigation; Dai (2005)determining the controlling factors of coal mine gas outburst through Fault Tree Analysis; Zhao (2010) the reason discussion of coal mine fire accident based on Fault Tree Theory; Gu (2010) the factors analysis of

pipeline leakage based on Fault Tree Analysis; You (2010) the application of fault tree analysis in the pulp transportation pipes burst accidents; Zhou (2010) the application of fault tree analysis in Quantitative Management of Coal Mine Safety; Yang (2010) the application of fault tree analysis in mine fire; Liu(2011)the cause analysis of electric shock accidents of operators in construction site. To this end, Fault Tree theory is chosen to analyze the reasons of highway rear-end collision accidents, to identify the main cause of the accident and to propose preventive measures, and can also point out the focus and direction for highway traffic safety management.

The basic concepts and principles of Fault Tree Analysis

Fault tree analysis is widely used in security systems engineering, which is a strong logical analysis, a graphical method of deductive reasoning, that is the method, from the results to the reasons analysis [3]. The method is from a possible accident (top event), top-down, layer by layer to find the direct and indirect cause events of the top event until the underlying cause events (basic events), and use logic diagram to express logical relationship between these events [4], and an inverted tree is formed. It is called the fault tree.

Fault tree is consisted of all kinds of symbols and logic gates, such as the event symbols, the logic gate symbols and transfer symbols, the specific form see literature [4,5,6].

If the cut set is no longer a cut set after a basic event out of it, we called the cut set as minimal cut set. There are two methods which are Boolean algebra simplification and ranks to resolve minimal cut set. This article will make use of Boolean algebra to simplify the fault tree and get a union set of a number of intersections, each intersection is a minimal cut set in fact.

Analysis of the structure importance is to analyze the impact on top events from the basic events, and without taking into account the probability of occurrence of each basic event, or assume that the probability of occurrence of basic events are the same, general use IΦ(i) said. A fault tree is consisted of N basic events, assuming there is m_i kinds when xi is from 0 to 1, the state of top event is from 0 to 1, and the coefficient of the structure importance of the basic event X_j is defined:

$$I\varphi(i) = \frac{m_i}{2^{n-1}} = \frac{1}{2^{n-1}} \sum_{j=1}^{2^{n-1}} [\varphi(1_i, X_j) - \varphi(0_i, X_j)] \quad (1)$$

Therefore, we can solve minimal cut set and determine the structure importance of the basic events through fault tree analysis, to find out the main factors and establish security measures to prevent accidents.

The construction of highway rear-end accidents tree model

First of all, to make a investigation about all causes of accidents related top events from man, machine and environment, highway rear-end collision is the top event of the tree in this article. Many reasons are involved in highway rear-end accidents, through the wide access to information and consulting the related experts' advice from transportation management department, factors are summarized from three areas of people, vehicles, and environmental, see figure 1.

Then, using deductive reasoning analysis method, constructing the logical relationship between those factors influenced a highway rear-end accident step by step, and setting up a general model for rear-end fault tree analysis, as shown in figure 1. In the figure, T stands for the top event of rear-end fault tree; T_i (i = 1，2，……，n) stands for intermediate event; X_j (j = 1，2，……，n) stands for different intermediate events and basic events that may lead to rear-end accidents.

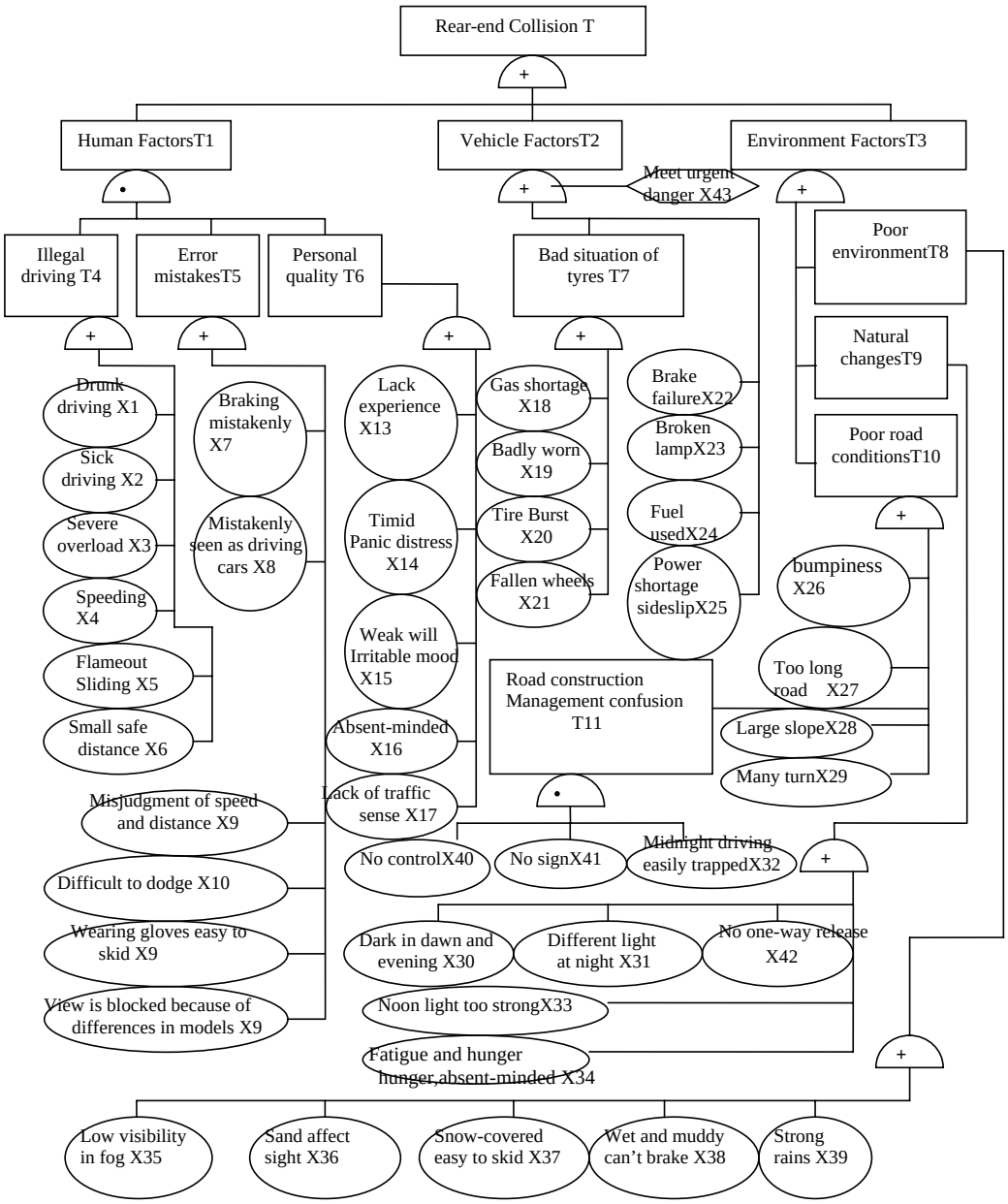


Fig. 1: The general model of rear-end fault tree analysis

The main factors analysis of 5 cars rear-end accident on Chuda highway in Yunnan

Figure 1 shows that many basic events are involved in a generic model of rear-end accidents tree,relationships between them are more complex, the number of minimal cut sets are more diverse and complex, the workload of solving completely is very large, however, there are the following two methods to meet the actual needs. First, as the a specific rear-end incident is concerned, basic events are not the same as those in Figure 1, that is, not all the basic events listed play a more prominent role, so we can remove those do not exist or weak impact on rear-end events, to simplify

the actual rear-end accidents tree correspondingly, this not only meets the practical needs, and to improve the processing speed, finally the number of the minimal cut set (path set) are reduced and the problem is solved more easily. Second, we can get all the minimum cut sets by means of the advantages in data management and processing speed of computer [7]. The qualitative analysis of fault tree which is combined with chuda highway rear-end accident is shown as following.

The solution of fault tree minimal cut sets. At 0:30 on January 8, 2011, 5 cars rear-end collision occurred at 3 kilometers of xiazhuang toll station, in the direction from Chuxiong to Dali on chuda highway, and resulting in 38 people trapped. The traffic police department initially determined the cause of the accident is a bus carrying 38 passengers bus rapid speeded in motion, and the driver fatigued and nodded off, which caused the body unresponsive, did not have time to stop, and hit the two big truck in front. By this time, a bus and a truck followed by the bus had no time to avoid, and 5 cars rear-end collision occurred. X4、X6、X10、X13、X16、X22、X25、X27、X31、X32、X43 are the basic events which influence and control the rear-end accident. See figure 2.

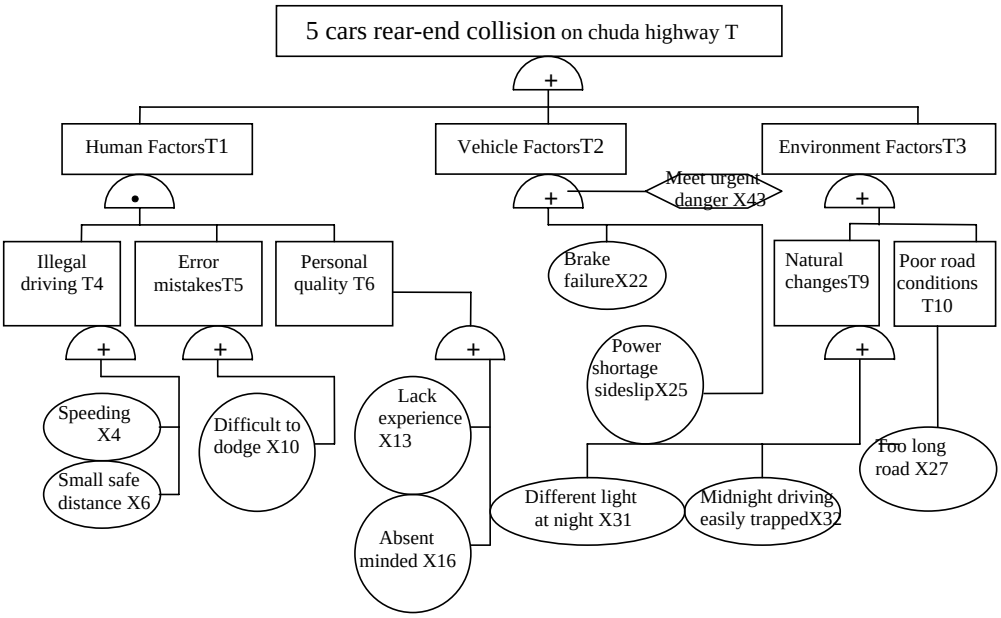


Fig.2: The accident tree of 5 cars high-speed rear-end on chuda highway in Yunnan

According to the logical relationship between events in figure 2, fault tree can be expressed in the Boolean algebra:

$$T1=T1+T2+T3=T4*T5*T6+X43(X22+X25)+T9+T10$$

$$=(X4+X6)X10(X13+X16)+X22*X43+X25*X43+X31+X32+X27$$

$$=X4*X10*X13+X4*X10*X16+X6*X10*X13+X6*X10*X16+X22*X43+X25*X43+X31+X32+X27$$

Finally, 9 groups of minimum cut sets of the accident tree are obtained as the following:

G1={X27};G2={X31};G3={X32};G4={ X4，X10，X13};G5={X4，X10，X16};

G6={X6，X10，X13};G7={X6，X10，X16};G8={X22，X43};G9={X25，X43}.

Obviously, when any of the minimal cut sets occurs, the top event (rear-end accidents) is inevitable.

The structure importance analysis of basic events. The frequency of the basic event in the nine groups of the minimum cut set, reflects the impact degree of the basic events on the rear-end accidents, and this is also known as the structure importance. The greater the structure importance of the basic events is, the greater influence on the top event is. According to equation (1) and the degree of influence on the accident from the basic events, through the analysis of minimal cut set analysis:

(1) X27, X31, X32 is minimum cut set of the single event, thus

$I\Phi$（27）=$I\Phi$（31）=$I\Phi$（32）> $I\Phi$(others)

(2)X43 appears in the minimum cut set of two events twice, and X22, X25 appears only once, thus

$I\Phi$（43）> $I\Phi$（22）=$I\Phi$（25）

(3) X10 appears in the minimum cut set of three events for four times, and X4, X6, X13, X16 appears in the minimum cut set of three events twice, thus

$I\Phi$（10）> $I\Phi$（4）=$I\Phi$（6）=$I\Phi$（13）=$I\Phi$（16）

(4) Use the formula $I\phi(j)=\sum_{Xj\in Gr}\frac{1}{2^{n_j-1}}$ to calculate.

$$I\Phi（43）=\sum_{Xj\in Gr}\frac{1}{2^{n_j-1}}=\frac{1}{2^{2-1}}+\frac{1}{2^{2-1}}=1$$

$$I\Phi（10）=\sum_{Xj\in Gr}\frac{1}{2^{n_j-1}}=\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}=1$$

$$I\Phi（22）=\sum_{Xj\in Gr}\frac{1}{2^{n_j-1}}=\frac{1}{2}$$

$$I\Phi（4）=\sum_{Xj\in Gr}\frac{1}{2^{n_j-1}}=\frac{1}{2}$$

To sum up, the order of structure importance of basic events: IΦ（27）=IΦ（31）=IΦ（32）> IΦ（43）= IΦ（10）>IΦ（4）=IΦ（6）=IΦ（13）=IΦ（16）= IΦ（22）=IΦ（25）

The main factors analysis of the accident. The results can be obtained from the above, the most important reasons of this accident are long road leading to fatigue (X27), different light at night(X31), midnight driving easily trapped(X32); meeting urgent danger(X43), difficult to dodge(X10), speeding(X4), small safe distance(X6), lack experience(X13), absent minded(X16), brake failure(X22), power shortage and sideslip(X25)are followed by. These are the main reasons of 5 cars rear-end collision on chuda highway in Yunnan.

Precautions

From the above analysis of fault tree models and case study, we can get the highway rear-end collision accident causes and laws. On the basis of these, it is more important to find out measures to actively prevent highway vehicle rear-end accidents.

(1) Carrying out highway vehicle rear-end accident investigation and analysis, fully and accurately grasp the freeway rear-end collision accident characteristics and laws.

(2)Strengthen propaganda, to create a law-abiding atmosphere. Ensure highway traffic safety, it needs to start from the highway publicity, so that the community people consciously abide by laws and regulations.

(3)It needs to cooperate perfectly with travel, weather, and other closely related departments, to further improve the work of fog, rain, snow and other inclement weather accident prevention, control, processing.

(4)Regulating the construction and eliminating the risk caused by construction operations.

(5)Checking and eliminating hidden dangers and perfect the road traffic safety facilities.

Conclusions

In view of the diverse and complex factors of highway rear-end accident, FTA is an effective method, which could analyze systematically and quantitatively. Based on a comprehensive analysis of various factors and the principles and rules of fault tree, the papers built up a general universal model of Fault Tree Analysis of rear-end. The model was applied to the occurred incident. On the one hand, it helps to ensure the completeness and reliability analysis; on the other hand, it can increase the efficiency of the accident investigation and is of great significance for the prevention and control of accidents.

References

[1] X. Wu, X. Ge, H. Huang: Research of Avoiding Automobile Rear-end Collision and Safety Traffic on Highway. Journal of Xiamen University, Vol. 3 (2009), p.119

[2] Y.Li, B.Liu: Highway Rear-end Accident Cause Analysis. Journal of Tianjin Automobile, Vol. 2 (2000)

[3] J.Yao: Research of Marine Accidents based on FTA Method. Journal of Dalian Maritime University, Vol. 25 (2010): 348-352

[4] Y.Zhuang, P.Lei: Safety Accident Emergency Management. Beijing: the press of Chinese economy (2009).

[5] J.Jiang: The Accident Investigation Forecast and Analysis Technology. Beijing: Chemical Industry Press (2004).

[6] Y.Niu: Safety Production Techniques. Beijing: Chemical Industry Press (2005).

[7] X.Dai: Finding out the Major Factors of Controlling Gas Outburst by the FTA Method, Vol. 24 (2005): 333-340.

Reputation Management of Art Communication in Internet

Li Cui [1] and Juan Han [2 a]

[1] Jiujiang University, Jiangxi, 332005, China

[2] University of Shanghai for Science and Technology, Shanghai, 200093, China

[a]juanhan81@gmail.com

Keywords: the pornographic contents, reputation, Internet, art communication

Abstract. In order to avoid the transmission of the pornographic contents which are harmful to most of people, a novel method based on the reputation management is studies in the paper. Based on a set of the characteristic parameters, the reputation of a node is calculated for evaluating whether a node is the illegal node. The method is easy to be realized, and supervises the art websites efficiently for the art communication in Internet.

1. Introduction

With the development of modern Internet technology, many art websites are built in recent years. These art websites are gradually charged with the main function of the art communication instead of the traditional communication media. Of all art websites in China, a small proportion of them are built by the national government or the local governments, while most of them are built by the investors, the entrepreneurs, and the fanciers. Because of no official website grading system in China, all people can read the complete contents in the websites, who include children and teenager. Generally speaking, the art websites that are built by the governments have the impeccable governance mechanism that can assure that the young people can not read adult contents. However, in another group of the art websites, which are built by the investors, the entrepreneurs, and the fanciers, often show some adult contents that are read by the young people easily. For example, the body art is the important part of modern photographic art and painting art, but it is not fit for children. For the amateurs, the pornographic contents are also difficult to distinguish from the body arts. Some managers of the art websites want to develop their popularity so that they add some pink contents and links in their pages. These contents and links are harmful to most of people, especially children and teenager. Therefore all art websites have to be supervised according to the law. For this purpose, a method based on the reputation management is studies in the paper. The method is easy to be realized, and supervises the art websites efficiently for .the art communication in Internet.

2. Definition of Reputation

From the point of view of the network, the network for the art communication includes the source websites, followers and the reader. The source websites are the content sources. The followers are defined that who paste the contents to other websites, bbs, blog and twitter. The readers are defined as the final receivers of the transmission contents. We can see the art communication network as a peer-to-peer network [1]. Moreover, we can use the reputation [2] [3] [4] to define every node in the art communication network.

