Jasni Mohamad Zain
Wan Maseri bt Wan Mohd
Eyas El-Qawasmeh (Eds.)

Communications in Computer and Information Science 179

Software Engineering and Computer Systems

Second International Conference, ICSECS 2011
Kuantan, Pahang, Malaysia, June 2011
Proceedings, Part I

Part 1

Springer

Communications
in Computer and Information Science 179

Jasni Mohamad Zain   Wan Maseri bt Wan Mohd
Eyas El-Qawasmeh (Eds.)

Software Engineering and Computer Systems

Second International Conference, ICSECS 2011
Kuantan, Pahang, Malaysia, June 27-29, 2011
Proceedings, Part I

Springer

Volume Editors

Jasni Mohamad Zain
Wan Maseri bt Wan Mohd
Universiti Malaysia Pahang
Faculty of Computer Systems and Software Engineering
Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
E-mail: {jasni, maseri}@ump.edu.my

Eyas El-Qawasmeh
King Saud University
Information Systems Department
Riyadh 11543, Saudi Arabia
E-mail: eyasa@usa.net

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-22169-9 e-ISBN 978-3-642-22170-5
DOI 10.1007/978-3-642-22170-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930423

CR Subject Classification (1998): H.4, H.3, D.2, C.2, F.1, I.4-5

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Message from the Chairs

The Second International Conference on Software Engineering and Computer Systems (ICSECS 2011) was co-sponsored by Springer is organized and hosted by the Universiti Malaysia Pahang in Kuantan, Pahang, Malaysia, from June 27-29, 2011, in association with the Society of Digital Information and Wireless Communications. ICSECS 2011 was planned as a major event in the software engineering and computer systems field, and served as a forum for scientists and engineers to meet and present their latest research results, ideas, and papers in the diverse areas of data software engineering, computer science, and related topics in the area of digital information.

This scientific conference included guest lectures and 190 research papers that were presented in the technical session. This meeting was a great opportunity to exchange knowledge and experience for all the participants who joined us from all over the world to discuss new ideas in the areas of software requirements, development, testing, and other applications related to software engineering. We are grateful to the Universiti Malaysia Pahang in Kuantan, Malaysia, for hosting this conference. We use this occasion to express thanks to the Technical Committee and to all the external reviewers. We are grateful to Springer for co-sponsoring the event. Finally, we would like to thank all the participants and sponsors.

Jasni Mohamad Zain
Wan Maseri Wan Mohd
Hocine Cherifi

Preface

On behalf of the ICSECS 2011 Program Committee and the Universiti Malaysia Pahang in Kuantan, Pahang, Malaysia, we welcome readers to proceedings of the Second International Conference on Software Engineering and Computer Systems (ICSECS 2011).

ICSECS 2011 explored new advances in software engineering including software requirements, development, testing, computer systems, and digital information and data communication technologies. It brought together researchers from various areas of software engineering, information sciences, and data communications to address both theoretical and applied aspects of software engineering and computer systems. We do hope that the discussions and exchange of ideas will contribute to advancements in the technology in the near future.

The conference received 530 papers, out of which 205 were accepted, resulting in an acceptance rate of 39%. These accepted papers are authored by researchers from 34 countries covering many significant areas of digital information and data communications. Each paper was evaluated by a minimum of two reviewers.

We believe that the proceedings document the best research in the studied areas. We express our thanks to the Universiti Malaysia Pahang in Kuantan, Malaysia, Springer, the authors, and the organizers of the conference.

Jasni Mohamad Zain
Wan Maseri Wan Mohd
Hocine Cherifi

Organization

Program Co-chairs

Yoshiro Imai	Kagawa University, Japan
Renata Wachowiak-Smolikova	Nipissing University, Canada
Eyas El-Qawasmeh	King Saud University, Saudi Arabia

Publicity Chairs

Ezendu Ariwa	London Metropolitan University, UK
Jan Platos	VSB-Technical University of Ostrava, Czech Republic
Zuqing Zhu	University of Science and Technology of China, China

Table of Contents – Part I

Software Engineering

Network

Bioinformatics and E-Heatlth

Biometrics Technologies

Web Engineering

Neural Network

Parallel and Distributed

E- Learning

Ontology

Image Processing

Table of Contents – Part II

Information and Data Management

Engineering

Software Security

Graphics and Multimedia

Databases

Algorithms

Signal Processings

Table of Contents – Part III

Software Design/Testing

E- Technology

Ad Hoc Networks

Social Networks

Software Process Modeling

Miscellaneous Topics in Software Engineering and Computer Systems

Use of Agents in Building Integration Tool for Component-Based Application

Ebtehal Alsaggaf[1] and Fathy Albouraey[2]

[1] Faculty of Computing and Information Technology, King Abdulaziz University KAU,
Jeddah, Saudi Arabia
Ebte-alawi@hotmail.com
[2] Faculty of Computing and Information Technology, King Abdulaziz University KAU,
Jeddah, Saudi Arabia
fathy55@yahoo.com

Abstract. Software agents, the one of the most exciting new developments in computer software technology that can be used for both quickly and easily building of integrated enterprise systems. In order to solve the complex problems, agents work cooperatively with other agents in heterogeneous environments to constitute Multi-Agent System (MAS). The concept of building the best component-based application, which contains components from both CORBA & DCOM, becomes very attractive. So, in this research, we will define how component CORBA can interact with DCOM by using multi-agent system. This will able to make us freedom to choose the best attributes from both systems and combine them to build the best possible application that specifically suits my environment. In this paper we are interesting to introduce a new idea of how to build a software integration tool for component-based application by using MA.

Keywords: CORBA, DCOM , COM, mobile agent, MAS.

1 Introduction

The software applications become more complex and computer hardware becomes more powerful and more affordable distributed computing in future. A component is a software object that implements certain functionality and has a well-defined interface that conforms to a component architecture defining rules for how components link and work together. Component-based software is emerged as an important developmental strategy focusing on achieving systems development. It can be defined as a unit of software that implements some known functions and hides the implementation of these functions behind the interfaces that it exposes to its environment [1].

For the components to be able to interact with each other they must comply with the rules of their underlying middleware technology. However, it is difficult, if not impossible, for two components, hosted on different component architectures, to interact with each other. The incompatibility problems stem from the differences of the underlying models and the way they present and use the software components.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 1–15, 2011.

Component technology offers many advantages for scientific computing since it allows reusability, interoperability, maintability, adaptability, distribution that can used easily, efficiently in application development. It speeds the development of applications, operating systems or other components. It enables the developers to write distributed applications in the same way of writing non distributed applications. As a result, scientists can focus their attention to overall application design and integration [2].

Many reusable components are available on the Internet. Finding a specific component involves searching among heterogeneous description of the components in the Internet. There are three most widely-used component standards, which are Component Object Model (COM/DCOM), Common Object Request Broker Architecture (CORBA) and Java/Remote Method Invocation (Java/RMI) [3].

The (DCOM) and (CORBA) are two models that enable software components with different descent to work together. The aims of component-based software development are to achieve multiple quality objectives, including interoperability, reusability, implementation transparency and extensibility [4].

Anyway, Agents are applied as interaction entities to mediate differences between components [8]. The term agent means a variety of things to a variety of people, commonly it is defined as independent software program, which runs on behalf of a network user. It can run when the user is disconnected from the network [4]. However, mobile agent-based computing, being high-level and flexible, can be a useful tool in rapid prototyping due to its high level of abstraction and ease of use.

This research will review the two technologies DCOM and CORBA in their properties and then comes to the differences between them utilizing some dedicated papers. Also we would like to study how to make Integration between CORBA and DCOM technologies in the way in which DCOM and CORBA are differ and resemble, organizing the comparison according to some studied criteria. We will focus in the programming differences according to their function and programming language using MAS. Finally, it will introduce MAS model for building a component-based application.

1.1 CORBA Overview

CORBA is an effective resource in allowing interoperability of heterogeneous systems in a Distributed System. CORBA is emerging as a standard for distributed computing and has a lot of advantages that make use of distributed computing.

CORBA was designed by the Object Management Group (OMG) to primarily provide an object-oriented interoperability of applications in heterogeneous distributed system. The use of Object-Oriented design, analysis, and development using CORBA allows greater reusability across systems. Advantages of Object-Oriented features such as inheritance, encapsulation, redefinition and dynamic binding are implemented in CORBA. They are also effectively easier to extend and modify without affecting other applications and objects.

CORBA encapsulates applications and provides a common infrastructure to communication using the CORBA ORB and is used to receive requests and locate server objects. It encourages the development of open applications that can be integrated to larger systems [3, 4]. Also its advantages include: location transparency, programming language transparency, Operating System transparency and Computer Hardware transparency. The following figure illustrates the primary components in the OMG Reference Model architecture.

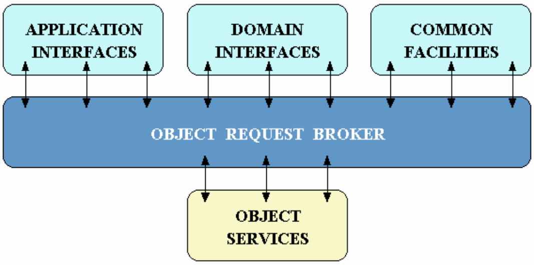


Fig. 1. OMG Reference Model Architecture

1.2 COM and DCOM Overview

COM refers to both a specification and implementation developed by Microsoft Corporation which provides a framework for integrating components that supports interoperability and reusability of distributed objects by allowing developers to build systems by assembling reusable components from different vendors which communicate via COM.