The reputation is the opinion of a group of entities toward a person, a group of people, or an organization on a certain criterion. It is an important factor in many fields, such as education, business, online communities, or social networks [3] [5]. In the art communication network, we

define a series of characteristic parameters for every node. All characteristic parameters are linearly independent with each others. So we treat the set of characteristic parameters as a Euclidean space. An N dimensions Euclidean space $\boldsymbol{R}^{N}$ that includes the characteristic parameters is defined as Table.1.Obviously, the maximum number of parameter combination is as follows:

$$K_{\max} = C_{N-1}^{0} + C_{N-1}^{1} + C_{N-1}^{2} + \cdots + C_{N-1}^{N-1} = 2^{N-1}, \quad (1)$$

where K_{max} is the maximum number of parameter combination.

Table.1. Definition of character parameters in nodes

Description of characteristic parameter	Parameter	Constraint condition
Number of pink image in latest examination	$\boldsymbol{\varepsilon_1}$	$\varepsilon_1 \geq 0$
Number of pink text latest examination	$\boldsymbol{\varepsilon_2}$	$\varepsilon_2 \geq 0$
⋮	⋮	⋮
Days how long no pink content is in the website	$\boldsymbol{\varepsilon_N}$	$\varepsilon_N \geq 0$

Based on Tab.1, we can define the reputation of every node as follows:

$$\boldsymbol{Rep}=\alpha_1\boldsymbol{\varepsilon_1}+\alpha_2\boldsymbol{\varepsilon_2}+\ldots+\alpha_N\boldsymbol{\varepsilon_N}, \quad (2)$$

where $\alpha_1, \alpha_2, \ldots, \alpha_N$ are impact factors that are defined as the influence of every characteristic parameter.

3. Reputation Management

After defining the reputation, we study on the management method of the reputation. At first we need to initialize $\boldsymbol{Rep_i}$ and $V_i(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \ldots, \boldsymbol{\varepsilon_N})$ when the node i is added into the art communication network. In the supervision node, a node list has to be built and real-time refreshed. An algorithm is proposed to initialize the parameters and reputation when a node is added into the art communication network, which is shown as Algorithm 1.

Algorithm. 1. Initializing a node when node is added into art communication network

When Node i is added into art communication network

Initializing $V_i(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \ldots, \boldsymbol{\varepsilon_N})$

$\boldsymbol{Rep}_i \leftarrow$ GetReputation(V_i)

Inserting $V_i(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \ldots, \boldsymbol{\varepsilon_N})$ and $\boldsymbol{Rep}_i$ into Nodelist

After initializing $\boldsymbol{Rep_i}$ and $V_i(\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \ldots, \boldsymbol{\varepsilon_N})$, the supervision node may use some methods to detect the nodes in the art communication network, such as the data mining, the string matching, the image recognition, and the manual review. If some characters that are predefined in V_i are found out, the supervision node edits the characteristic parameters in V_i and calculates $\boldsymbol{Rep_i}$. If $\boldsymbol{Rep_i}$ is larger than the presetting threshold value, the supervision node will mark the node i as the illegal node. The process of judgment based on the reputation is shown as Algorithm 2.

Algorithm. 2. Detection of illegal node

Calculating $\boldsymbol{Rep_i}$ of node i

If $\boldsymbol{Rep_i} > T_{Rep}$ then

Marking node *i* as illegal node
End If

For the different node, the methods of calculation on the reputation are also different from each other. The source websites are key objects in the network management, which are main content sources in the art communication network. Therefore, the occurrence frequency and number of the pink texts and images in the source websites are the important influence of the reputation. The followers are another group of objects which need to be supervised because the followers are secondary content sources in the art communication network. They retransmit not only the contents from the source websites in the art communication network but also the contents from other information provider. Therefore the retransmission frequency of the pink texts and images are the important influence of the reputation for the followers. In the calculation of reputation, we have to define the different values of impact factors to assess different nodes.

4. Conclusions

In order to manage the nodes in the art communication network for avoidance of transmission of the pornographic contents, a novel method based on the reputation management is studies in the paper. An N dimensions Euclidean space $\boldsymbol{R}^{N}$ that includes the characteristic parameters is defined firstly. Based on the characteristic parameters, the reputation of a node is calculated for evaluating whether a node is the illegal node. The method is easy to be realized, and supervises the art websites efficiently for .the art communication in Internet.

Reference

[1] Z. Despotov, and K. Aberer: Computer Networks Vol. 50 (2006), p.485--500

[2] K.Aberer and Z. Despotovic: In proceedings of the 9th International Conference on Information and Knowledge Management, McLean, USA (2000)

[3] S.Buchegger and J. L. Boudec: In proceedings of P2PEcon 2004, Berkeley, USA (2004)

[4] N. Stakhanova, S. Ferrero, J. Wong, and Y. Cai: In proceedings of 17th International Conference on Parallel and Distributed Computing Systems, San Francisco, USA (2004)

[5] R. Jurca and B. Faltings: In SIGCOMM2005, Philadelphia, USA (2005)

Study on Content Clustering in E-Journal Operation

Juan Han[1a] and Li Cui[2]

[1] University of Shanghai for Science and Technology, Shanghai, 200093, China

[2] Jiujiang University, Jiangxi, 332005, China

[a]juanhan81@gmail.com

Keywords: e-journal, content clustering, profit model, personalized requirement.

Abstract. The e-journal has become an important communicative carrier in the information age. The income from the readers who require the personalization service according to their hobbies and professions is studied. On that basis, we study on the content clustering for the e-journal and its profit model. From our analysis, the clustering of content can not only enrich the operator profit model of the e-journal, but also have the beneficial effect on the development of the e-journal.

1. Introduction

With the development of the digital technology and Internet, the e-journal has become an important communicative carrier [1]. The modern e-journal not only includes the texts and the images, but also uses the audios, the videos, and the flash movies as the medium of the messages. Both the novel peer-to-peer networks and traditional client-server networks are applied to publish the e-journal. A user can subscribe the different journals from the e-journal database according to his hobby and requirements. In the profit model of the e-journal, the advertisement is the main income, while the basis of the profit from the advertisement is the number of users. Therefore the first aim of the e-journal is improve the amount of distribution. Moreover, the e-journal has both the traditional profit model and the novel profit model based the networked publication system, such as the download income, the income from the network flow, and so on. However, because of the operation principle, the limitation of technology, the profit model, and the influence of the government policy, many companies can not realize the earning after taxation. How to enhance the competition of the company of the e-journal becomes a hot topic of the digital communications and the e-commence [2] [3].

2. Analysis of Income Source

From the analysis of income sources, operating earnings of e-journal are four parts as follows. The first is the advertising revenue; the second is paid from readers; the third is the copyright transfer fee from the large digital libraries; the fourth is paid from the authors (page charges). The advertising revenue increases rapidly due to the rapid economic development. The recent statistics [4] shows that the Chinese populations who are above the age of 15 have an average of 8.3 years of education, and more than 70 million people studied in universities or colleges. The target consumer markets for them are also expanding, especially in the luxury market, high-tech product market, tourism markets, and so on. These increasing markets provide more opportunities for the development of e-journal. In addition, highly educated people are easier to fit for the emerging networked society. So the income for reading the digital publication has become another major source. With the realization of the national digital library projects and the commercialization of digital publishing platforms, copyright transfer fee has gradually become an important income

source. The page charges are mainly from the academic e-journal, which only are a small part of the operating profit.

We analyze for the income from the readers who require the personalization service according to their hobbies and professions. On that basis, we study on the content clustering for the e-journal and its profit model. The contents of the e-journal are similar to the contents of the printed journal. In order to fit for the requirements of the different readers, the content of every issue has a certain degree of depth and scope. However, because the different readers have the different requirement of interesting and profession, seldom readers are interested in all articles in a journal. In the sales model of the e-journal, the vouchers and e-banks are often used. The sale of e-journal can be also based on an article or an issue while the printed journal is often sold based on an issue. Therefore, the profit model based on the content clustering becomes a new model for the earning of the e-journal.

3. Content Clustering for E-Journal

The subdivision is the basis of the content classification. The subdivision is often defined as a process that a company divides the consumers into some groups of buyers according to their different requirements. In our opinion, the purpose that a company needs the subdivision is to reunite the subdivided consumers. The subdivision variables include the demographic factors, the consumer psychology, and so on. The contents of the computer e-journal are divided into several classes that are shown as Fig.1. In the tree structure in Fig.1, we can see that the journals are divided into several categories in accordance with the requirements of readers, while every category is also subdivided into some small sub-classes.

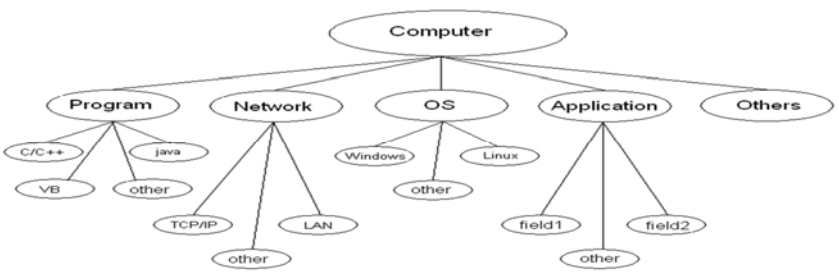


Fig.1. Classification of computer journals

The content subdivision of the printed journals is to facilitate the readers to easily find what they need. Regardless of an article or several articles in a journal in which the reader is interested, this reader has to buy a book of journal. So some readers will abort to buy the journal because of the individual pecuniary condition of the readers or much useless contents in this journal. The e-journal can make up for the shortcomings of the printed journal. The readers can purchase the selected contents in an e-journal according to their requirements or buy a book of e-journal. The method not only enriches the reader's consumption patterns, but also provides a novel profit model for the companies that operate the e-journal.

In contrast to the printed journal, the e-journal has the higher operating cost and the more complex platforms. However, the e-journal has some advantages that the printed journal can not reach. These advantages support the personalized service for the readers according to their requirements and hobbies. The database of the e-journal can cluster the personalized requirement for each reader, so as to optimize the choosing contents for the reader's interest. The income is also increased.

After realizing the personalized requirement of the reader, the company also can supply the related value-added services for the different requirement. For example, the operators can carry out

the content integration of journals for more high-end needs of the reader, and support customized service packages for the readers, such as the multimedia information support, the hyperlink across several related but different journals, and so on.

4. Conclusions

In summary, the method that uses content clustering and the charging for a text in a e-journal based on the content subdivision can increase investment of reading from the reader effectively. The method not only increases the income, but also expands the number of readers. Other incomes are also promoted because of the increasing readers. Therefore, the clustering of content can not only enrich the operator profit model of the e-journal, but also have the beneficial effect on the development of the e-journal.

References

[1] M. Sitko, N. Tafuri, G. Szczyrbak and T. Park: Serials Review Vol. 28 (2002), p.176--194

[2] E. N. Thomas: Electronic Journal Collection Management Issues Vol.16 (1997), p.58--65

[3] T. F. Eamon: On the Street: Electronic Journals and Electronic Content Management Systems Vol. 86 (1998), p.82--84

[4] China education development report, http://www.moe.edu.cn/

A Flexible Workflow Management System Architecture Based on SOA

Huifang Li [1,a], Cong Chen[2,b]
School of Automation, Beijing Institute of Technology (BIT),
Haidian District, Beijing, 100081, P. R. China
[a]huifang@bit.edu.cn, [b]chencongbit@163.com

Keywords: SOA; Workflow management system; ESB; Flexibility;

Abstract. Today's fast-paced and competitive business environment requires enterprise's Workflow Management System (WfMS) is capable of adapting to changing business requirements more efficiently, easily and quickly. However, the traditional WfMS can't meet these requirements. This paper proposes an implementing architecture for WfMS to improve its flexibility using Service Oriented Architecture (SOA). This architecture divides WfMS into several service modules which communicate with each other through ESB (Enterprise Service Bus). Services can be integrated with any enterprise legacy systems and/or new applications easily, and the integration will not depend on the language, operation system and database which the applications use. The characteristics of loosely coupled services make WfMS more flexible, and have capabilities for faster and efficient response to new business changes.

1. Introduction

WfMC (Workflow Management Coalition) describes that workflow is concerned with the automation of procedures where documents, information or tasks are passed among participants according to a defined set of rules to achieve, or contribute to an overall business goal [1]. The goal of workflow management is to make sure that the proper activities are executed by the right person at the right time. Workflow is the computerized facilitation or automation of a business process in whole or part.

The workflow management system has been widely used, quickly developed and brings huge benefits to the enterprises. With the economic globalization, more and more customers' requirements and ever-changing marketplace require enterprises to reduce the product cost and development cycle. While the traditional workflow management systems are monolithic [2]. It is difficult to extend, and strongly depends on the programming languages, operating systems, network protocols and database used. These defects drive people research and design a new architecture for WfMS.

The emergence of SOA and Web Service technology brings a good chance to solve the above problems. The core idea of SOA is to separate the application system's function from its concrete realization [3]. SOA is essentially based on a set of services, which have many characteristics such as loosely coupling, reusability, good contract interfaces. This makes any SOA-based systems easily upgraded and extended. It will be a good idea to combine SOA approach and workflow technology [4-7]. In this paper we focus on how to construct a SOA based workflow management system according to workflow reference model from WfMC.

This paper is organized as follows: firstly we introduce the concepts, general development framework of SOA, ESB architecture and its corresponding functions. Secondly, we propose a SOA-based flexible architecture for WfMS, demonstrate its realization and summarize this paper.

2. SOA Summary and its Framework

Advantages of SOA. The concept of SOA is widely accepted as a software architecture design paradigm which promises the design and implementation of flexible systems and facilitates the change of business processes quickly. SOA leverages the alignment of business processes and information technology (IT). In SOA, main design element is "services". Services are self-describing and network-enabled components with well-defined interfaces that are implementation independent. Services can be assembled into business processes, made available for consumption by users, systems or other services. Applications use these services and communicate with each other through them. To reuse these services, W3C gives a SOA model in Fig 1:

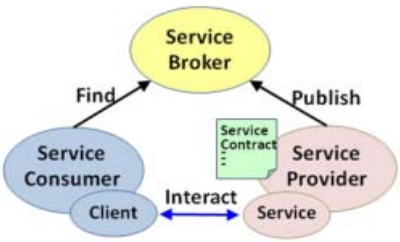


Fig. 1 SOA Model.

The three participators, service provider, service consumer and service broker interact with each other through three basic operations, such as publishing, finding and binding. Service provider publishes services to service broker, and service consumer looks up the required services through service broker and binds the found services.

SOA can be used to combine business services and IT resources. Based on components, SOA can transfer business process into a set of services linked to each other. Service consumers can use or access these services through network when necessary. SOA aligns the IT resource with business processes by changing its IT architecture, and this brings many benefits as follows:

(1) Maintaining the consistency of IT and business makes it easy to construct a reusable business application system with flexible structure.

(2) SOA provides an abstract layer through which enterprises can keep on using its IT investment, so as to get maximum utilization of IT assets.

(3) In SOA, systems are constructed by orchestrating different services which is loosely coupled, platform independent and access transparent. This makes systems much easier to integrate, manage and evolve, and impact on the changes of infrastructure and implementation can be minimized.

SOA lifecycle. Two key points of SOA are services and architecture. The process of SOA implementation is to determine the application architecture and define the application service components. The implementation of SOA consists of four stages: modeling, assembly, deployment and management, which is also called SOA life cycle, and this is very important for SOA governance. Firstly, after analyzing the business process, its services based model is constructed. Secondly, developers can construct new services and/or reuse existing ones, and then assemble them to create composite applications for process implementation. Then, all the resources constituting SOA should be deployed into a safety and integrated environment. Finally, all deployed SOA based systems must be managed and monitored from the IT/business point to identify problems and improve them.

SOA Development Framework. Service oriented in theory, is based on the concept of wrapping applications with well-defined interfaces so that the applications can be turned into a series of services. The wrapping process creates an abstraction layer which hides all the complex details of the application implementation. To achieve information resources sharing and applications cooperating with each other by SOA, a basic development framework for SOA is necessary, and is shown as Figure 2, which can be divided into five layers as follows:

(1) *Application layer*

All heterogeneous and off-sit application systems are in this layer. There may be some new systems and/or legacy systems within an enterprise and/or across the partners of the enterprise. It's difficult for these systems to directly communicate with each other without calling interfaces.

(2) *Component layer*

Every function component from application layer is encapsulated to form the corresponding component service for service layer.

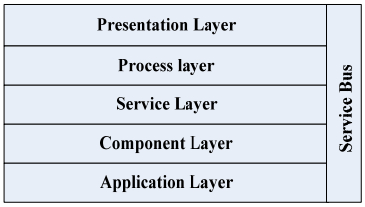


Fig. 2 The SOA development framework.

(3) *Service layer*

Service layer calls component layer and provides the corresponding services for the upper layer according to required granularity. It is the core layer in SOA.

(4) *Business process layer*

In this layer, business processes are constructed by all kinds of encapsulated services and all the services are well managed for later service reuse and maintainence.

(5) *Presentation layer*

This layer is not only used to display the final integration result of a global business process in a client end, for the enterprises distributed in different areas, to make real time inquiry, but also provides users with interfaces for calling process service so as to realize seamless business integration.

Even through workflow technologies and products have developed a lot in the past few years, but the traditional workflow process model binds business processes and enterprise resources, and this leads to the over tightness combination of business model and organization & resource model, insufficient support for cross-organizational workflows, and therefore it cannot meet the requirements for enterprise dynamic changes and developments. So it is necessary to investigate some new workflow techniques, which are more flexible and facilitate for high efficient implementation of business processes.

Workflow Reference Model. WfMC proposed workflow system reference model as in Figure 3.

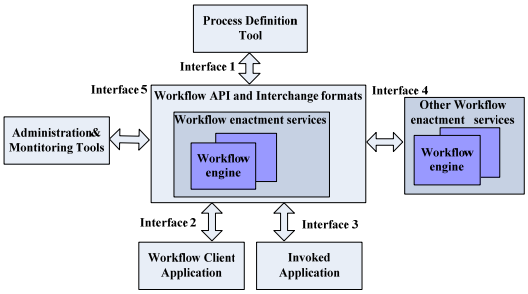


Fig. 3 Workflow reference model.

(1) Interface 1 (process definition tools): it provides users with a visual tool which is used to create the process definition model.