COM defines an application programming interface (API) to allow both the creation of components and use in integrating custom applications or to allow diverse components to interact. However, in order to interact, components must adhere to a binary structure specified by Microsoft. As long as components adhere to this binary structure, components written in different languages can interoperate.

DCOM is the distributed extension to COM (Component Object Model) that builds an object remote procedure call (ORPC) layer on top of DCE RPC to support remote objects. It is best to consider COM and DCOM as a single technology that provides a range of services for component interaction, from services promoting component integration on a single platform, to component interaction across heterogeneous networks. COM specifies that any interface must follow a standard memory layout, since the specification is at the binary level; it allows integration of binary components possibly written in different programming languages such as C++, Java and Visual Basic [1].

1.3 CORBA and DCOM Comparison

Both DCOM and CORBA frameworks provide client-server type of communications. To request a service, a client invokes a method implemented by a remote object, which acts as the server in the client-server model. The service provided by the server is encapsulated as an object and the interface of an object is described in an Interface Definition Language (IDL). The interfaces defined in an IDL file serve as a contract between a server and its clients. Clients interact with a server by invoking methods described in the IDL. The actual object implementation is hidden from the client.

CORBA also supports multiple inheritances at the IDL level, but DCOM does not. Instead, the notion of an object having multiple interfaces is used to achieve a similar purpose in DCOM. CORBA IDL can also specify exceptions [9].

The following terminologies will be used to refer to the entities in both frameworks.

- Interface: A named collection of abstract operations (or methods) that represent one functionality.
- Object class (or class): A named concrete implementation of one or more interfaces.
- Object (or object instance) :An instantiation of some object class.
- Object server: A process responsible for creating and hosting object instances.
- Client: A process that invokes a method of an object[4].

In both DCOM and CORBA, the interactions between a client process and an object server are implemented as object-oriented RPC-style communications. Figure 2 shows a typical RPC structure. In DCOM, the client stub is referred to as the proxy and the server stub is referred to as the stub. In contrast, the client stub in CORBA is called the stub and the server stub is called the skeleton. Sometimes, the term "proxy" is also used to refer to a running instance of the stub in CORBA.

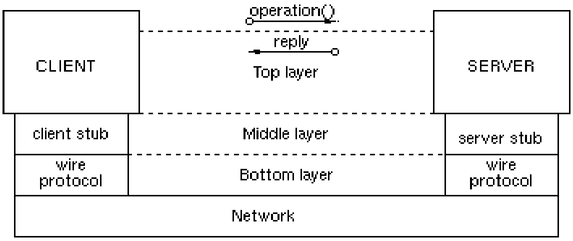


Fig. 2. RPC structure

Their main differences are summarized below. First, DCOM supports objects with multiple interfaces and provides a standard QueryInterface() method to navigate among the interfaces. This also introduces the notion of an object proxy/stub dynamically loading multiple interface proxies/stubs in the remoting layer. Such concepts do not exist in CORBA. Second, every CORBA interface inherits from CORBA::Object, the constructor of which implicitly performs such common tasks as object registration, object reference generation, skeleton instantiation, etc. In DCOM, such tasks are either explicitly performed by the server programs or handled dynamically by DCOM run-time system. Third, DCOM's wire protocol is strongly tied to RPC, but CORBA's is not. Finally, we would like to point out that DCOM specification contains many details that are considered as implementation issues and not specified by CORBA. As a result, they used the Orbix implementation in many places in order to complete the side-by-side descriptions. [5]

Table 1. Summary of Corresponding Terms and Entities

	DCOM	**CORBA**
Top layer: Basic programming architecture		
Common base class	`IUnknown`	`CORBA::Object`
Object class identifier	`CLSID`	interface name
Interface identifier	`IID`	interface name
Client-side object activation	`CoCreateInstance()`	a method call/`bind()`
Object handle	interface pointer	object reference
Middle layer: Remoting architecture		
Name to implementation mapping	Registry	Implementation Repository
Type information for methods	Type library	Interface Repository
Locate implementation	SCM	ORB
Activate implementation	SCM	OA
Client-side stub	proxy	stub/proxy
Server-side stub	stub	skeleton
Bottom layer: Wire protocol architecture		
Server endpoint resolver	OXID resolver	ORB
Server endpoint	object exporter	OA
Object reference	OBJREF	IOR (or object reference)
Object reference generation	object exporter	OA
Marshaling data format	NDR	CDR
Interface instance identifier	IPID	object_key

1.4 Software Agents

Software agents, one of the most exciting new developments in computer software technology, can be used for quickly and easily building integrated enterprise systems. The idea of having a software agent that can perform complex tasks on our behalf is intuitively appealing [9]. The natural next step is to use MAS that communicate and cooperate with each other to solve complex problems and implement complex systems. Software agents provide a powerful new method for implementing these information systems.

Mobile agent-based computing is an attractive, though not widely accepted model for structuring distributed solutions. The most distinctive feature of this model is the mobile agent: a migrating entity, with the capability to transfer its current state and code to a different network location. Compared to remote communication, migration could reduce network traffic. Furthermore, mobile agents can function independently of their dispatching host and contact it later only to return a small set of results. Relevant application domains for mobile agents are distributed information retrieval, monitoring and filtering [7].

There are several reasons for the quite limited acceptance of the mobile agent technology. First, it's quite difficult to identify a distributed problem whose solution can be based on mobile agents only, instead of an equivalent or even better "classical" message-passing or Web Services solution. Another major concern is security: how to protect agents and servers from one another. Nevertheless, mobile agent-based computing, being high-level and flexible, can be a useful tool in rapid prototyping. Due to its high level of abstraction and ease of use, it can also be applied as a teaching tool in introducing students to distributed computing [8].

However, applications require multiple Agents that can work together. A MAS is a loosely coupled network of software agents that interact to solve problems [6, 7]. The difficulty arises from the need to understand how to combine elements of various content languages and interaction protocols in order to construct meaningful and appropriate messages [10] but, it has the following advantages [7]:

1. A MAS distributes computational resources and capabilities across a network of interconnected Agents. A MAS is decentralized and thus does not suffer from the "single point of failure" problem in centralized systems.
2. A MAS allows for the interconnection and interoperation of multiple existing systems.
3. A MAS efficiently retrieves, filters, and globally coordinates information from sources that are spatially distributed.
4. In MAS, computation is asynchronous.
5. A MAS enhances overall system performance, efficiency, reliability, extensibility, robustness, maintainability, responsiveness, flexibility, and reuse.

2 Related Work

2.1 COM-CORBA Interoperability

To make distributed objects work in a heterogeneous environment, developers must bridge the gap between Microsoft COM/DCOM and the industry CORBA standard. This is the first complete, up-to-date guide to doing so. It starts with easy-to-understand descriptions of both COM and CORBA, exploding the myth of complexity that surrounds these technologies. Next, it delivers a step-by-step guide to building your own working, scalable and transparent COM/CORBA systems, integrating Windows and UNIX. The CD-ROM includes MS-Access source code for all examples, plus trial versions of IONAs Orbix COMet, the first commercial bridge for linking COM and CORBA modules, and OrbixWEB 3.0 tools for building Internet-based CORBA Server applications [1].

2.2 Multi-technology Distributed Objects and Their Integration

Most of the work in the area, they surveyed concerns bridging CORBA and DCOM. This is expected considering the widespread deployment of Microsoft's operating systems and the acceptance of CORBA as the most mature middleware architecture. Moreover, the early presence of a variety of COM components and ORB products from commercial companies led developers to use those products. As a result the bridging between CORBA and DCOM was an urgent need.

They can distinguish two basic approaches for bridging, the static bridging, and the dynamic bridging. Under static bridging, the creation of an intermediate code to make the calls between the different systems is required. The disadvantage of the static bridge is that any changes on the interfaces require a change in the bridge. In dynamic bridging there is no code depended on the types of calls.

The implementation belongs to commercial companies which have released many bridge tools, compliant with OMG's specification. Some of these products are PeerLogic's COM2CORBA, IONA's OrbixCOMet Desktop, and Visual Edge's ObjectBridge. All the above products realize one of the interface mappings that OMG specifies [2].

2.3 A Component-Based Architecture for Multi-Agent Systems

In a large system, some problem solving required agents that have the BDI set, the MAS is not only heterogeneous but also has a heavy overhead on the system execution. This complexity must be resolved at the architecture level so that in an implementation the complexity does not arise. In [6], they introduced a formal multilayered component-based architecture towards developing dependable MAS. They had not investigated any specific approach for verifying the MAS design yet. However, it seems feasible that they could provide a uniform platform for both programming and verifying MASs, if they provided reasoning rules for Lucx.

After we studied the above researches, it can be feasible for us to design MAS as integration tool for component-based application by investigating specific widely component standards, which are DCOM and CORBA.

3 The Current Work

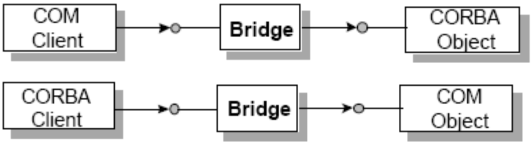
There are several phases to develop the best component-based application, the first phase is analysis phase which is concerned on user requirements and then present system model which is corresponds to use cases in object oriented design. The second phase is design which specifies the different roles to be developed in the software system, and their interactions. It is composed of the role model and the interaction model. The third phase is implementation. The last phase is testing phase which defines the types of used tests. These phases called Software Development Cycle.