(2) Interface 2 (workflow client applications): through this interface, users have access to worklist and get tasks. Workflow engine assigns tasks to its client end applications.

(3) Interface 3 (invoked applications): it aims to establish a connection between workflow engine and other applications, and the engine can get workflow relevant data.

(4) Interface 4 (other workflow enactment services): it targets to solve the interoperability among heterogeneous workflows, such as basic data exchange operations among different workflow systems.

(5) Interface 5 (administration and monitoring tools): it is used to monitor & manage workflows, such as role & user management, resource control and process supervisory.

3. A WfMS Architecture Based on SOA

WfMS Architecture Based on SOA. According to workflow reference model from WfMC, this paper proposes a flexible WfMS architecture, which is based on SOA and shown in Figure 4. In this architecture, every module is realized by a service, and services can be published and found in the UDDI registry through service management module. All modules communicate with each other by ESB and all enterprise legacy systems can be integrated by ESB. The proposed service-oriented workflow management system framework is composed of process definition service, task list management service, monitoring and management service, workflow engine service etc [8]**.**

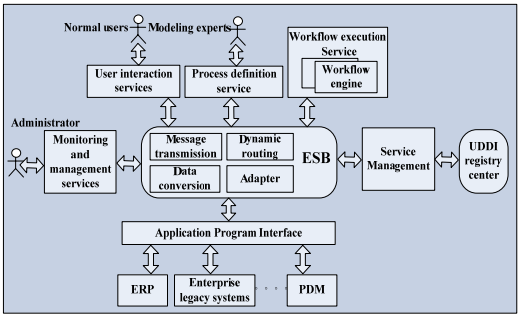


Fig. 4 The architecture of WfMS systems based on SOA.

(1) *ESB*

ESB is the core of this architecture, and it serves as the most important infrastructure to provide support for data interoperability and system integration.

(2) *Workflow enactment service*

Workflow enactment service consists of one or more workflow engines. It provides the run-time environment for process instances and navigates process execution. It interprets process definition and controls the process instances and activity sequences, adding work items to user worklists and invoking application tools when necessary. Workflow enactment service not only maintains workflow control and relevant data, but also interacts with external resources.

(3) *User interaction service*

User interaction service includes user interface and woklist handler. Users can login in systems and get their corresponding authority by user interface. User interface is a separate software component, and responsible for the interaction between users and enactment services, while worklist handler is a component for managing the interaction between workflow participants and workflow enactment service and thus progresses business processes.

(4) *Monitoring and management services*

Monitoring and management services include 3 functions:

•User mangement operations: Establish and/or delete the authorities of users or workgroups.

•Role management operations: Define, delete, modify role relationships and set new role attributes.

•Process supervisory operations: Change the state of processes or activity instances for a specified type, such as gets the details of process or activity instances, enables/disenables a particular version of process definition, and terminates all process instances.

(5) *Service management*

Service management serves as a service broker, which receives service request, and then look up the requested service from UDDI registry. All published services can register in UDDI through service management.

(6) *Application program interface*

It provides the integrated interfaces for enterprise legacy systems and other applications such as PDM, ERP. These applications are packaged as services to be called by other components.

SOA facilitates aligning existing IT infrastructure and systems to achieve end-to-end enterprise integration. By adopting an SOA approach and implementing it using supporting technologies,

companies can build flexible systems that implement changing business processes quickly, and make extensive use of reusable components. SOA supports an information environment built upon loosely coupled, reusable and standards-based services. This approach can minimize the interdependencies between applications, so adding new applications or replacing modules or changing operations within individual business processes do not impact the other applications or modules. This will enable faster modifications to business application, easy integration of applications and rapid deployment of new solutions resulting in business agility. Enterprise will respond to changing business requirements maximizing its adaptability, while adaptability and agility will result in high enterprise flexibility.

Principle and Architecture of ESB. ESB is an enterprise level SOA, which is loosely coupled. ESB provides an open, standard-based message mechanism, as well as the interoperability between rough granularity applications and other components, and then meets the integration requirements for large heterogeneous IT environments. Its architecture is shown as Figure 5.

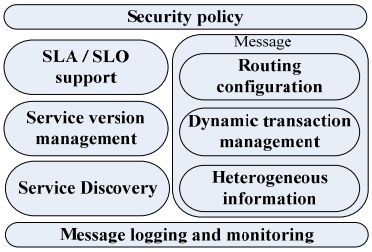


Fig. 5 ESB architecture.

(1) The principle of ESB

The fundamental principle of ESB is to integrate technologies, like SOA, web services and XML, into a unified and distributed architecture so as to construct an integrated infrastructure which can be easily deployed and managed. With ESB, new or SOA based applications can be integrated, and the existing applications, can also be integrated together by disassembling and packaging legacy systems to make them provide service interfaces. Furthermore, ESB bridges the new and existing application systems within and/or cross the enterprises, it serves as a service broker.

(2) The architecture of ESB

ESB is defined as a hub and has the following functions:

•*Data conversion and adapter:* ESB constructs connection between heterogeneous components by its predefined interface and contract.

•*Buffer:* ESB is responsible for converting business logic and data format between services to make services reusable.

•*Asynchronous message:* All services communicate with each other by SOAP based message.

•*Service discovery:* Users can dynamically discover the services they want by UDDI standards.

•*Intelligent routing:* ESB intelligently searches for the adaptable services through distributed management function and web service pool.

•Cross-platform interoperability: ESB uses open standard and non-exclusive technologies to achieve interoperability between different platforms.

4. Implementation

The implementation of our WfMS architecture mainly involves three aspects:

(1) *Realization of workflow engine and process design*

We can design the business process and realize the workflow engine by jBPM. jBPM is an open source, flexible and extensible framework based on JavaEE. It's a lightweight engine. The process of developing a WfMS as follows:

•Designing a business process definition by process designer of jBPM;

•Selecting a database which supports the Hibernate;

•Deploying the defined definition into a server;

•Executing the process by services API of jBPM.

(2) *Realization of Web Service*

Service component is the most important element in SOA. Web service is the most popular technology for implementing SOA. We develop web services by XFire. XFire is an open source service engine which is based on Java. Comparing with other service engines, it has simple configuration and high efficient. The process of developing a web service is as follows:

•Creating a java program which has a java interface class and corresponding implementation class;

•Creating a xfire servlet and defining a request path by modifying the web.xml file;

•Creating the service.xml file which defines the name and implementation class of a service;

•Publishing the defined service.

(3) *Realization of ESB*

We develop ESB by Apache Synapse, which is a lightweight, useable integration broker based on xml. It's the best choice for developing ESB.

5. Summary

There exist many inadequacies of traditional WfMS in supporting dynamic business environments, resulting from rigid process definition and too tightly organization & resource binding. By applying SOA approach into the design and implementation of WfMS, this paper proposes a flexible architecture for WfMS, which supports the easy integration of new and existing applications and processes. We also summarize some key prospects of the architecture implementation. Comparing with existing architecture for WfMS, the proposed architecture will, in one hand enable an enterprise to connect people, process and information in such a way that it becomes more flexible and responsive to the dynamics of its environment and competitors, and in the other hand enable it to align IT and its corresponding business goals, and eventually to promote an simple, adaptable and manageable enterprise architecture. With the further applications of SOA into WfMS, the flexibility of WfMS will be improved a lot, and this will better satisfy more complex requirements for enterprise applications.

References

[1] Workflow Management Coalition, The Workflow Reference Model Document Number WfMC-TC-1003, 19-Jan-95, 1.1, Workflow Management Coalition (1995).

[2] Gartner, Growing IT's Contribution: the 2006 CIO Agenda, Gartner EXP, American (2006).

[3] CHEN. Jinjun, Liu. Jianxun and S.C. Cheung: *A workflow engine-driven SOA-based cooperative computing paradigm in grid environments*, International Journal of High Performance Computing Applications, vol. 22 (Aug. 2008), p. 284-300.

[4] M. Zyla and D.Canban: *Dependability Analysis of SOA Systems*, IEEE, 3rd International Conference on Dependability of Computer Systems, Szklarska Poreba (June. 2008), p. 301-306.

[5] LIU Xiao-Lin and ZHENG You-cai: *Service-Oriented Workflow Technology*, Computer Engineering and Applications (in Chinese), vol. 42 (2006), p. 226-228.

[6] Ning Xiao and Badr. Y: *Conception and Development of Workflow Engine Based on Service Oriented Architecture*, IEEE, 1st International Conference on Digital Information Management (2007), p. 544-549.

[7] YANG Ming Kui, LIANG Hong Bing: *A Service-based Workflow Management System in Grid Environment*, IEEE, 19th International Conference on Advanced Information Networking and Application (2005), p. 293-297.

[8] WANG Shu-yang: *Workflow Management Design Based on SOA*, Journal of Changchun University of Technology (in Chinese), vol. 30 (Aug. 2009), p. 406-411.

Knowledge Based Interactive Smart Camera

Rustam Rakhimov Igorevich[1,2 a], Pusik Park[1,b], Jongchan Choi[1,c] and Dugki Min[2,d]

[1]SoC Platform Research Center, Korea Electronics Technology Institute, #68 Yatap, Bundang, Seongnam, Korea

[2]School of Computer Science and Engineering, Konkuk University, Hwayang-dong, Seoul, 133-701, Korea

[a]rustam@keti.re.kr, [b]parksik@keti.re.kr, [c]choijc@keti.re.kr, [d]dkmin@konkuk.ac.kr

Keywords: Smart Camera, Multimodal Interaction, Machine Learning, Logistic Regression, Neural Network, One-vs-all Support Vector Machine (SVM), iVision, Median Filter, Kalman Filter.

Abstract. In quick developing ubiquitous computing era smart systems are embedding into objects which are used in our daily lifecycle. Efficient and intelligent interaction becomes more important in pervasive computing. Especially when the user command and context should be recognized, interaction needs usage of additional intelligence. In this paper we propose the interactive iVision camera device, which adapts to exploitation environment and user context. This camera is able to predict user context. After training phase iVision camera comes up with quick advice for more reliable interaction. For the experimental purpose some of the basic machine learning algorithms was applied and according results were received.

Introduction

Recently society start facing smart environments where user context becomes important and interactions between user and machine becomes way too different than it was decade ago. Especially the introduction of MMI (Multimodal Interaction) changed total picture of the human machine interaction. W3C made MMI standard to give a structured view for different UI entities and represent them as modality components.

This research related to our previous works about hand tracking and gesture recognition applications based on stereoscopic camera solution. Wrong-distance problem started occurring, when the system for hand tracking and gesture recognition were tested by inexperienced users. The reason is that new user didn't have knowledge about actual distance range allowed by camera. To solve this problem adaptive distance solution using grayscale image histogram analysis method was applied [1]. Even though adaptive distance method gave expected results, new users required some additional advices in usage of our system. Every time when new inexperienced user was testing our system for the first time, they needed some advices. To overcome absence of adviser we decide to apply some of the machine learning techniques to provide user with useful hints in usage of hand motion interface.

Three methods of machine learning were applied on demand to implement interactive camera solution. Multi-class classification (with logistic regression kernel), neural network learning and Multiclass SVM methods were selected to apply for this problem.

This research work organized as follows: first of all some of the background and related works are listed out. After that application of Multiclass Classification, Neural Networks learning and Multiclass SVM methods are demonstrated in according order. A numerical result of applying machine learning methods comes along.

Related works

Similar work was done for Kinect Natal project where Xbox developers from Microsoft implemented user adviser for wrong-distance problem. It has two optional informing messages for user: “too far way” and “too close” [2]. Since Kinect generates depth map with has high accuracy it can easily track the user in certain area. But even in case of kinect, when user stands at the edge of kinect detectable area, the application struggles to perform hand tracking. To generate advice message for user, Kinect solution applies simply area thresholding method.

Chan et al. [3] is the first who applied AI to a smart home automation system for use by the elderly and disabled people. His system could keep up with twelve rooms with sensors deployed in it, to collect context information.

Another researchers Jorge and Goncalves[4] focused on automated health monitoring of elder people in their everyday life using artificial intelligence tool. They used ubiquitous computing devices to collect data from the elderly health information. They first tightly relate ubiquitous computing devices with AI methods to predict next status. Main issue was distinguishing patients who require constant medical monitoring and other type of patients who might need special needs.

In Rezauld et al. [7] an integrated framework with potential application of ANNs was involved for smart home. Central computer operates a number of ANN models in parallel, where different automated devices have different architectures. Data bus is responsible for collecting raw data from different modules such as caregiver, entertainment, security & safety etc. Developed middleware section extracts and selects the important features from the acquired data, and then passes these features to the relevant ANNs in the computer.

Problem definition

Interaction in Ubiquitous Computing. Ubiquitous systems need high level and natural way of interaction with human. The beauty part of ubiquitous technologies related to the fact how it becomes part of our daily routine, while we didn’t fully realize it cognitively. It shows the fact how much the user interface have become natural towards to machines.

The main contribution for the Ubiquitous Computing is hardware implemented iVision camera, which calculates depth map of the upfront scene image, by comparing two shifted images from two cameras. Our hardware architecture performs few more operations over the extracted depth maps, such as thresholding, masking, searching for blob and listing its properties, classification of blobs and rearranging them in order.

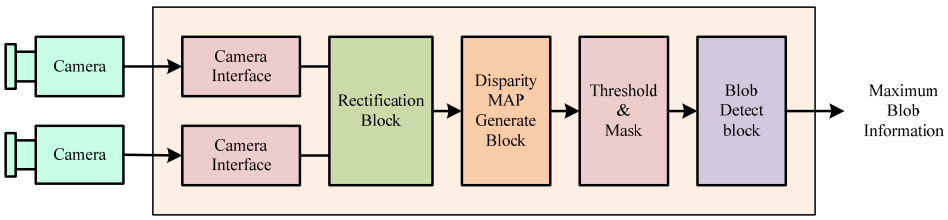


Fig.1. Basically Proposed Gesture recognition hardware architecture

The block “Threshold & Mask” is masking the disparity map image by fixed value for threshold. It separates the background and the fixed objects in sequential image frames. Typically, objects for the gesture recognition, such as outstretched hand and finger, locates closer to the camera compared to other parts. The disparity map indicates for the distance between the camera and the object, which also called as a depth map. Finally, the Blob detection block searches the masked image and finds blob and calculate center position coordination.

This proposed architecture was implemented on top of the FPGA (Xilinx Vertex5), and it gives satisfied results. Next, iVisionTest application was designed and implemented. Now it has access to

the FPGA generated information, for the further processing. It provides software engineers with sort of possibilities to perform additional operations, providing additional interactions through the network with other smart devices.

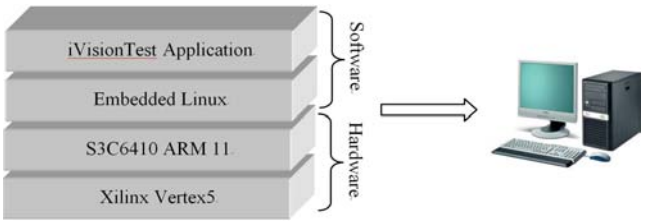


Fig.2. Overal View of Main Components

iVisionTest application contains two interfaces to access stored blob information on FPGA board: SDRAM and I2C interfaces. When the blob information is read from the registers of FPGA they should be compound into variables. The preprocessing of blob properties requires application of simple operations such as shifting and some other logistic operations. When blob and other necessary data are extracted, further processing can be performed. In our case testing goal was encapsulate these variables into packets and transmit them through the network to the remote server, for further performance analysis.

Fig.3. Result of Applying (Left)Median and (Right) Kalman Filters for Blob Center illustration

Before sending the raw blob information different kind of smoothing filters such as Median and Kalman filters were also applied to track blob center information. Results of filtered coordinates also were sent to the remote server. Some are the snapshots from the iVisionTest application are illustrated in Fig.3. Implementation of Kalman filter gave good result for smoothing the mouse pointer. Using kalman filter user of iVision system can track his/her hand very smoothly.

When the system for hand tracking and gesture recognition were tested by inexperienced users the distance problem occured. Every time when new inexperienced user was testing our system for the first time, they need some advices. To overcome the absence of adviser we decide to apply some of the machine learning techniques to provide user with useful hints in usage of hand motion interface. To do this we used supervised learning algorithms which can be used for expert systems. First of all training data were collected and saved.

Applying Machine Learning Techniques

As it was mentioned before three methods of machine learning were applied: Multi-class classification, Neural Network learning and SVM Multiclass. In further, more details about implementation of these algorithms will be described. These three algorithms able to cover the solution for the problem related to horizontal noise problem generated by hardware layer during depth map calculation process. For both of these methods four features were selected: center

coordinate (contains x and y values), width and number of pixels in biggest blob. Ultimate goal was how to classify these input values to five output values such as: forward, little forward, backward, little backward and stay there. Accordingly we number these values from 1 to 5.

Heuristic functions of first two selected methods (Multiclass classification and neural networks) are bases on sigmoid function. That meant output should be represented in binary form, where Y will contain vector with five values where corresponding indexed value will contain 1, others are 0.

$$\text{Ex: } 4 = \begin{bmatrix}0\\0\\0\\1\\0\end{bmatrix} \quad 3 = \begin{bmatrix}0\\0\\1\\0\\0\end{bmatrix}$$

Multi-Class Classification. For multi-class classification we used One-Vs-All method, where in every training cycle the learning class will be considered as a positive all other classes as a negative.

$$h_\theta^{(i)}(x) = P(y = i|x;\theta) \qquad (i = 1,2,3) \tag{1}$$

Since multi-class classification is uses logistic regression, heuristic function is defined by sigmoid function.

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{2}$$

Cost function of logistic regression can be defined as:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right] \tag{3}$$

To perform Gradient Descent and minimize cost function J(θ), value of θ should be recursively calculated.

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) - x_j^{(i)} \tag{4}$$

After implementation of these algorithms in Octave programming environment, we receive successful results. Learning process was performed 5 times for five different classes, where for every class 50 iterations were performed. The learning process was performed on 5000 training data, where every class has 1000 different training data. In result we received learning accuracy equal to 83.529882%

Neural Networks Learning. Neural networks are the imitation of human brain. Threre are a lot of information in web for related topics [5]. In this work we applied and tried neural networks with two different architectures.