3.1 System Analysis

Implementing distributed computing presents many challenges with respect to middleware in general and CORBA and DCOM specifically. Depending on both the business and technical problems that need to be solved, the greatest probability is make Integration between CORBA & DCOM to build the best component-based application from hybrids of the best technology and tools available at the time and the concept of building application which contains components from both CORBA & DCOM, is giving you the freedom to choose the best attributes from both technologies [1,2].

How do we make this Integration?

For many organizations, a business and technology need exists for "Integrating" between CORBA and DCOM. This generally means providing a bridge between CORBA and COM, and two mappings.



In order to transparently couple components from DCOM and CORBA, some of bridging software is needed to handle the translation of types and object references between the two systems.

What actually required building abi-directional Bridge between CORBA & DCOM. ?

We should to be able to do the following [1]:

- Transparently contact object in one system to other.
- Use data types from one system as though they were native type in the other system.
- Maintain identity & integrity of the types as they pass through the bridge in order to reconstitute them later.

We can distinguish two basic approaches for bridging, the static bridging, and the dynamic bridging

Static bridging**:** This provides statically generated marshalling code to make the actual call between the object systems. Separate code is needed for each interface that is to be exposed to other object system. Static bridging also implies that is an interface-specific package (DLLs, configuration files, ect.) which needs to be deployed with client application.

Dynamic bridging: This provides a single point of access between the two systems which all calls go through. No marshalling code is required to expose each new interface to the other object system.

In either case, a proxy is needed on the client side to intercept and pass on the call to remote machine, with a stub on the server side to receive it. (Of course, if the call is being made in-process, it will occur directly between the calling object and the target object, with no proxy or stub required) Hence, all that is required to provide a bridge between CORBA and DCOM is to provide some thing (bridge call) which belongs to the current object system and sent the bridge call to MAS.

Under static bridging in the two technologies, the creation of an intermediate code to make the calls between the different systems is required. That intermediate code would be called the bridge object which could be sent and receive the call function to and from MAS. In dynamic bridging, the operation is based on the existence of a dynamic mechanism which can manage any call in spite of the interfaces[1, 2].

To make the Integration:

- First: We will build the bridge between the client and server for both CORBA & DCOM which is integrator model between them.
- Second: We will build the bridge between two components for both CORBA & DCOM which is also integrator model between them.

The same bridge use in both ways. In our model, the integrator (bridge) is a MAS, which can contain set of software Agents that may run on one computer, and may be distributed on different computers in the network. The system contains different Agents having different functions and tasks.

Based of the features of Multi-Agent system MAS, it is natural to introduce MAS in our system where it can be applied to bridging and mapping the two most widespread technologies.

When the component need to call another component in the same techniques nothing to do else if the component need to call another component in different techniques then send a message to MAS.

3.2 System Architecture

The objectives of design this system model is to make Integration between CORBA and DCOM technologies in the way in which DCOM and CORBA are differ and resemble, organizing the comparison according to some studied criteria .This system consists of Combine Agent, Mapper Agent, Manager Agent, Agent library, DCOM component and CORBA component. The sequence of the work among Agents as the following:

1. The Manager Agent receives the massage of the call function which sent by CORBA component (for example), the main job is achieved by the Manger Agent. The functions of Manger Agent are:

    - Receive this call and determine the kind of technology.
    - Send this call to Mapper Agent.
    - Manages all active Agents.
    - Last, sent the mapping call to other technology.

2. The Mapper Agent. Separate the formula and sent all sub formula to corresponding agent, as the following:

- The Interface Agent takes and reads its Interface, comparing CORBA Interface to its corresponding DCOM from library Agent and then written the equivalent one of new mapping Interface
- The function Agent takes and reads its function, comparing CORBA function to its corresponding DCOM from library Agent and then written the equivalent one of new mapping function.
- The data type Agent takes and reads its data, comparing CORBA data type to its corresponding DCOM from library Agent and then written the equivalent one of new mapping data type

3. Agent library has three tables: function table [Tablt3] and data type table [Tablt2] and Table of the corresponding terms and entities [Tablt1]. These tables are be fixed and stored in the database. The function table has two columns, one for DCOM functions and the other column is for the corresponding function in CORBA. The same technique will be applied for data type table in which each data type of DCOM arranged to its corresponding data type in CORBA , Table of the corresponding terms and entities summarizes the corresponding terms and entities in the two architectures.
4. Agent library will come then after receiving a message from Interface Agent, function Agent or The data type Agent .For example, if Interface Agent will send message to Agent library that will execute a query to the database using JDBC (Java Data Base Connectivity). The result of query will be represented as object. This will result in an array of objects (may be one object).
5. The Combine Agent takes the result of mapping from data type Agent, function Agent and Interface Agent. Then combine them to building anew DCOM formal and sent this formal to the Manger Agent which sending them to the DCOM component.

Table 2. Data Type Table

DCOM	CORBA
short	short
unsigned short	short
long	int
Unsigned long	int
double	double
float	float
char	char
Boolean	Boolean
byte	byte

Table 3. Function Table

DCOM	CORBA
IUnknown	`CORBA::Object`
QueryInterface	-
Addref	-
Release	-
CoCreateInnstance	a method call/`bind()`
UUID	-
CLSID	interface name
get()	get()
set()	set()

6. The DCOM component receives the call formula to complete the operation, and then produce the result of function which will return back to MAS.
7. The MAS will be applied the same technique for a DCOM result to produce a new CORBA formal after mapping in it (which may be the replying of call function).

At the end, The Integration between CORBA & DCOM technologies will be completed (see Figures 3, 4).

4 Discussion

This section will briefly discuss some of the advantages of our system model. Some of these advantages are applicable to all distributed computing technologies, including CORBA and DCOM. In additional, it will also add the advantages of MAS that we have mentioned in section 1.4, thus our system model can able to make how to combine elements of various Interfaces , data types and functions in order to construct meaningful and appropriate messages .So it will have the following advantages:

1. **The components will able to interact with each other** for two components, hosted on different component architectures. in appendix section, We describe the details of design interaction diagram for the whole system and for each agent .
2. **Speeding up development processes**: Since applications can be built from existing pre-built components, this helps to maintain the speed up of development process tremendously.
3. **Improving deployment flexibility**: Organizations can easily customize an application for different areas by simply changing certain components in the overall application.
4. **Lowering maintenance costs**: Certain functions of an application can be grouped into discreet components, which can be upgraded without retrofitting the whole application.
5. **Improving scalability:** Since applications are built from many objects, the objects can be redeployed to different machines when needs arise or even multiple copies of the same components can run simultaneously on different machines.

Moreover, it will achieve multiple quality objectives for developing it, including:

6. **Interoperability and reusability** by using block interfaces which have all functions and their parameters for each agent in the MAS model
7. **Implementation transparency and extensibility** were done by using all functions and their parameters for each agent and also by representing the database schema which include three tables that stored in it (see appendix section).

Finally, we will build a software integration tool for component-based application by using MAS.