Compared to previously described method neural networks uses first derivative of sigmoid function for Back Propagation algorithm. According to heuristic function with gradient sigmoid neural networks have regularized cost function (5). Sigmoid function and regularization equation used to calculate cost function.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[-y_k^{(i)} \log\left(\left(h_\theta\left(x^{(i)}\right)\right)_k\right) - \left(1 - y_k^{(i)}\right) \log\left(1 - \left(h_\theta\left(x^{(i)}\right)\right)_k\right)\right] + \frac{\lambda}{2m}\left[\sum_{j=1}^{25}\sum_{k=1}^{4}(\theta_{j,k}^{(1)})^2 + \sum_{j=1}^{10}\sum_{k=1}^{25}(\theta_{j,k}^{(2)})^2 + \sum_{j=1}^{5}\sum_{k=1}^{10}(\theta_{j,k}^{(3)})^2\right] \tag{5}$$

During our research we tried neural network in different structures. For comparison we will describe only two of them, where first architecture of neural network has simple architecture. It consist three layers and one hidden layer, where hidden layer contains 25 nodes, to map four inputs to five outputs. After performing back propagation algorithm our training accuracy was 68.898661%.

It shows that, neural network with one hidden layer is not proper for training this system. On the next Neural Network with double hidden layer was designed for current problem. (Fig.4)

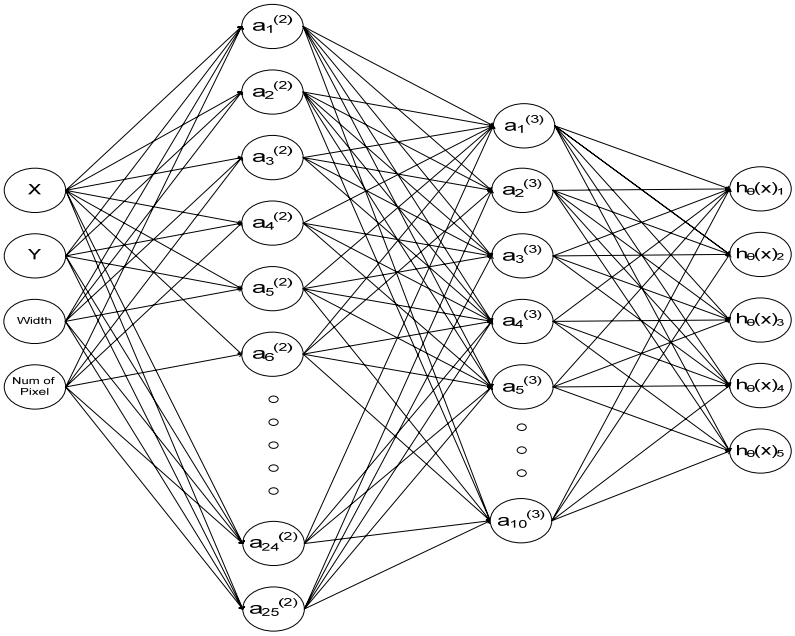


Fig.4. Illustration of NN with two Hidden Layers

When Neural Network with two hidden layers was applied to the defined problem it gave better result compared to Neural Network with one hidden layer. By using trial and error method Neural Network with 25 nodes in first hidden layer and 10 nodes in second hidden layer were selected. This network gave 72.14% accuracy with given training set data, which is 3.25% accurate compared to Neural Network with one layer. Other tries with different layers and nodes of Neural Network gave results around these values.

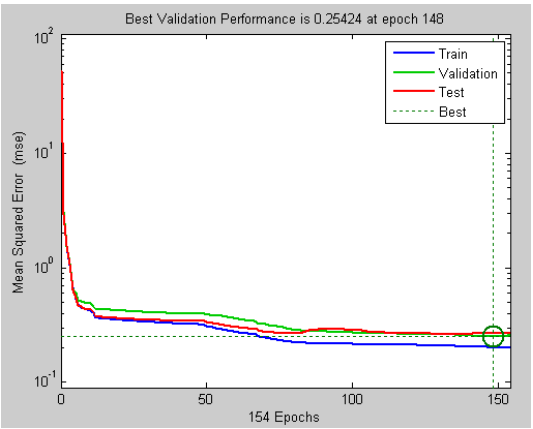


Fig.5. Illustration of Degrading Mean Squared Error

For the sake of minimizing cost function with 154 epochs was made, final Cost function was equal to 0.25424 when the mean squared error was almost equal to zero. Performance of degrading mean square error rate is illustrated in Figure.5.

Multiclass SVM. Multiclass SVM (Support Vector Machine) [6] method is also applied to the training set which was used for previous two algorithms. Originally SVM belongs to binary classifiers, but easily can be combined to handle multiclass cases. Simple and effective combination trains One-Versus-All classifiers for the N-class case. Gaussian kernel was selected as a base function

of SVM.

$$K_{gauss}(x^{(i)}, x^{(j)}) = e^{-(\frac{\|x^{(i)}-x^{(j)}\|}{2\sigma^2})} \quad (6)$$

Comparatively to previous two methods Multiclass SVM gave highest accuracy on training set. Since implementation of SVM One-Versus-All was already has been implemented, we just used solution provided by LIBSVM [8] library. After experiment we got accuracy equal to 97.5015%.

Conclusion and Future Extensions

Currently our system uses pre-generated feature coefficients, since we didn't provide our system with real time learning. This requirement comes from main principles of neural network applications, where preferably neural networks should be applied in an off-line fashion. iVision system already provides network access, we can apply usage of cloud computing techniques to overload some of the learning processes to design semi real-time learning system. When user runs iVision system for the first time in a new environment it will use pre-generated feature coefficients. Later when iVision is in non-actively using mode, it can send new captured values of current user to the cloud server. Cloud server will perform re-training operation based on received training data. When gradient descent algorithm performance completed with minimal value of cost function, cloud system respond new feature coefficients to the iVision system.

Acknowledgment

This research was supported by Industrial Source Technology Development Programs and "The next generation core technology for Intelligent Information and electronics" project funded by the Ministry of Knowledge Economy (MKE) of Korea.

References

[1] Igorevich, R.R., Pusik Park, Dugki Min, in: *Application of Information and Communication Technologies (AICT), 2010 4th International Conference on*, "Hand gesture recognition algorithm based on grayscale histogram of the image". 28 October 2010, Tashkent

[2] http://support.xbox.com/en-US/kinect/more-topics/errors/standing-too-close-too-far-away

[3] Chan, M., C. Hariton, P. Ringeard, E. Campo, "Smart House Automation System for the Elderly and the Disabled", *IEEE international conference on Systems, Man and Cybernetics*, 1995, Vol. 2, pp. 1586-1589.

[4] Jorge, D. and Goncalves, V. "Ubiquitous Computing and AI Towards an Inclusive Society", *Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing: providing for the elderly*, 2001, pp. 37-40.

[5] Arbib, Michael A. (Ed.). *The Handbook of Brain Theory and Neural Networks*. (1995)

[6] Kai-Bo Duan and S. Sathiya Keerthi. "Which Is the Best Multiclass SVM Method? An Empirical Study".*Proceedings of the Sixth International Workshop on Multiple Classifier Systems*. (2005)

[7] Rezauld Begg and Rafiul Hassan. In: Designing Smart Homes The Role of Artificial Intelligence. " *Artificial Neural Networks in Smart Homes*", (2006)

[8] Library from http://www.csie.ntu.edu.tw/~cjlin/libsvm/

2012 International Conference on Affective Computing and Intelligent Interaction
Lecture Notes in Information Technology, Vol.10

Performance Evaluation of Modern Intel x86 Processors through Computer Capacity

Boris Ryabko[1,a] and Andrey Fionov[1,b]

[1]Siberian State University of Telecommunication and Information Sciences
Institute of Computational Technologies SB RAS
Kirov St. 86, Novosibirsk 630102 Russia

[a]boris@ryabko.net, [b]a.fionov@ieee.org

Keywords: performance evaluation, benchmarks, computer architecture, instruction set

Abstract. We apply a notion of computer capacity as a novel approach to evaluation of computer performance. To show the benefits of the suggested approach, we determine the capacities of modern processors of Intel x86 family and compare the results to the metrics obtained from some available benchmarks. We show that our theoretical evaluation of computer performance conforms well to that given by the benchmarks.

Introduction

Computer capacity, as a new theoretical metric for computer performance evaluation, was first suggested in [1, 2] and then in [3], where it was applied to Knuth's MMIX machine. Later the concept was extended to modern computer architectures that incorporate such features as cache memory, pipelines and parallel processing units, as well as multicore and hyperthreading technologies. Estimations of computer capacity to some older x86 processors, starting from 80286, were reported in [4]. In the present paper we provide estimations of computer capacities for modern x86 processors, ending in i7 devices.

Why is it important to search for new methods of computer performance evaluation? It is clear that various aspects of performance are the key goals of any new computer design. Simple performance metrics, such as the number of integer or floating point operations executed per second, are not adequate for complex computer architectures we face today. A more appropriate and widely used approach is to measure performance by execution time of specially developed programs called benchmarks. The main issues of benchmarking are well known, we only mention a few. First, it is very difficult, if ever possible, to find an adequate set of tasks (in fact, any two different researchers suggest quite different benchmarks). Then, when a benchmark is used at the design stage, it must be run under a simulated environment which slows down the execution in many orders of magnitude, making it difficult to test various design decisions in the time-limited production process. As a consequence, the designers reduce the lengths and the number of benchmarks, which raises the question of conformity with real applications. Quite often, benchmarking is applied to already made devices for the purposes of evaluation and comparison. Here, the benchmarks produced by a hardware manufacturer may be suspected of being specially tuned just to facilitate sales. The benchmarks suggested by independent companies are prone to be outdated when applied to technologically novel devices. All these appeal to objectivity of evaluation results. The performance figures obtained in this way may be suitable for one kind of applications but useless for another.

Evaluation of computer performance by means of computer capacity is a completely different approach which allows to circumvent the difficulties outlined above. The new approach is based on

calculation of the number of different tasks that can be executed in time *T*. This is quite similar to determining the channel capacity in information theory through the number of different signals that can be transmitted in a unit of time [5]. The number of different tasks does not depend on any particular application field and is determined only by the computer architecture which, in turn, is described by the instruction set, execution times of instructions, structure of pipelines and parallel processing units, memory structure and access time, and some other basic computer parameters. All these parameters can be set and adapted at the design stage to optimize the performance.

Describe briefly the essence of computer capacity and the way of its estimation. Denote by $I = \{u_1, u_2, \ldots, u_s\}$ the instruction set of a computer (processor). An admissible sequence of instructions $X = x_1 x_2 \ldots x_t$, $x_i \in I$, seen as a process in time, is called a computer task. The term "admissible" means that the instruction sequence *X* can be executed up to the last element without errors in computation (so-called exceptions), such as division by zero or illegal memory reference. We consider two tasks *X* and *Y* as different if they differ at least in one instruction, i.e., there is an *i* such that $x_i \neq y_i$.

Denote the execution time of instruction *x* by $\tau(x)$. Then the execution time $\tau(X)$ of a task *X* is given by

$$\tau(X) = \sum_{i=1}^{t} \tau(x_i).$$

The number of different tasks whose execution time equals *T* may be written as

$$N(T) = |\{X : \tau(X) = T\}|.$$

Note that in modern processors, $\tau(x)$ can be measured in the number of processor cycles.

The computer capacity $C(I)$ is then defined as

$$C(I) = \lim_{T \to \infty} \frac{\log N(T)}{T}. \qquad (1)$$

The definition of computer capacity is quite general. It does not restrain us from using one or other model of computer task formation. We may apply restrictions on instruction sequences, consider dependence of instruction execution times upon preceding instructions, and so on. Generally, the calculation of the limit in Eq. 1 becomes a complicated combinatorial problem. But as a first step, we can use a simple method suggested by Shannon in [5] for finding the capacity of noiseless channel where code symbols had different durations. When we use this simple method, we assume that all sequences of instructions are admissible. Clearly, by doing that we obtain an upper bound of capacity, which we denote by $\widehat{C}(I)$, because the number of admissible instruction sequences $N(T)$ cannot be larger than the number of all possible sequences, denoted thus by $\widehat{N}(T)$. Despite this simplification, we take proper account of the effects of caches, pipelines and parallel processing. More specifically, following [5], for the instruction set $I = \{u_1, u_2, \ldots, u_s\}$ we may state that the number of all possible instruction sequences must satisfy the difference equation

$$\widehat{N}(T) = \widehat{N}(T - \tau_1) + \widehat{N}(T - \tau_2) + \cdots + \widehat{N}(T - \tau_s).$$

Here $\widehat{N}(T - \tau_j)$ is the number of instruction sequences of duration *T* ending in instruction u_j.

It is well-known from the theory of finite differences that asymptotically, as $T \to \infty$, $\widehat{N}(T) = Z_0^T$, where Z_0 is the greatest positive root of the characteristic equation

$$Z^{-\tau(u_1)} + Z^{-\tau(u_2)} + \cdots + Z^{-\tau(u_s)} = 1.$$

So from the definition of computer capacity Eq. 1 we have

$$\widehat{C}(I) = \log Z_0.$$

In what follows we will estimate $\widehat{C}(I)$ as a first approximation of real computer capacity, realizing that there are more complicated and more exact methods of finding $C(I)$.

Consider some examples. Let the first computer has only two instructions and execution time of each instruction is one clock cycle. So we have $I_1 = \{u_1, u_2\}$, $\tau(u_1) = \tau(u_2) = 1$ and the characteristic equation is $2Z^{-1} = 1$. Hence $Z_0 = 2$ and the computer capacity $C(I_1) = \log 2 = 1$ bit per cycle. Now add a third instruction with duration 2 cycles: $I_2 = \{u_1, u_2, u_3\}$, $\tau(u_1) = \tau(u_2) = 1$, $\tau(u_3) = 2$. The characteristic equation is $2Z^{-1} + Z^{-2} = 1$, its greatest root $Z_0 = 2.414$. The capacity $C(I_2) = 1.27$ bit per cycle, it is greater than $C(I_1)$ due to "more rich" instruction set I_2.

In practice, the computer instructions are often built of operation codes and operands, which may be references to internal registers, memory, or some immediate data. The key point is that to find the computer capacity we must consider the instruction set containing all operations with all combinations of operands. Let, for example, the computer have 8 registers, 2^{16} memory locations, and can perform two operations op1 and op2 of the following format: (op1 reg reg) and (op2 reg mem), where reg is one of 8 registers, and mem is a reference to one of 2^{16} memory locations. Let op1 require 1 cycle and op2 2 cycles. Then the characteristic equation will be

$$\frac{8 \cdot 8}{Z} + \frac{8 \cdot 2^{16}}{Z^2} = 1$$

The solution $Z_0 = 757$ and $C(I_3) = 9.56$ bits per cycle.

Of course, the characteristic equations for modern x86 processors are very huge since their instruction set is quite rich. Nevertheless, the way of constructing the equations is straightforward and follows the considered examples. The results of computer capacity estimation and comparisons with metrics of benchmarks are given in the next section.

Computer Capacity of x86 Processors

The basic characteristics of the processors that are essential for our calculations are provided in Table 1. In Table 2 the computer capacities are given. Since the clock frequencies for different processors are different, the capacities obtained initially in bits per cycle, are multiplied by the number of cycles per second (i.e., the clock frequency) to produce comparable results in bits per second.

Now we compare the estimates of computer capacities shown in Table 2 to the results of benchmarks. On the Internet, we have found the results of three sets of benchmarks applied to the same processors. The sites were selected as "first found" without any filtering. The benchmarks are: SiSoftware Sandra 2010 Pro, PCMark Vantage (both from `http://www.tomshardware.com`) and PassMark (from `http://www.passmark.com`).

Table 1. Main Characteristics of Considered x86 Processors

Computer CPU	Num. of Cores	Clock Freq. [MHz]	RAM Size [M byte]	L1 Cache [K byte]	L2 Cache [K byte]	L3 Cache [K byte]
Pentium E5300	2	2600	2048	2×32	2048	–
Pentium E6500	2	2930	2048	2×32	2048	–
Core 2 Duo E7500	2	2930	4096	2×32	3072	–
Core 2 Duo E8400	2	3000	4096	2×32	6144	–
Core 2 Quad Q8300	4	2500	4096	4×32	4096	–
Core 2 Quad Q9450	4	2660	4096	4×32	12288	–
Core 2 Quad Q9550	4	2830	4096	4×32	12288	–
Core i7 870	4	2930	4096	4×32	4×256	8
Core i7 920	4	2670	4096	4×32	4×256	8
Core i7 960	4	2930	4096	4×32	4×256	8

Table 2. Computer Capacities of Considered x86 Processors

Computer CPU	Computer Capacity [Gbit/s]
Pentium E5300	280
Pentium E6500	315
Core 2 Duo E7500	315
Core 2 Duo E8400	323
Core 2 Quad Q8300	538
Core 2 Quad Q9450	572
Core 2 Quad Q9550	609
Core i7 870	906
Core i7 920	825
Core i7 960	989

The marks assigned by different benchmarks are not mutually compatible since they are measured in quite different abstract units. So any direct comparison of the marks with the values of computer capacities from Table 2 is meaningless. Nevertheless, to advocate the practicality of computer capacity, we make simple normalization of the numbers and plot the results as the diagrams shown in Figs. 1–3. The normalization was done by considering the mark for Pentium E5300 as the unity within each benchmark. Similar normalization was done for computer capacity.

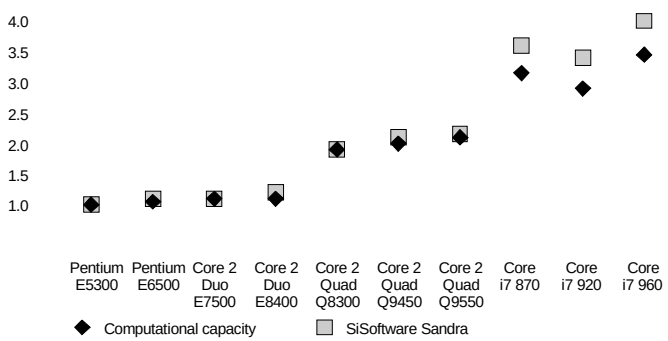


Fig. 1. Computer capacity against SiSoftware Sandra 2010 Pro

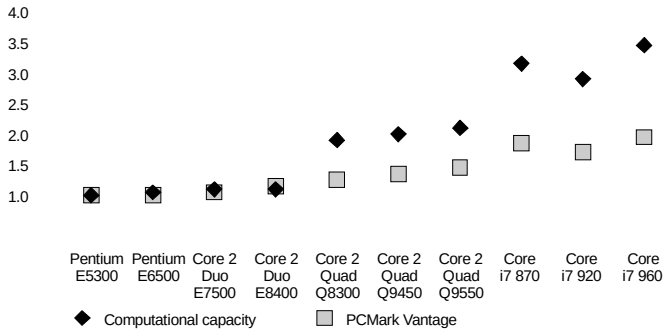


Fig. 2. Computer capacity against PCMark Vantage

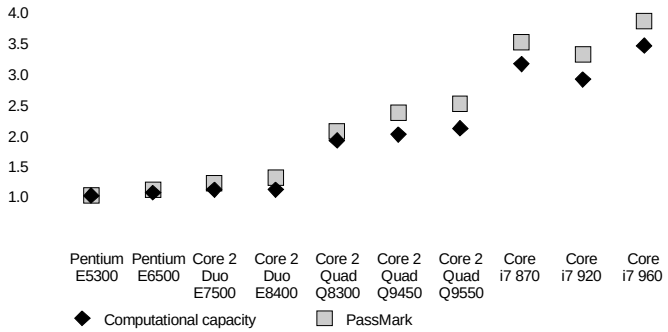


Fig. 3. Computer capacity against PassMark

We can see that the graphs of benchmarks generally show the similar tendency in performance evaluation as the computer capacity when proceeding from one processor to another. Figure 2 demonstrates that the PCMark Vantage benchmark does not take advantage of 4-core processors (probably, it uses only two cores). So the computer capacity, as a new measure of computer performance, is rather adequate and, as it was noted earlier, can be easily computed.

Summary

We presented the application of a new approach to evaluation of computer performance based on the notion of computer capacity. A distinctive feature of the method is that it does not require to have a hardware for testing in order to produce an evaluation of performance. The comparison of several generations of x86 processors shows that the suggested notion of computer capacity quite adequately reveals the difference in computational capabilities of existing devices, which is confirmed by comparison with the data from benchmarks. We believe, it can become a useful tool for computer designers in their search for the best architectural parameters of computer systems.

References

[1] B. Ryabko: Applications of Information Theory to analysis of efficiency and capacity of computers and similar devices, in Proc. IEEE Region 8 SIBIRCON-2010, Irkutsk Listvyanka, Russia, July 11–15 (2010), pp. 11–14.

[2] B. Ryabko: Using Information Theory to study efficiency and capacity of computers and similar devices, in *Information*, 1 (2010), pp. 3–12.

[3] B. Ryabko: On the efficiency and capacity of computers, in *Applied Mathematics Letters*, vol. 25 (2012), pp. 398–400.

[4] A. Fionov, Yu. Polyakov and B. Ryabko: Application of computer capacity to evaluation of Intel x86 processors, in 2nd Int. Congress on Computer Applications and Computer Science (CACS-2011), November 15–17 (2011), Bali, Indonesia.

[5] C. E. Shannon: A mathematical theory of communication, in *Bell Sys. Tech. J.*, vol. 27 (1948), pp. 379–423.

Positional Conformity Degree Checking Method of Spatial Data Quality*

Dou Shiqing[1,2,a], Du Jiliang[2,b] ,Yu Fujun[3,c]

[1]College of Geoscience and Surveying Engineering CUMTB, Beijing, China

[2]College of Resource and Environment Engineering, Heilongjiang Institute of Science and Technology, Harbin, China

[3] Jixi Coal Mining Corporation, Jixi, China

[a] 37929972@qq.com,[b]3982159@qq.com,[c]149319929@qq.com

Keywords: positional conformity degree; quality; spatial data; assessment

Abstract. Traditional checking and assessment methods of spatial data quality emphasize on unit production. During the process of checking and assessment for unit production, the known higher precision data are required as reference. Under the condition of lacking of those data, the checking and assessment work can not proceed. But it is possible to process the check of positional conformity degree from the association aspect of multi-source data. The spatial data checking can be finished by the positional conformity degree .In the paper, the novel concept and application of positional conformity degree were putted forward and the theoretical system and methods were constructed. In order to solve the positional precision checking problem for DLG, the paper adopts the positional conformity degree between DLG and the DOM and DEM. Two DLG products of Western Mapping Project were selected for experiment, and the detailed procedure was depicted in the paper.

Introduction

The positional conformity degree was defined as the conformity degree of same features in same areas in different spatial data. The positional conformity degree checking of multi-source spatial data is the consistency of the conformity degree checking to determine the data whether there is systematic error or exceed the required precision. It concludes the check of plane positional conformity degree and elevation positional conformity degree. It can be used to check the unknown precision spatial data and also can be used to assess the whole spatial data quality. The positional conformity degree checking is proceeded by the conformity of obvious point, line and surface features, and it can realized the spatial data checking and assessment by computer in order to improve the spatial data quality [1].

Traditional checking and assessment methods of spatial data quality emphasize on unit spatial data production. During the process of checking and assessment for unit spatial data production, the known higher precision data are required as reference. Under the condition of lacking of those data, the checking and assessment work can not proceed. But the spatial data acquisition have the characteristics of high speed, diversified forms, and multi-source, multi-scale, multi-temporal and multi-theme, so it is possible to process the check of positional conformity degree from the association（namely the same features in same area of different spatial data source should overlay completely） aspect of multi-source data. If these are in the same area, different spatial data sources overlay completely that it have a high positional conformity degree which shows that multi-source

* Supported by Fund of Heilongjiang Provincial Education Department(No. 11544043).

data has a high relative position accuracy, the produced mass and multi-source spatial data have higher quality or the relative position accuracy of the produced multi-source spatial data is low, or even errors. The spatial data checking can be finished by the positional conformity degree, so it is important to study the checking methods of positional conformity degree.

Spatial data includes vector data, image data and DEM data, so the positional conformity degree checking of multi-source spatial data can be divided into four categories:

- The positional conformity degree checking between vector data and vector data;
- The positional conformity degree checking between vector data and image data;
- The positional conformity degree checking between image data and image data;
- The positional conformity degree checking between DEM and vector data.

Ⅰ *The checking methods and processes of positional conformity degree*

The positional conformity degree checking of multi-source spatial data measure the degree of same geographic elements in geographic coordinates space by some methods, during the checking process, usually by calculating positional conformity degree of obvious point, line or area feature to check positional conformity degree. The positional conformity degree of different types of data can be calculated by different methods.

A. Checking Methods

1) The positional conformity degree checking based on Generalized Hough Transform [3].

The positional conformity degree checking based on Generalized Hough Transform method can be used between the DLG and DOM conformity checking, the basic idea is based on the vector elements boundaries of DLG to establish two-dimensional form, and then through the Hough transform to detect the corresponding image area, at last through concrete evaluation indexes to check the conformity checking of the images and vector graphics, in order to obtain the accuracy of vector data. The positional conformity degree checking based on Generalized Hough Transform common indicators are: degree of polymerization, the position offset index, in favor of the index.

2)The positional conformity degree checking based on information theory measurement

We can see from the information theory, mutual information describes the contained degree of matching information, in the quality assessment process, through the comparison between the evaluation data set of and the reference data set to assess the quality of spatial data products, the better the positional conformity degree of evaluate the data and reference data is, the greater the calculation of the two mutual information is. Conditional information need to compare the reference data (signaling) and the target data (received signal) for matching and comparison, it represents the loss of information channels. The greater the loss the signal is, the greater the conditional information entropy is, the worse the positional conformity degree of evaluation data and reference data is, the worse the quality of the object to be evaluated is. Therefore, mutual information and conditional information entropy can be used as the index of data overall quality.

The positional conformity degree checking based on information theory measurement can be carried out between the vector data. The calculation of information and mutual information base on the characteristics feature entities of the two data sets, computational process is as follows:

a)scale-scale conversion

If the multi-source data scale is inconsistent, you need to do scale-scale conversion, making corresponding entity with the same scale.

b) The extraction of vector elements

The positional conformity degree checking may contain DOM and DEM data. The positional conformity degree checking of DOM needs the extraction of significant points, line elements or closed areas or edge of areas. The positional conformity degree checking of DEM needs extraction of feature lines.

c)The calculation of Mutual information and conditional information

The calculation of Mutual information and conditional information is based on the vector features entities between two data to complete. First, finding the corresponding pixel pair among two data sets, and then calculating the mutual information. To find the corresponding pixels can be completed by manual or buffer approach, specific information can refer the reference.

The formula of mutual information is

$$I_h(D_1, D) = \sum_{p \in h} I(angle(p); angle(h(p))) + \sum_{p \in h} I(length(p); length(h(p))) + \sum_{p \in h} I(form(p); form(h(p))) + \sum_{p \in h} I(position(p); position(h(p))) + \sum_{p \in h} I(connected(r_i, r_j); connected(h(r_i), h(r_j))) \quad (1)$$

In the formula, p is a matching pixel pair: angle, length, form, position and connected, is angle, length, shape, location and topology respectively [2].

D1 and D2 is respectively the collection of two data sets.

The calculation of conditions information: the conditions information equal to information minus mutual information.

3)the positional conformity degree checking based on error entropy with uncertainty band

BOS (buffer overlay statistics)method is based on the comparison between the data set of the unknown quality and the and the data set of the known or relatively high quality, finding the integrity line element data or panel data, including error, loss of error and other indicators. These indicators are set by the target data (data sets to be evaluated) X and the reference data set Q with the buffer (buffer) and overlay analysis (overlay) comparing obtain them, the steps of the calculation of line element uncertainty indicators are as follows [6][7] :

a)Buffer generation

Error entropy uncertain theory determine the with the Epsilon width as the width of the buffer, the line element of the data sets X and Q do the buffer. The results can get two other data sets XB and QB.

b)Overlay analysis

Two lines - surface overlay analysis: Data sets X and data sets QB, data sets XB and data sets Q, and thus get two new hybrid data sets XQB and XBQ.

c)The statistics of index value

- integrity

It can calculate the length of Q's line element into XB with the data set XBQ, and compare it with the total length of Q's line element. It indicates the approximate level of X and Q. The formula is

$$Completeness(X) = \frac{Length(Q \subset XB)}{Length(Q)} \quad (2)$$

- contained error

It can calculate the length of X's line element which fall outside QB with the data set XBQ , and compare it with the total length of Q's line element. This index indicates X contains the false data which related to Q s'. The formula is

$$Commission(X) = \frac{Length(X \not\subset QB)}{Length(Q)} \quad (3)$$

- loss error

It can calculate the length of Q's line element which fall outside XB with the data set XBQ, and compare it with the total length of Q's line element. It indicates X contains the loss data which related to Qs'. The formula is

$$Ommission(X) = \frac{Length(Q \subset XB)}{Length(Q)} \quad (4)$$

d) The calculation of accuracy region, false region, loss region and the area of QB, and calculate the uncertainty indicators of panel.

- integrity

$$Completeness(X) = \frac{A_{accuracy}}{A_{QB}} \quad (5)$$

- contained error

$$Commission(X) = \frac{A_{false}}{A_{QB}} \quad (6)$$

- loss error

$$Ommission(X) = \frac{A_{missing}}{A_{QB}} \quad (7)$$

4)The positional conformity degree checking based on all kinds of error indexes

The positional conformity degree checking based on all kinds of error indexes calculate through the point, line and area feature mainly.

a)point feature

The point error calculation of all checking points in the X and Y directions, for point feature is

$$M_x = \pm\sqrt{\frac{\sum_{i=1}^{n}(x_i - X_i)^2}{n-1}}, \quad M_y = \pm\sqrt{\frac{\sum_{i=1}^{n}(y_i - Y_i)^2}{n-1}}$$
$$M_{xy} = \pm\sqrt{M_x^{\,2} + M_y^{\,2}} \quad (8)$$

In (8), M_x and M_y are the error in points X, Y direction, x_i、y_i is the sample points on the graph X, Y coordinates , X_i、Y_i is the corresponding point coordinates of the same name; M_{xy} is the point error.

b)linear feature

The positional conformity degree for linear feature mainly calculate the maximum linear error, the average error, the center distance error, etc which corresponds to the linear target node and calculate the maximum linear error, the average error, the center distance error of all the linear target node.

- node maximum error:

$$d_{\max} = \max(\sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2}) \quad (9)$$

In (9), (x_i, y_i) , $(x_i^{'}, y_i^{'})$ is the corresponding node coordinates respectively.

- Average error:

$$d_{average} = \sum(\sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2}) / n \quad (10)$$

In (10) (x_i, y_i) , $(x_i^{'}, y_i^{'})$ is the corresponding node coordinates respectively and n is the number of nodes.

- center distance error:

$$d_s = \sqrt{\left(\frac{x_1^{'} + x_2^{'}}{2} - \frac{x_1 + x_2}{2}\right)^2 + \left(\frac{y_1^{'} + y_2^{'}}{2} - \frac{y_1 + y_2}{2}\right)^2} \quad (11)$$

In (11) d_s is the center distance error, (x_1, y_1)、(x_2, y_2) and $(x_1^{'}, y_1^{'})$、$(x_2^{'}, y_2^{'})$ is the corresponding node coordinates of linear feature.

c)Area feature

The positional conformity degree checking for area feature mainly use center distance, the fitness of area and precision indexes of the relative area.

- center distance error:

$$d_s = \sqrt{(x_s^{'} - x)^2 + (y_s^{'} - y)^2} \tag{12}$$

In (12) d_s is the corresponding polygon center point distance, $(x_s^{'}, y_s^{'})$ 、 (x_s, y_s) is the corresponding polygon center point coordinates

- the fitness of area:

$$\omega_s = \frac{A_i}{B_i} * 100\% \tag{13}$$

In (13), ω_s is the fitness of Polygon area, A_i is the corresponding polygon intersection area, B_i and is the evaluated polygon area.

- the accuracy of relative area:

$$Z_i = \left(1 - \frac{|A_i - B_i|}{B_i}\right) * 100\% \tag{14}$$

In (14), Z_i is the relative accuracy of polygon area , A_i 、 B_i is the corresponding Polygon area respectively.

B. The checking process

The positional conformity degree checking methods and processes are shown in Fig.1.

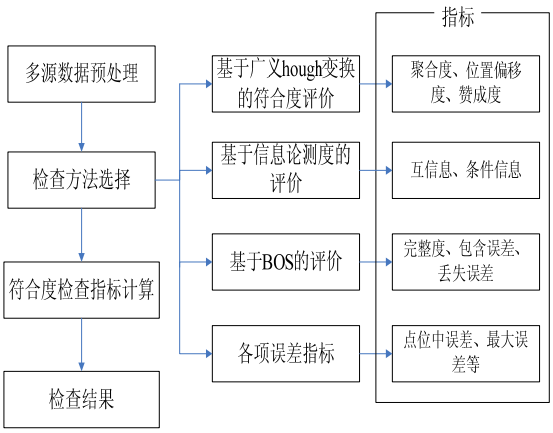


Figure1 Positional Conformity Degree Checking Methods and Process

1)Multi-source data preprocessing

Preprocessing is required before the positional conformity degree checking of multi-source data (different scale DLG, DOM, DEM, etc.) .Pre-work includes the conversion of coordinate systems and the conversion of different scales, making the data to be examined and referenced data have the same coordinate system and the same scale .

2)The choice of checking method

The positional conformity degree checking methods of multi-source data includes generalized Hough transform method, the method based on information theory, entropy-based error and uncertainty with the error indicator method, this paper mainly use the positional conformity degree checking methods based on all kinds of the error index.

3)The Calculation of the evaluation index

4)Giving checking report based on the checking index

Ⅱ The positional conformity degree checking of Western 1:50000 DLG

As Western data in the acquisition process have the characteristics of multi-source data with multiple coverage , When it checks the DLG product quality, flat precision, height accuracy should take the positional conformity degree checking methods . The flat location accuracy checking is completed by the flat location accuracy checking of DLG and DOM, the elevation position accuracy is completed by the elevation position accuracy checking of DLG and DOM.
In this paper the positional conformity degree checking of the two DLG products as an example to explain Western DLG positional accuracy checking method.

A.The plane positional conformity degree checking of DLG

The plane positional accuracy degree checking of DLG can achieve through the plane positional conformity degree checking of DLG and DOM. In this paper, the plane positional accuracy degree checking of DLG can take the conformity degree checking method which based on all kinds of error indexes, Specific targets take the bias of the corresponding point feature in the X and Y direction and the corresponding line elements nodes in the X and Y direction of the and apply point and the node point error.
The plane positional conformity degree checking of DLG and the corresponding DOM region process are shown in Fig.2:

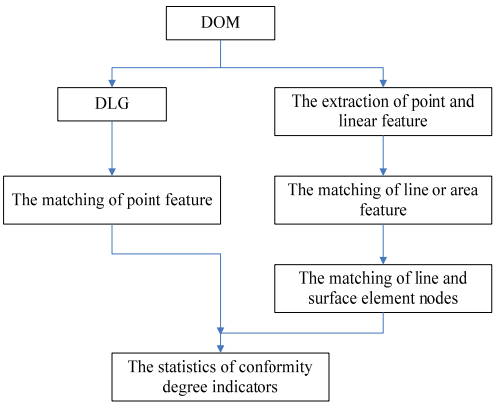


Figure2 Flow of Plane Conformity Inspection

1)the extraction of 1:50000 DOM point and linear feature
The extraction of point feature take manual selected extraction approach, the extraction of line features take human interaction way to complete.
The extractions of DOM linear elements take the following method:

- The endpoint of the given characteristic line: Picking up the DOM which describes the characteristic of the linear elements as tracking seed points.
- Characteristic line tracing: it begins from the point marked seed, searching in another seed point direction, spreading in the direction of its eight neighbors points, judging each new added pixel value, selecting the closest pixel point, the newly added point as the next seed point, it is the proliferation of new computing, and so on, until another tag with the initial seed point form an arc.

2)The matching of point, line or area feature [3].
The matching of the obvious features of point, line matching elements can be carried out in the two separate vector graphics which can find the corresponding pixels between the DLG and the extraction of elements. To find the corresponding pixel by buffer method can be completed. The matching point feature can be calculated by error index directly. The matching line, area make up by

a number of nodes of elements, so the matching of line and surface element nodes are required before the calculation of the error.