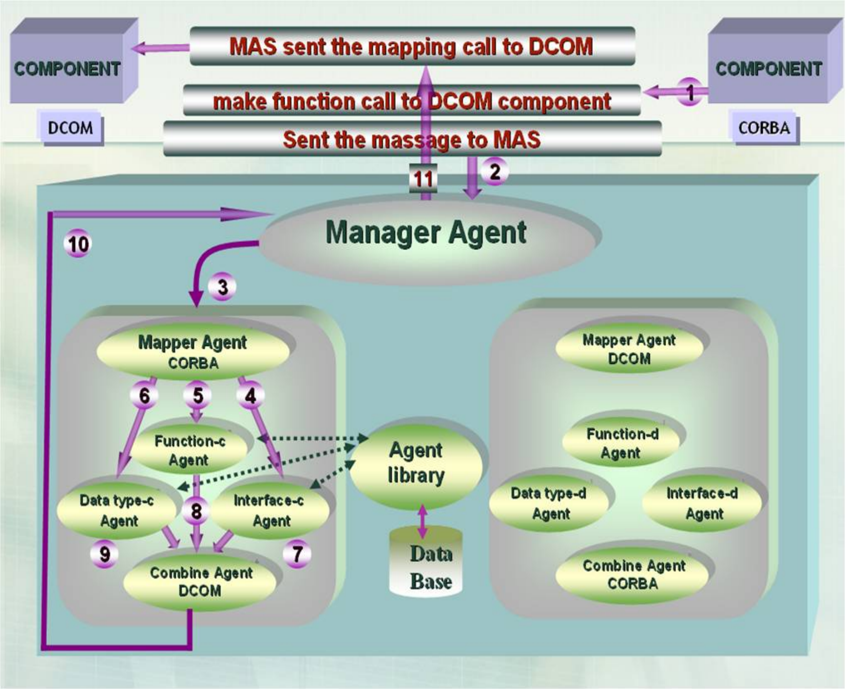


Fig. 3. CORBA to DCOM steps in MAS

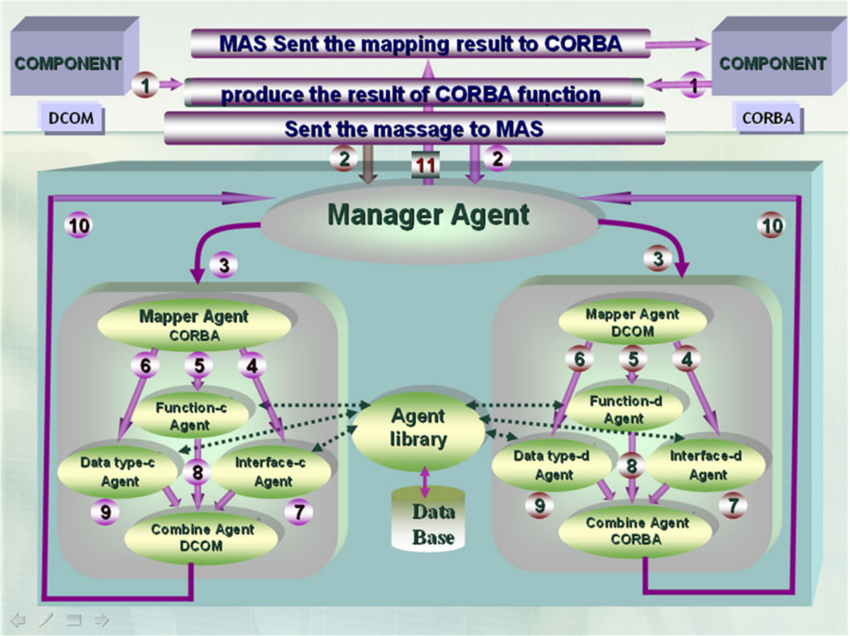


Fig. 4. DCOM to CORBA steps in MAS

5 Conclusion and Future Work

As a result the bridging between CORBA and DCOM was an urgent need. For software component to integrate with each other, it is difficult if not impossible for two objects conforming to dissimilar technologies to interact with each other. The above model is the way which uses the software agents to build a software integration tool between CORBA and DCOM technologies which will give us the freedom to choose the best attributes from both systems and combine them to build the best possible component-based application.

Benefits of our system that they able to transparently contact object in one system to other rely on the features of MAS and it also use data types from one system as though they were native type in the other system by Agent library with DB. This will maintain identity and integrity of the types as they pass through the bridge in order to reconstitute them later. This will achieve multiple quality objectives, including interoperability, reusability, implementation transparency and extensibility.

In the future work, we want to improve our model to support integration tool for all Component technologies to build a software component-based application by using MAS. MAS will be modifying by adding new two mapping for each new technology which it adding to it. Finally, the system that we have designed stills a basic model. So, we will interest to complete the system developing and implementing these.

References

1. Geraghty, R., Joyce, S., Moriarty, T., Noone, G.: Com-Corba interoperability. Prentice Hall PTR, Upper Saddle River (1999)
2. Raptis, K., Spinellis, D., Katsikas, S.: Multi-technology distributed objects and their integration. Computer Standards & Interfaces 23(3), 157–168 (2001)
3. The CORBA Programming Model (2008), `http://download.oracle.com/docs/cd/E15261_01/tuxedo/docs11gr1/tech_articles/CORBA.html`
4. IBM, Is web services the reincarnation of CORBA? (2001), `http://www.ibm.com/developerworks/webservices/library/ws-arc3/`
5. Pritchard, J.: COM and CORBA side by side: architectures, strategies, and implementations. Addison-Wesley Professional, Reading (1999)
6. Wan, K.Y., Alagar, V.: A Component-Based Architecture for Multi-Agent Systems. IEEE Computer Society, Los Alamitos (2006)
7. Agent Technology, Green Paper, Agent Working Group, OMG Document ec/2000-03-01, Version 0.91 (2000)
8. Adey, R.A., Noor, A.K., Topping, B.H.V.: Advances in Engineering Software (2010)
9. Manvi, S.S., Venkataram, P.: Applications of agent technology in communications: a review. Computer Communications 27(15), 1493–1508 (2004)
10. Shepherdson, J.W., Lee, H., Mihailescu, P.: mPower—a component-based development framework for multi-agent systems to support business processes. BT Technology Journal 25(3), 260–271 (2007)

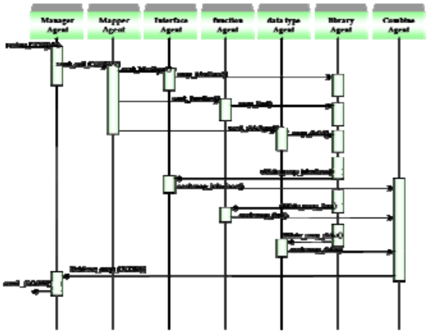
Appendix: System Detailed Design

Detailed design phase consists of the following steps:

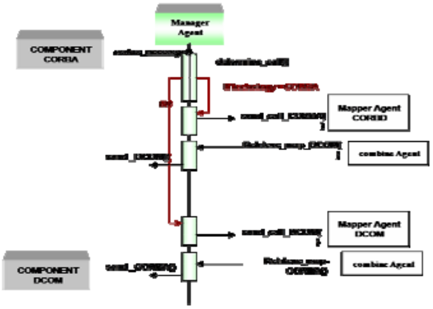
- Design Interaction diagram for the whole system and for each agent which presents the agents and messages between them.
- Design Block interfaces by written all functions and their parameters for each agent
- Design database schema.