3)The matching of line and surface element nodes

The matching line and surface elements usually exists inconsistencies in the number of nodes; firstly a line should be standardized, so that the two pixels to be compared have the same number of nodes. Line vertical projection method can be standardized, projected point coordinates can be calculated according to analytical geometry theory, the calculation process is as follows:

From Fig. 3, it can calculate the coordinates of the projection point which based on analytic geometry theory.

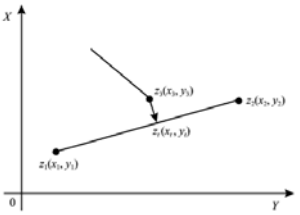


Figure3: Computation of Projection Points Coordinates

$$\begin{cases} x_t = x_2 + \dfrac{n}{m}\Delta x_{21} \\ y_t = y_2 + \dfrac{n}{m}\Delta y_{21} \end{cases} \tag{15}$$

In (15), $m = \Delta x_{21}^2 + \Delta y_{21}^2$, $n = \Delta x_{21}\Delta x_{32} + \Delta y_{21}\Delta y_{32}$

During the process of projection, if the generated nodes and original projection nodes are within the fuzzy tolerances, the nodes will be removed from the projector. After this treatment, the number of nodes on two corresponding lines will be the same.

4)The statistics of the matched error indicators

The node bias and error statistics of matched corresponding points elements and line elements are shown in TAB.I.

TABLE1 ERROR STATISTICS

Number	**Point code**	**Point type**	**X deviation**	**Y deviation**	**Product code**	**Point code**	**Point type**	**X deviation**	**Y deviation**
1	1	point	3.08	2.19	2	1	point	1.91	2.89
	2	point	2.4	2.2		2	node	2.12	2.45
	3	node	3.80	3.12		3	node	2.02	3.17
	4	node	2.01	2.49		4	node	1.80	2.34
	5	node	4.13	3.15		5	node	2.21	2.08
Error statistics	Mx	3.5631			Mx	2.2554			
	My	2.9788			My	2.9243			
	Mxy	4.6442			Mxy	3.6930			

In TAB I, The type of the midpoints respects point features in the table; the nodes respect the corresponding nodes of linear elements.

According to the regulations of "digital topographic map series and the basic requirements " in GB/T18315-2001 : the plains and hilly areas of 1:50000 scale topographic maps in the plane point map error can not be greater than 0.5mm, mountains and mountain locations in the error-bit map can not be larger than 0.75mm. AS the terrain of western region is complex, most belonging to the second topography, therefore the plane point error of DLG western can not be greater than 0.75mm.

According to error propagation law, the conformity degree checking with the relative positional error tolerance should be less than $\sqrt{2\times 0.75^2}$ = 1.0607mm, the field is $1.0607\times 5 = 5.3$ m.
According to the above provisions, plane points error of two DLG to be examined are within limits respect DLG positional accuracy meet the quality requirements[5][6].

B.The elevation positional conformity degree checking of DLG

The positional conformity degree checking of DLG can achieve through the conformity degree checking of DLG and DOM.
In this paper, the checking is carried out on the above the two DLG elevation positional accuracy ,the checking principle is: Firstly reading Elevation coordinates (X, Y) of the DLG product, using the DEM coordinates of the node (X, Y) and its four known grid points around the (X, Y, Z) value, using interpolation algorithm to calculate the elevation of the node elements; then comparing the calculated elevation with elevation values of elevation point reading from the DLG attribute table, calculating the elevation error of each node; finally knowing elevation error.
The formula of the elevation error is:

$$m_H = \sqrt{\frac{\sum (H_i^{'} - H_i)^2}{n-1}} \tag{16}$$

In (16), m_H is the elevation error of the detection point, $H_i^{'}$ and H_i are detection point elevation of DLG and DEM on the same name, n is the number of test points.
The elevation positional accuracy checking statistics of DLG in the two experiments is shown in TAB.II.

TAB.II. STATISTICS OF ELEVATION-CONFORMITY ERRORS

Product code	**The elevation positional conformity degree between DLG and DEM**
	M_H
1	12.1
2	14.3

According to the regulations of "digital topographic map series and the basic requirements " in GB/T17941.1-2000,the error of grid points should not be greater than contour lines elevation error which meets the corresponding topographic mapping requirements . The corresponding scale on the relevant provisions of the elevation error of "Digital topographic map series and the basic requirements "(GB/T18315-2001) is shown in TAB.III.

TAB.III REGULATION OF ELEVATION ERROR

scale	**Check**	**ground**	**hilly**	**mountain**	**High mountain**
1：50000	contour	3.0	5.0	8.0	14.0

AS the terrain of western region is complex, Elevation error tolerance adapt the mountains standard .According to error propagation law, the relative elevation error tolerance should be less than $\sqrt{2\times 14^2}$ =19.7990 m.
According to the above provisions, elevation points error of two DLG to be examined are within limits that respect DLG elevation positional accuracy meet the quality requirements[5][6].

Summary

In the paper, the novel concept and application of positional conformity degree were putted forward. In the test part of this paper, the necessity and feasibility for Western Mapping Project was analyzed, in order to solve the positional precision checking problem for DLG, the paper adopts the positional conformity degree between DLG and the DOM and DEM. Two DLG products of Western Mapping Project were selected for experiment, and the detailed procedure was depicted in the paper.

References

[1] Chrisman, N.R: *Obtaining Information on Quality of DigitalData*, Proeeedings of the AutoCarto London Conference, Jan 1986, p. 350-358.

[2] Volker walter: *Dieter Fritsch. Matching spatial data sets: a statistical approach*, INT. J. Geographical Information Science, vol. 13(5), 1999, p. 445-473.

[3] T. Mao and B. Zhang: *The assess method of spatial data quality based on Generalized Hough Transform*, University of Surveying and Mapping, vol. 21 (4), 2004, p. 266-268.

[4] GB/T18315-2001 (Digital topographic map series and basic requirement), The National Quality and Technical Supervision.

[5] GB/T18316-2001 (The acceptance and quality assessment of digital mapping product inspection) The National Quality and Technical Supervision.

[6] Walter V, Fritsch D. Matching spatial data sets: statistical approach [J]. Geographical Information Science, vol. 13(5), 1999, p. 445-450.

[7] Zhang Yongbin, Aimin Fan: *Spatial data quality assessment based on error entropy uncertainty band*, Journal of Heibei University, vol. 24(2), 2002, p.121-128.

Rapid Features Detection Using Improving Algorithm for Self-Localization in a DSP Board

Xing Xiong[a], and Byung-Jae Choi[b]

School of Electronic Eng., Daegu University, Jillyang

Gyeongsan-city Gyeongbuk, 712-714, KOREA

[a]GaleWing@gmail.com, [b]bjchoi@daegu.ac.kr

Keywords: SURF(Speeded-Up Robust Factures), DSP(Digital Signal Processor), Box filters

Abstract. With the shortcomings of large data amount and long time consuming in the conventional image feature points extract algorithms , an improved algorithm based on SURF for rapid features detection in a DSP board is presented in this paper. The algorithm only used the last octave to find a large number of the feature points.

Introduction

SURF [1] is the algorithm which could obtain the feature points faster than SIFT (Scale-invariant feature transform) [5]. However, in some low-speed processer such as DSP board, SURF's still a very large amount of computation and cost a lot of time, and feature points cannot be obtained in real-time.

Methods and Techniques

In paper [1], the SURF uses the box filter and integral image. Due to the use of both, the SURF can apply box filters of any size at exactly the same speed directly on the original image and even in parallel.

The SURF's scale analyze with constant image size.

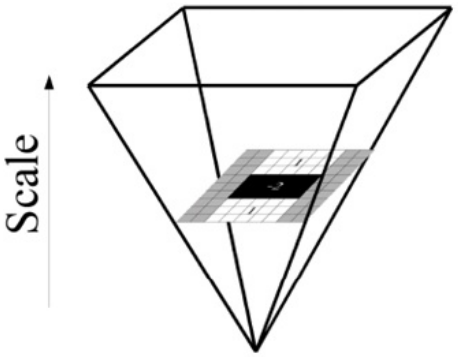


Fig. 1. The scale space is analysed by up-scaling the filter.

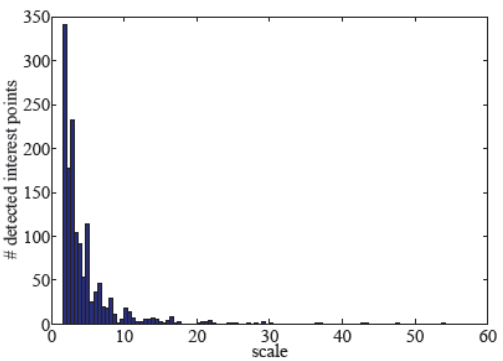


Fig. 2. Histogram of the detected scales.

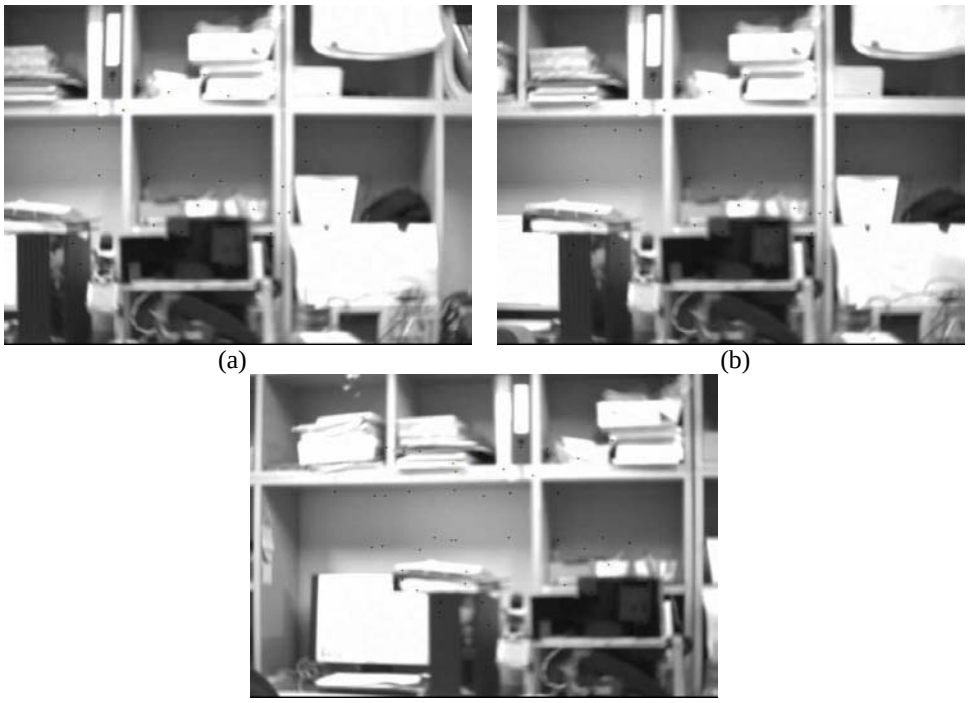
In the paper [1, 2], the more octaves may be analyzed, but the number of detected interest points per octave decays very quickly, the figure 2. In a large number of experiments, the gray value of the extracted feature points is very small, its means that brightness of the most feature points is dark.

However, because of fluorescent and material in the ceiling, the brightness of the most points in the ceiling is light.

In order to get the feature points whose gray value is high, the NMS (Non-Maximum Suppression) method after get Fast-Hessian matrix is changed to Non-minimum suppression. So the numbers of detected interest points per octave increase very quickly, just the opposite with the figure 2. In order to get a large of the feature points, the normal order of the scale are changed. Not only to remove the impact of the image edge, reducing the amount of calculation, but also can increase the interest points while the scale is large.

Simulation Result

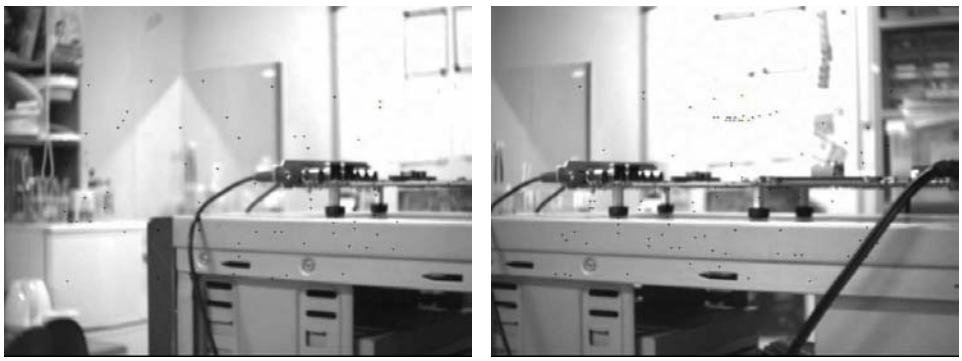
In order to verify the correctness of the algorithm, large number of experiments carried out. Here, we show some figure:



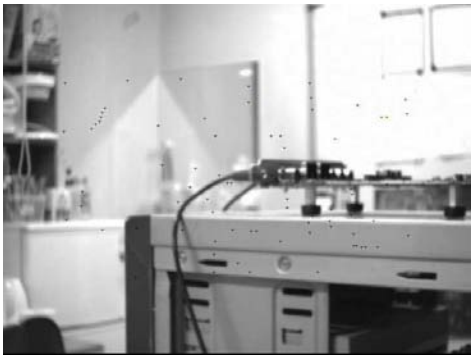
(a) (b)

(c)

Fig.3 Some simulation result after camera moved (where the black dots represent feature points)



(a) (b)



(c)

Fig.4 Some simulation result after camera moved (where the black dots represent feature points)

In above two sets of the pictures, each group of three pictures was obtained in short period of time and camera moved. From these pictures, when camera moves in a small range, its feature points remain especially in the brighter areas. This proves the method is useful.

Summary

In the paper, we proposed a improve method to detect interest points. Although the proposed method reduces the computational complexity and octave, the image after camera moved contained still a lot of the same feature points. In the future, we will extract descriptor of the feature points and match.

Acknowledgment

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology under Grant 2010-0006588.

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "*SURF: Speeded Up Robust Features*", Computer Vision and Image Understanding(CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

[2] Paul Viola and Michael Jones, "*Rapid Object Detection using a Boosted Cascade of Simple Features*", Accepted Conference On Computer Vision And Pattern Recognition 2001.

[3] HAN Bing and B. Boyd, "*Direct Replacement Algorithms of Fast Computing Integral Image in SURF*", Journal of Projectiles, Rockets, Missiles and Guidance, Vol. 31, No. 3, Jun 2011.

[4] WANG Jun-ben, LU Xuan-min and HE Zhao, "*An Improved Algorithm of Image Registration Based on Fast Robust Features*", Computer Engineering & Science, Vo1.33, No.2, 2O11

[5] David G. Lowe, "*Distinctive Image Features from Scale-Invariant Keypoints*", International Journal of Computer Vision, 2004

The Development of Corridor Identification Algorithm Using Omni-directional Vision Sensor

ARTHAYA Bagus[1,a], and WU Mellisa[2,b]

[1]Parahyangan Catholic University, Bandung 40141, INDONESIA

[2]Alumni of Parahyangan Catholic University, Bandung 40141, INDONESIA

[a]bagusmooi@gmail.com, [b]mellisa.wu.89@gmail.com

Keywords: corridor identification, omnidirectional vision sensor, image processing, mobile robots.

Abstract. Robot is one of applications in automation that has been implemented and developed in the last decade. The use of robots can be found in many manufacturing processes, and other activities like assembly, transportation, medical, and research. Based on its mobility, robot can be divided into two types: static robot (fixed at one place) and the mobile robot (can move from one point to another). One example of mobile robots is AMR (Autonomous Mobile Robot) which has a navigation system in guiding the robot to move towards the desired places. When an AMR placed on a corridor, it requires a system of sensors to provide feedback about the surrounding environment so that AMR can perform desired movements when moving in the corridor. In this study, the sensor used is omni-directional vision sensor. This sensor is capable of capturing images that provide information of its 360º surroundings in one image. This image requires further identification process in order to get the information contained in the image and use it as feedback for AMR. This identification process can be done by designing an algorithm to translate and produce decisions based on images captured by the sensor. The first step of this algorithm is to divide the 360º environment into four quadrants and then further information from each quadrant is used to determine the type of corridor. There are 4 types of cases that may occur in a quadrant as follows, straight road, dead ends, turn point, and T-junction. Based on the cases in quadrant, some conditions that may be found in the hallway of this research are blind alley, intersection, right turn, left turn, and other conditions. The designed algorithm is implemented in several images taken using the omni-directional vision sensors. Decision on conditions of corridor stated by the result of the algorithm is in accordance with the conditions of the corridor in the real situation. Weaknesses of the algorithm is the existence of specific constraints such as environmental contrast, the maximum distance of detection, etc

Introduction

Mobile robot is now getting wider attention from researchers all over the world especially for explorations purpose such as mapping or scanning new terrains, poisonous areas, unstructured environments, radio-active contaminated places, and so on. One of popular examples of mobile robot is Autonomous Mobile Robot (AMR) which is useful for exploring an unknown area such as desert, forest, new setup building, etc. before a physical access is performed. Meanwhile, fixed position robots are typically used in production lines, such as in the assembly line, welding line, painting booth, and so on. An AMR can move around its surrounding autonomously to do some activities without any intervention from human beings.

Nayar [3] explores the usefulness of omni-directional camera (or a new type of camera with a hemispherical field of view). He was able to present a pure perspective view from an omni-directional image given any user-selected viewing direction and magnification.

On the other hand, Gaspar *et al.* [4, 5] addressed the problems of mobile robot perception in the context of navigation. They utilized panoramic view to develop a 3-D model of scanned corridor by a mobile robot equipped with hemispherical reflector and digital camera. They also continuously tested the performance of the topological navigation system.

In their work, Kawanishi *et al*. [6] explore the advantages of omni-directional vision sensor for determining the mobile robot velocity by measuring camera motion. They tag some corresponding points in the robot surrounding and compare the images taken before and after the robot movement. This method works effectively and was able to model a passage way and a room of an experimental environment.

In order to map a new area within a building, Goh and Lee [7] introduced an adaptive omni-directional vision system by inserting a wide-angle lens into the existing omni-directional vision system. Significant improvements were shown especially in mapping a closed indoor space.

The first AMR prototype that's capable of mapping an unknown maze [2] was developed in our laboratory. It is a very simple mobile robot that consists of two stepper motor driving wheels and the control of robot motion is still done by a desktop PC. The data transfer between the robot and the control computer was carried out by hard-wire communication. However, this mobile robot has been able to map an unknown map successfully.

The later development of this robot is a similar type but it is equipped with ultrasonic sensor and blue tooth communication features [1]. The main advantage of this mobile robot is that it can autonomously map un unknown maze and continuously work although there is communication interrupt or break down and it also was able to implement recursive back tracking algorithm with priority.

This research is addressed in taking the benefits of omni-directional image used for identifying corridor status when an AMR has to map an unknown maze. The basic idea is based on researched done by Gaspar *et al.* [4, 5] but its goal is getting the most possible status of one particular position/labyrinth in a corridor. Information about the corridor status can later on be used as feed forward information by the robot controller to decide which way to be taken in performing its task.

Maze Mapping Task

One method that can be used to map an unknown maze is the recursive back tracking algorithm. This algorithm ensures the robot to move/to visit any labyrinth inside the map. In exploring an unknown maze, this algorithm is employed and from the starting node (a cell in the left bottom corner of the maze) the robot moves cell per cell according to the priority assigned to it. Example of a tree of nodes is presented in Fig. 1. When the right side tree is explored first, it will move from Start to B-G-J-O, whereas O is a dead end. When back track is applied, it brings the robot back to J and then it moves forwards to N. Again N is a dead end and the same cycle starts over again resulting path as follows: N-J-G-B-F-B-Start and so on.

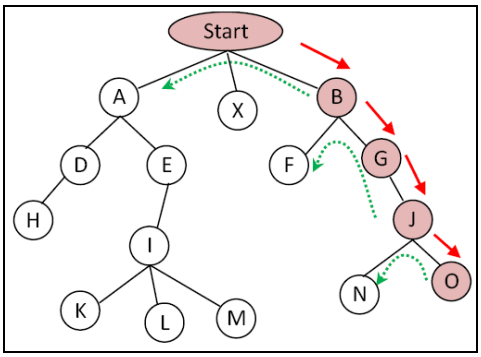


Figure 1. Example of cells network that should be explored

When approaching a crossing point, the robot does not have enough information about this point, unless it arrives at that point. To go to one node, the robot should make a decision to turn left or right or go straight forward. This decision can only be made when the robot has arrived exactly at the cross point. In some important tasks of exploration, this action is considered to be late. Anyhow this method was able to completely map an unknown maze and to develop 2-D digital map as depicted in Fig. 2. To deal with this situation, an image processing technique that captures the corridor condition is introduced before the robot arrives at one crossing point. The sooner the decision can be made, the better the explorations task will be executed.

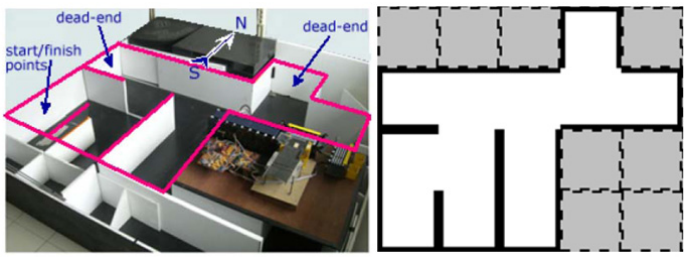


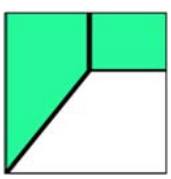
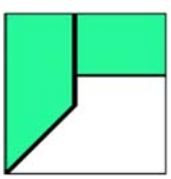
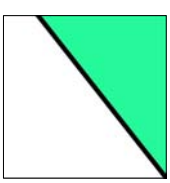
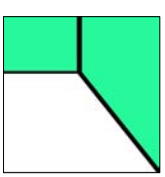
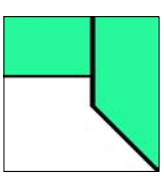
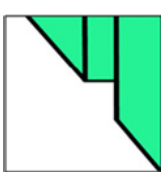
Figure 2. The maze setup containing 5x4 cells and its digital map

Algorithm Development of Corridor Identification

Omni-directional vision system provides the robot with complete information of its 360º surroundings in one image/view. This view is captured by a camera with a hemispherical field of view and cannot be directly used to control the motion of the robot. This image needs some processing actions as will be explained later.

The research starts with the development of omni-directional vision sensor system as an additional "seeing" device for the AMR and also as a modeling feature of the corridor. The image processing algorithm consists of main two parts, i.e. (1) algorithm for processing image data and (2) algorithm for identifying corridor condition. The image processing delivers digital pictures of front view as well the rear view of the robot. In this research, only front view will be further processed. This picture is processed by the second algorithm and this algorithm reports any possible condition of the corridor. Table 1 exhibits all possible views captured by the image processing system. Introducing this information to the robot controller in advanced will be much helpful before it should make a decision.

Table 1. All Possible Corridor Feature that Might be Faced by an AMR

Possible Case	Front views			
	(a) Straight way	(b) End corner	(c) End turning	(d) Crossing point
1. Left side Cases				
2. Right side Cases				

The pictures in Table 1 are all possible conditions in a corridor that might be met by the robot. These conditions may appear within a corridor containing some crossing points, turning points and dead-end as depicted in Fig. 3. When an AMR entering this kind of corridor lies inside an unknown maze, the system should analyze one or more cases and deliver the correct decision.

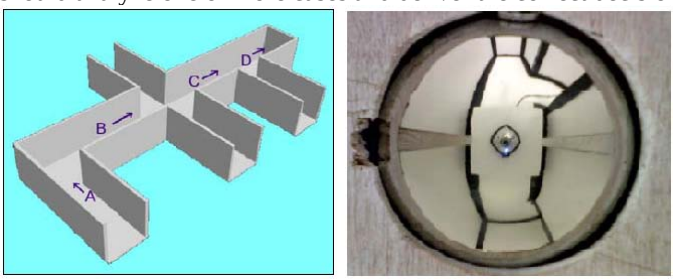


Figure 3. The Example of a corridor containing all possible crossing points, turning ways and dead end as well as the omni-directional capture

When the robot is moving from point A, image processing system will deal with picture in Table 1 (1,b) and (2,c) and the conclusion will be turning right within a few cm forward. While when it's in point B, picture in Table 1 (1,d) and (2,d) will be seen and decision has to be made whether to turn the left or to the right.

Image Conversion and Cropping. In this research, the picture capturing is still done manually as this camera system has not been assembled with the mobile robot yet. Every picture taken is then converted into the form of panoramic view as depicted in Fig. 4. It is clearly seen that some parts in the picture are not really needed for the analysis and then a cropping action is performed. Cropping some parts of the image provides the only important digital image that really needed in the calculation and reduces the calculation time. In the left part of the picture, a spherical view of a corridor is captured and the conversion in a panoramic view is shown in the right part. Panoramic view contains a lot of unwanted information and cropping action needs to be done.

Figure 4. Omni-directional picture and its panoramic view

In Fig. 5, an omni-directional view is converted into cropped panoramic view with identification of all four quadrants from the original view. From this image, we need only information of lines that construct the type of corridor features as depicted in Table 1.

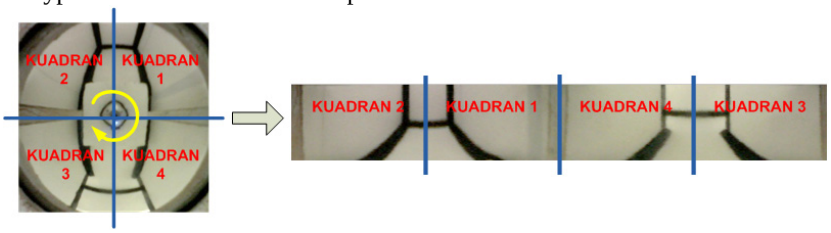


Figure 5. Conversion of omni-directional picture into a cropped panoramic view

Corridor Identification Technique. From every quadrant of panoramic image, one can conclude the existence of corridor features as tabulated in Table 1. Using standard calculation procedures in image processing technique, edge lines that construct a corridor can be identified as shown in Fig. 6.

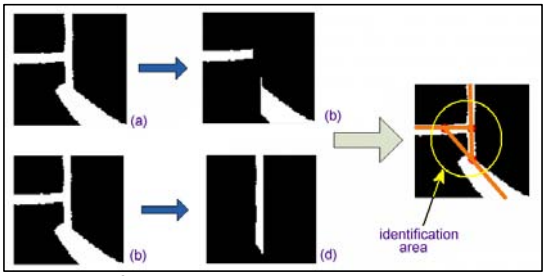


Figure 6. The step of conversion from quadrant image into possible type of constructed feature

The first step is identification of vertical line in the image. The existence of vertical line and the number it appears will determine the features may available (e.g.: all features in Table 1 except features (1,a) and (2,a)). More than 1 vertical line will present features (1,d) and (2,d). Other than those two cases, then features (1,b and 1,c) and (2,b and 2,c) will appear.

The second step is distinguishing features (b) and (c) is done by comparing the size of identification area. At a certain distance of picture capturing, each of these features will show a certain size. By giving the threshold value, features (b) and (c) then can be differentiated.

Table 2: The caption of a table should appear at the top of the table

Case	Left- feature	Right- feature	Corridor condition
1	(1,a)	(2,a)	Straight forward
2	(1,a)	(2,d)	Cross point 2 directions, straight forward and turn right
3	(1,b)	(2,b)	Dead-end
4	(1,b)	(2,c)	Turn right soon
5	(1,c)	(2,b)	Turn left soon
6	(1,d)	(2,a)	Cross point 2 directions, straight forward and turn left
7	(1,c)	(2,c)	Cross point 2 directions, turn right and turn left
8	(1,d)	(2,d)	Cross point 3 directions, straight forward, turn right and turn left

This simple algorithm is then tested in simple software equipped with interactive GUI (Graphical User Interface). The opening of the software will show an interface as depicted in Fig. 7. User can input any image taken from omni-directional sensor directory, and they will be delivered the most possible corridor feature as depicted in this figure as well.

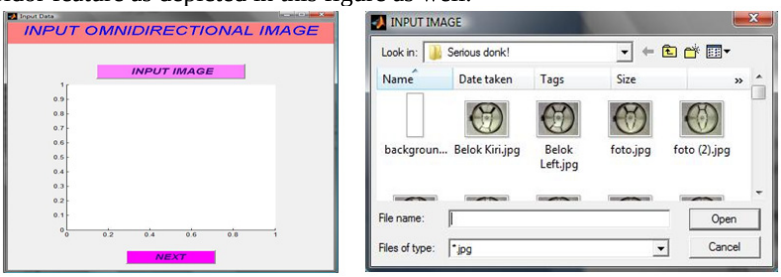


Figure 7. The GUI when user starts the software and can chooses one image

An image chosen from a directory is shown in Fig. 8. This image is then processed according to the

technique presented above. The result shows that at the front side view there are features (1,b) and (2,b) and at the back side there are features (2,c) and (3,c).

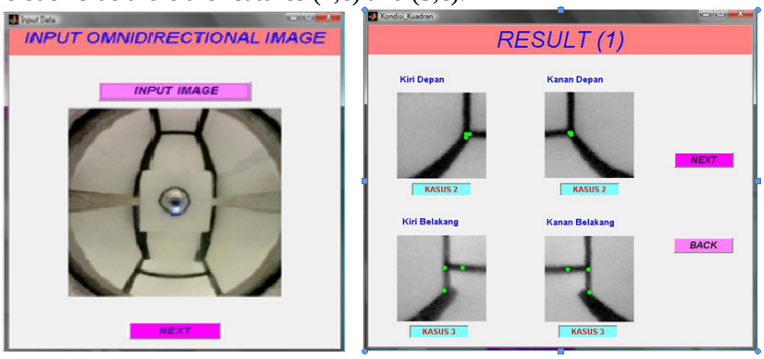


Figure 8. Omni-directional picture chosen from the file directory

Conclusions

We have demonstrated the possibility of omni-direction vision sensor contribution in mobile robot maze mapping especially in indentifying what corridor feature is coming. Instead of waiting until the robot arrives in a cross-point to identify the corridor feature, this technique can provide identification far before it arrives in the position and it is very useful when the robot should make some advancing decisions.

There are 8 possible corridor features that might appear when omni-directional image is processed. The problems faced in identifying the feature are the quality of image captured, the focus of the image and the image distortion as we used a spherical form of mirror. Some improvements still have to be done when this technique is going to be implemented on the real mobile robot

References

[1] B. Arthaya *et al*. *On-line maze mapping by AMR equipped with ultrasonic sensor*, Proc. of the 4th Int. Sem. on Industrial Technology and Management, Lombok, Indonesia, 2010, p. 332-337.

[2] B. Arthaya *et al*. *Design of AMR Prototype and Motion Algorithm for Mapping an Unknown Maze*. Proceedings of the 10th Int. Conf. on Mechatronics Technology, Mexico City, 2006.

[3] S. K. Nayar, Catadioptric Omnidirectioanl Camera, *Proceeding CVPR '97 Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 1997, p. 482..

[4] J. Gaspar *et al.*, Vision-based navigation and environmental representations with an omni-directional camera. *IEEE Trans. on Robotics and Automation*, Vol. 16, No. 1, 2000, p. 890-898.

[5] J. Gaspar *et al.*, Toward Robot Perception through Omnidirectional Vision, *Preprint. Final version in Chapter 9 of "Innovations in Intelligent Machines - 1", Springer*, 2007, *http://www.springer.com/978-3-540-72695-1*.

[6] R. C. Gonzalez, , R. E. Woods, , and S. L. Eddins, *Digital Image Processing Using MATLAB*, Prentice Hall, Upper Saddle River, New Jersey, 2004.

[7] M. Goh. and S. Lee., Indoor Robot localization using Adaptive Omnidirectional Vision System, *IJCSNS International Journal of Computer Science and Network Security*, VOL.10 No.4, April 2010, p. 66 – 70.

[8] R. Kawanishi, A. Yamashita, and T. Kaneko, Estimation of Camera Motion with Feature Flow Model for 3D Environment Modeling by Using Omni-Directional Camera. *IROS'09: Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems, IEEE Press.*

Coordination of Ambulance Team Agents in Rescue Simulation Using Auction Strategy

Pooya Deldar Gohardani[1a], Peyman Ardestani[1b],Behrooz Masoumi[c]Mohammad RezaMeybodi[2d],Siavash Mehrabi[1e]

[1] Mechatronic Research Labratory, Islamic Azad University, Qazvin Branch, Qazvin, Iran

[2] Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, Iran

{[a]Pooya.Deldar,[b]Peyman.Ardestani, [c]Masoumi }@Qiau.ac.ir ,

[d]mmeybodi@aut.ac.ir, [e]Siavash.Mehrabi@Gmail.com

Keywords: Multi-agent, Robocop Rescue Simulation, Coordination, Auctioning.

Abstract. RoboCup Rescue Simulation System is a large-scale, real-time and multi-agent simulation of urban disaster. The problem introduced by RoboCup Rescue brings up several research challenges that go from Intelligent Robotics to Multi-Agent Systems (MAS) research. The goal is to coordinate and control the emergency services in the city to minimize damage and injuries resulting from the disasters. In this paper a new method based on market-based method auction strategy for coordination of the ambulance team agents is proposed which can simultaneously uses the power of centralized and decentralized approaches for better coordinating of worker agents. The results of simulations show that the proposed method has better performance in comparison with the other methods coordination systems that were used in RoboCup 2010 International competitions.

Introduction

A multi-agent system (MAS) is comprised of a collection of autonomous and intelligent agents that interact with each other in an environment to optimize a performance measure [1]. RoboCup Rescue Simulation System is a large-scale, real-time and multi-agent simulation of urban disaster [2]. For robotics and multi-agent researchers, RoboCup Rescue works as a standard platform that enables easy comparison of research results. The problem introduced by RoboCup Rescue brings up several research challenges that go from Intelligent Robotics to Multi-Agent Systems (MAS) research. These research challenges include real-time flexible planning, multi-agent coordination and team formation, path planning and navigation, heterogeneous resource allocation and machine learning at the team level. In fact, the main goal in this domain is minimizing the damage by helping trapped agents, extinguishing fiery collapsed buildings, and rescuing damaged civilians. There are several studies about cooperating and coordinating agents, communication, negotiation, distributed problem solving. Coordination of the agents is one of the highly rated subjects in this field which falls into three main categories: centralized, decentralized and a combination of previous methods.

Several coordination strategies are used in Robocop Rescue Simulation challenges. For example, ResQ Freiburg [3] implements a centralized mechanism for the Fire Brigade agents which send agents current and next targets by means of a leader. Caspian [4] combines the centralized and decentralized coordination approaches and regards the partitioning strategy as a social law, applying it to all the rescue agents. S.O.S [5] divides the FireBrigade agents into several teams, each team works together to accomplish the pre-assigned tasks. SBCe_saviour2004 [6] uses centralized coordination strategy and sets up a virtual center that coordinates all the decisions made by central agents. Impossibles [7] used an auction system for decentralized task allocation amongst their fire

brigade agents. Chou and Marsh [8] used a decentralized multi-agent coordination architecture based on contract net protocol.

In this paper a new method based on market-based method (auction strategy [9]) for coordination of the ambulance team agents is proposed which benefits from both centralized and decentralized approaches for better coordination of worker agents. The content of this article is as followed: After introduction, a brief description to RoboCup Rescue Simulator will be presented, then, Market based task allocation is explained, after that proposed algorithms are introduced, then, we will evaluate the proposed algorithms in RoboCup Rescue Simulator and compare them with some other participants of RoboCup 2010 competitions (as benchmark). Finally, The Conclusion is presented.

RoboCup Rescue Simulator

RoboCup Rescue simulator is a multi-agent system, first introduced in [10]. The goal of the system is to study the progress of rescue operations in a simulated part of a city (disaster space). The same as real life there are destructed buildings, burning buildings, blocked roads and injured civilians in the simulated disaster space.

There are three types of rescue agents with specific capabilities in rescue simulation system: ambulance agents are able to recover buried civilians, and transfer them to refuges where they can be tended to; fire fighter agents are able to extinguish fires, and police agents are able to clear blocked roads. An example of such simulated environment is illustrated in Figure 1.

The simulated environment is highly dynamic and Agents have to plan and decide their actions asynchronously in real-time, so single agents can make less remarkable effect on rescue processes without coordination and cooperation of other participant agents, therefor there is a real emphasis on coordinating agents to work better. There exists an evolution criterion in the simulation environment which names as score and obtained from Eq.1 [11].

$$score = (P + \frac{H}{H_{init}}) \times \sqrt{\frac{B}{B_{Max}}} \quad (1)$$

Where P is the number of civilians alive, H is the amount of health point of all agents and the ratio to the number of civilians alive initially, H/Hinit, shows the efficiency of operations, B is the area of houses that are not burnt, and BMax is the area of all houses.

As it is shown from Eq.1 the civilian health is very high on the final score, thus rescuing civilians as fast as possible with attention to their injury degree is a critical problem of this domain.

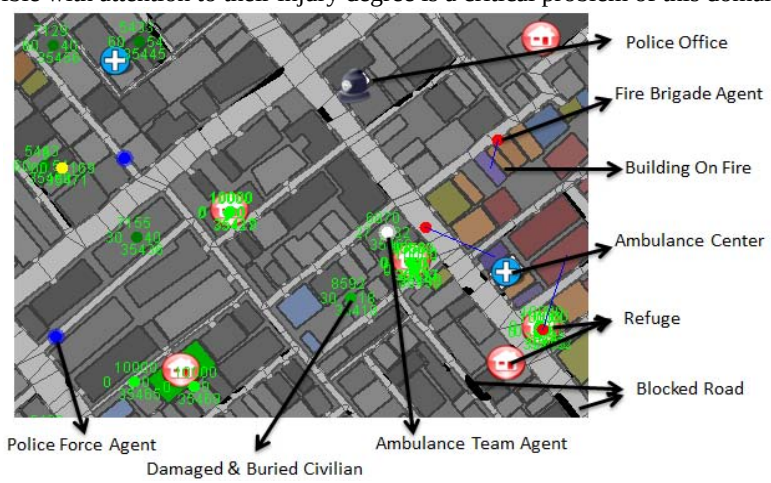


Fig.1. A Simulated disaster environment in Rescue Simulation system

Market-Based Multi-agent Coordination

There is a variety of coordination methods which falls into three main categories: Centralized methods, Decentralized and Market-based methods. Market-based methods are a combination of centralized and decentralized methods which benefits from both centralized and decentralized methods advantages [12].

As mentioned above, the principles of a market economy can be applied to multi-agent coordination. In this virtual economy, the agents are traders, tasks and resources are traded commodities, and virtual money acts as currency. Agents compete to win tasks and resources by participating in auctions that produce efficient distributions based on specified preferences. When the system is appropriately designed, each agent acts to maximize its individual profit and simultaneously improves the efficiency of the team.

Abstractly, an auction takes place between an agent known as the auctioneer and a collection of agents known as the bidders and the goal of the auction is for the auctioneer to allocate the good to one of the bidders [13].

There exist several protocols for auction settings which have differences in some dimensions as winner determination, bid awareness, and bidding mechanism [14]. As an example there could be an auction setting which uses first-price or second-price auctions for winner determination, open cry or sealed-bid methods as to different methods of knowledge awareness and one shot or ascending auctions as its bidding mechanism.

The Proposed Method for Ambulance Team Agents

In this section, we introduce the proposed method based on the concept of marketing to enhance the ambulance team agents' coordination. In this method, an auction takes place between the ambulance team agents. The goal of the auction is for the auctioneer to allocate an optimum number of ambulance team agents to an injured civilian so that the number of living civilians is maximized and minimizing the number of perished civilians at the end of the simulation. For implementing the auction model, it is assumed that two types of agents are defined: Leader agents and Worker agents. Leaders are auctioneers and workers are bidders that perform the tasks which are assigned to them. It should be mentions that a leader could act as a worker as well.

Considering the definition of a market-based method and as will introduced in next sections; in proposed method, first-price method is used for winner determination, one shot method is used for bidding mechanism and open cry method is used for their bidding awareness. The proposed algorithm, called Algorithm 1, and its structure and its pseudo code are given in more detail in Figure 2 and Figure 3 respectively.

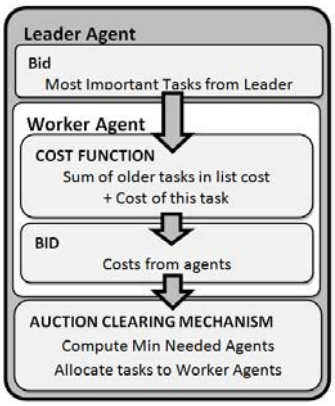


Fig. 2. Structure of the proposed algorithm

Auction Setup. In proposed method there is a constant, pre-defined schedule for beginning auctions that are called Auction Time.In each auction time there is only one agent (Leader) as the auctioneer. The Leader selects m most important (Level of importance is measured by a variable named Time to Death (TTD) which defines the remaining time cycle for a specific civilian to die, from the perspective of system or a mixture of some other parameters (Eq.1 and Eq.2)) injured civilians for auction and runs the auction. In another word Leader sends a list of m civilians and informs other agents to send their bid for an incoming auction. Workers take part in started auctions as bidders and submit their bids for every m civilian in auction considering the situation and status of themselves and status of target civilian.

Market with Leader Initiator Auction

Leader Agent

```
While(simulation is not ended){
if it is Auction Time {
Find m most valuable victims;
Put selected victims to sell as leader bids;
{
if it is Time to get worker bids {
Compute needed agents for each bade victims;
Find winner/winners of each bade victim;
Broadcast winners of all bade victims;
{
Do assigned tasks based on task list;
}
```

Worker Agent

```
While(simulation is not ended){
if it Time to get leader bid/bids {
compute cost /utility for each leader bid;
send bid for each leader bid;
{
If it is Time to get winner {
if I am winner{
add won victim to task list;
}
{
Do assigned tasks based on task list;
}
```

Fig.3. Pseudo code of the proposed algorithm

$$\text{Importance} = F(\text{TTD}, \text{BRD}, \text{ATA}) \tag{2}$$

$$F(\text{TTD}, \text{BRD}, \text{ATA}) = w_1 \times TTD + w_2 \times BRD + w_3 \times ATA \tag{3}$$

Here BRD and ATA respectively stand for Buried property (the amount of collapse on a victim) and Ambulance Team Around (number of ambulance teams nearby the victim).

Task Management byAgents. Considering that leader put m injured civilians in each auction, there is a probability that workers win in another auction before finishing their previous task. This could happen for many times and agents need to stack up their tasks in a list. This helps to calculate bids more precisely as well.

Bidding Mechanism. Each worker agent could announce its preparation for bidding after receiving the auction schedule message considering its current status and circumstances of the simulated environment.

This bid could be as a cost of service or a profit from service. In this article, cost of service is considered for bidding in a way that with m task in agent's task list next bid will be calculated by Eq.4.

$$\text{CostFunction} = \sum_{i=1}^{m} t_{rescue}\,(Civilian_i) + \sum_{i=1}^{m} t_{move}(i\,, i+1) \tag{4}$$

In this equation t_{rescue} is the time needed for rescuing i_{th} injured civilian and t_{move} is move cost-function from an injured civilian to next injured civilian.

Task Assignment (Winner Determination). If we use the *cost of service* method (as considered in this article), after calculating the number of needed agents, the leader will announce the agents, with minimum cost-function to rescue the civilian, as winners of auction and assigns the target to them.In another hand if we use the *profit from service* method, leader will assign the task to agents with higher bids. It should be noted that phrase "Compute needed agents" in Fig. 1 that represents the number of needed ambulance team agents needed for rescuing an injured civilian which is the minimum number of agents that is needed to rescue that civilian and bring him/her to refuge alive.

Simulations and Results. The proposed method is implemented for ambulance team agents of MRL team. MRL is one of the most successful teams that take part in International RoboCup Competitions and other competitions like IranOpen. In order to measure the functionality of this method, we have compared the test results with test results of roboAKUT [15], RiOne [16] which took place in 1st, 4th in RoboCup 2010 respectively [17]. Studied parameters in this comparison are shown in table 1.

Table 1. Comparison parameters

Parameter Name	Definition
Score	Final score of simulation
Alive Civilians	Number of alive victims at the end of simulation
Dead Civilians	Number of victims which die before simulation ended
Sum Of HPs	Sum of Health Points of victims at the end of simulation
Sum Of DMGs	Sum of damages of victims at the end of simulation

Score is a standard parameter that represents the total functionality of agents in n time cycles of simulation, higher value in score is better. The higher Alive Victims and Sum of HPs show the better performance. In contrast, less Dead Victims and Sum of DMGs are better. In order to obtain comparison parameters we ran simulation on Kobe, VC, Berlin and Paris maps with listed parameters in Table 2. As it is obvious by Table 2, Berlin and Paris maps are so bigger than two other ones and have more injured civilians, so their conditions are more critical.

To get precise results, each method executed 5 times and the average of those results listed in Table 3-7.

Table 2. Map parameters in our experiments

Parameter / Map	Civilians	Ambulance Teams	Refuges	Dimensions
Kobe	169	10	5	450×350
VC	162	10	5	430×440
Berlin	199	10	4	2200×1650
Paris	196	10	5	1000×1000

Table 3.comparison of different methods in different maps in terms of Score parameter.

Score	Method / Map	RiOne	roboAKUT	LIA_1	LIA_2	LIA_3	LIA_4	LIA_5
	Kobe	63.225	67.288	69.567	77.441	72.446	74.653	75.658
	VC	40.146	41.95	41.689	53.312	49.718	49.919	50.72
	Berlin	44.093	45.339	35.686	49.115	54.125	53.938	48.533
	Paris	41.475	44.124	36.67	43.081	54.096	48.898	46.098

Table 4. Alive Civilians comparison of different methods in different maps

Alive Civilians	Method / Map	RiOne	roboAKUT	LIA_1	LIA_2	LIA_3	LIA_4	LIA_5
	Kobe	63	67	69.333	77.2	72.2	74.4	75.4
	VC	40	41.8	41.6	53.2	49.6	49.8	50.6
	Berlin	44	45.2	35.6	49	54	53.8	48.4
	Paris	41.4	44	36.6	43	54	48.8	46

Table 5. Dead Civilians comparison of different methods in different maps

Dead Civilians	Method / Map	RiOne	roboAKUT	LIA_1	LIA_2	LIA_3	LIA_4	LIA_5
	Kobe	106	102	99.667	91.8	96.8	94.6	93.6
	VC	122	120.2	120.4	108.8	112.4	112.2	111.4
	Berlin	155	153.8	163.4	150	145	145.2	150.6
	Paris	154.6	152	159.4	153	142	147.2	150

Table 6. Sum Of HPs comparison of different methods in different maps (values are changed to (value - 100000) /5000)

Sum Of wHPs	Method / Map	RiOne	roboAKUT	LIA_1	LIA_2	LIA_3	LIA_4	LIA_5
	Kobe	56.099	77.269	58.857	61.33	63.302	65.455	67.123
	VC	27.217	28.641	8.865	16.345	18.259	18.467	18.817
	Berlin	16.991	35.362	14.339	25.898	29.933	35.027	32.932
	Paris	9.524	28.694	7.408	11.851	17.826	18.283	18.578

Table 7. Sum Of DMGs comparison of different methods in different maps (values are changed to (value-140000)/200)

Sum Of DMGs	Method / Map	RiOne	roboAKUT	LIA_1	LIA_2	LIA_3	LIA_4	LIA_5
	Kobe	58.435	59.635	56.846	52.167	55.021	54.801	53.635
	VC	73.584	71.572	70.25	64.471	65.377	66.029	65.246
	Berlin	91.925	93.851	98.705	89.98	85.096	86.02	90.4
	Paris	92.365	91.92	95.522	90.52	85.274	86.631	88.833

As it is shown from Table 3-7, there are different methods in LIA that the difference is in the number of injured civilians that leader puts in auction (LIA_1 one civilian, LIA_2 two civilians …). Due to the results it is obvious that LIA methods have better results comparing to other methods but according to the map conditions, if the map is not huge like Kobe and VC, then the LIA_2 method is the best one and if the map is huge then the LIA_3 has best overall result.

Considering Table 6, it seems that roboAKUT acts better than all other methods to save more HPs, but making a glimpse at Table 7 it will be shown that it got more DMGs too, which resulted in getting less overall Score.

Alongside these positive results this method has some shortcomings that need solutions. Some of these shortcomings are as follows:

- This system needs a high bandwidth for setting the auction, which is problematic in noisy and limited communication conditions.
- Delay in assignment. Because there are only a few targets for sale at auction there is a possibility that system fails to assign agents too all targets in auction time therefore some injured civilians remain unattended and may perish before ambulance teams could rescue them.
- In case of a big number of injured civilians (m) agents have to stack up their assigned tasks; this makes the calculation of cost function more complex and less precise in every step. As mentioned these data are used as bids in auctions which directly affect the precision of leader decision making system. This works as a destructive cycle and has negative effects on overall results.

As you could see in fig. 4 according to part 2 and 3 of this section LIA_3 method has better result by holding more auctions but it will demand more bandwidth due to a huge amount of messages needed for every auction

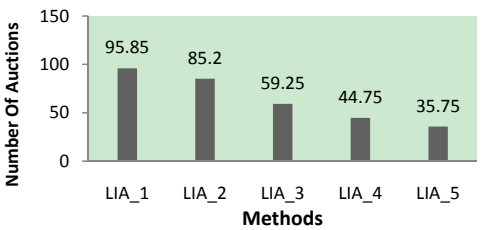


Fig.4. Comparison of Auction times in different proposed methods

There is a lot of load on the leader to hold auction, task assignment and communication with all workers; this makes the leader a bottleneck for this system due to a limited process time for each agent and may cause action timeout for leader which means the simulation kernel ignores the leader action commands and leader misses a cycle.

Conclusion

This paper presented an auction-based method for coordination of ambulance team agents in Robocop Rescue Simulation. A brief overview of market-based approaches had been presented and Robocop rescue simulation described as a great challenge for multi-agent coordination mechanisms. After presenting a detailed discussion on proposed method, comparisons between higher statuses in Robocop2010 and different methods of auctioning had been introduced. It had been shown that LIA_2 with m=2 injured civilians for each auction had best overall results in normal maps and LIA_3 with m=3 injured civilians had best overall results in huge maps.

In an overview the proposed method has the best results comparing other methods of other multi-agent coordination systems that were used in RoboCup 2010 International competitions but of course this method has its own shortcomings such as: requirement for high bandwidth communication, delay in assignment and need for a leader as a coordinator of other agents. In future works it is possible to address these problems with changing the marketing system.

References

[1] Weiss, G.: Multi-agent Systems A Modern Approach to Distributed Modern Approach to Artificial Intelligence. MIT Press, Cambridge, Massachusetts, London (1999)

[2] Kitano, H.: RoboCup Rescue: a grand challenge for multi-agent systems. Fourth International Conference on Multi-Agent Systems, IEEE Press, Boston, MA, USA (2000), p. 5-12

[3] Kleiner, A., Brenner, M.: ResQ Freiburg: Team Description and Evaluation. Proceedings of Robocup2004: The 8th RoboCup International Symposium. Lisbon, Portugal (2004)

[4] Sedaghati, M., Gholami, N., Rafiee, E., et al.: Caspian 2004 Rescue Simulation Team Description. Proceedings of Robocup 2004: The 8th RoboCup International Symposium. Lisbon, Portugal (2004)

[5] Amraii, S., Behsaz, B., Izadi, M., et al.: S.O.S 2004: An Attempt towards a Multi-Agent Rescue Team. Proceedings of Robocup 2004: The 8th RoboCup International Symposium. Lisbon, Portugal (2004)

[6] Nazemi, E., Fardad, M., Radmand, A., et al.: Message Management System in SBCe_Saviour Team. Proceedings of Robocup 2004: The 8th RoboCup International Symposium. Lisbon, Portugal (2004)

[7] Habibi, J., Fathi, A., Hassanpour, S., Ghodsi, M., et al.: Impossibles Rescue Simulation Team Description Paper RoboCup. Osaka, Japan (2005)

[8] Chou, W., Marsh, L., Gossink, D.: Multi-Agent Coordination and Optimization in the RoboCup Rescue Project. 18th World IMACS / MODSIM Congress. Cairns, Australia (2009), p. 1608-1614

[9] Milgrom, P. R., Weber, R. J.: A Theory of Auctions and Competitive Bidding. Econometrica, vol. 50, no. 5, Econometric Society (1982), p. 1089-1122

[10] Tadokoro, S., Kitano, H., Takahashi, T., et al.: The RoboCup-Rescue project: a robotic approach to the disaster mitigation problem. IEEE International Conference on Robotics and Automation, San Francisco, CA (2000), p. 4089-4094

[11] RoboCup Rescue Simulation League Agent Competition 2010 Rules and Setup, roborescue.sourceforge.net, http://roborescue.sourceforge.net/2010/rules.pdf

[12] Dias, M. B., Zlot, R., Kalra, N., Stentz, A.: Market-Based Multirobot Coordination: A Survey and Analysis. Proceedings of the IEEE – Special Issue on Multirobot Coordination, IEEE Press (2006),p. 1257-1270

[13] Shoham, Y., Leyton-Brown, K.: MULTIAGENT SYSTEMS Algorithmic, Game-Theoretic, and Logical Foundations. Cambridge University Press (2008)

[14] Wooldridge, M.: An Introduction to Multi-agent Systems. JOHN WILEY & SONS, LTD, London (2002)

[15] Akın, H., Yılmaz, O., Murat Se, M.: RoboAKUT Rescue Simulation League Agent Team Description (2010)

[16] Matsuda, Y., Ueno, H., Komukai, A., et al.: RoboCupRescue - Rescue Simulation League Team Description Ri-one (Japan) 2010

[17] RoboCup2010 Results, roborescue.sourceforge.net, http://roborescue.sourceforge.net /results/ index.html

Author Index

Author Index

Author Index