1 Interaction Diagram

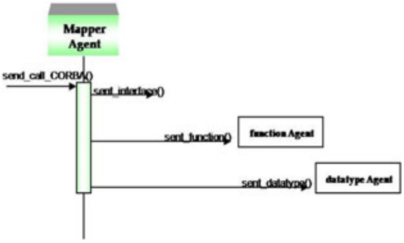
1. Interaction Diagram for the model



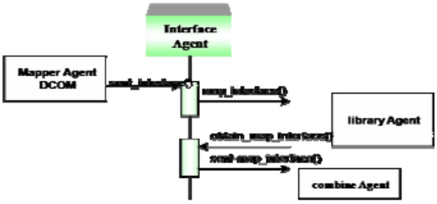
2. Interaction Diagram for the Manager Agent



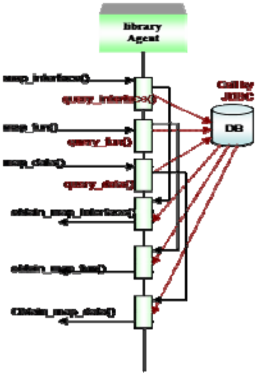
3. Interaction Diagram for the Mapper Agent



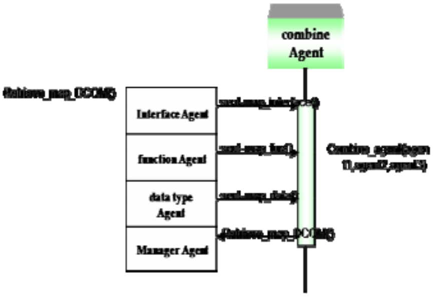
4.Interaction Diagram for the Interface Agent



5.teraction Diagram for the library Agent



6.Interaction Diagram for the combine Agent



2 Block Interfaces

: **Manager Agent Functions**

recive_massege()

determine_call()

send_call_CORBA()

send_call_DCOM()

Retrieve_map-CORBA()

Retrieve_map_DCOM()

send _CORBA()

send _DCOM()

: **Mapper Agent Functions**

sent_interface()

sent_function()

sent_datatype()

: **library Agent Functions**

query_interface()

query _fun()

query _data()

obtain_map_interface()

obtain_map_fun()

obtain_map_data()

: **Interface Agent Functions**

map_interface()

sent-map_interface()

: **function Agent Functions**

map_fun()

sent-map_fun()

: **data type Agent Functions**

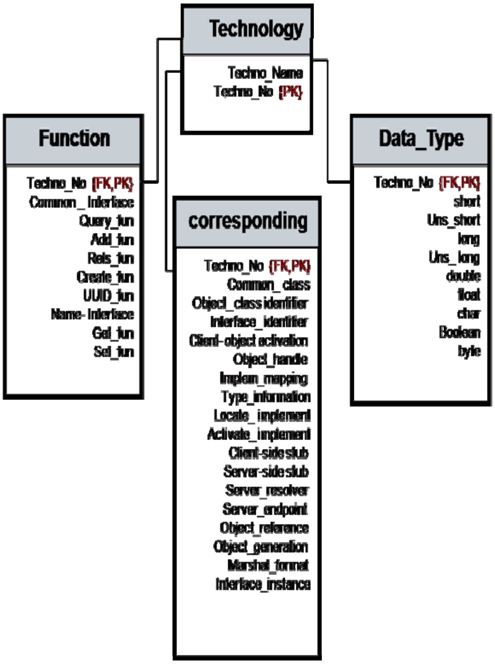
map_data()

sent-map_data()

: **Combine Agent Functions**

Combine_agent(agent1,agent2,agent3)

Send_result()



3 Database Schema

The above structure represent the data base tables in the model :

Towards Incorporation of Software Security Testing Framework in Software Development

Nor Hafeizah Hassan, Siti Rahayu Selamat, Shahrin Sahib, and Burairah Hussin

Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Durian Tunggal, Melaka,
Malaysia
{nor_hafeizah,sitirahayu,shahrinsahib,burairah}@utem.edu.my

Abstract. The aim of this paper is to provide secure software using security testing approach. The researchers have reviewed and analyzed the software testing frameworks and software security testing frameworks to efficiently incorporate both of them. Later, the researchers proposed to fully utilize the acceptance testing in software testing framework to achieve by incorporating it in software security testing framework. This incorporation is able to improve the security attribute needed during requirement stage of software development process. The advantage of acceptance test is to expose the system of the real situation, including vulnerability, risk, impacts and the intruders which provide a various set of security attribute to the requirement stage. This finding is recommended to establish a baseline in formulating the test pattern to achieve effective test priority.

Keywords: test pattern; software testing frameworks; security testing framework; security requirement; software process.

1 Introduction

In software industry, the testing process plays a crucial role, as people try to find defects of software [1]. The defects reflect the quality status of software on whether it should be ready to release. According to Standish's CHAOS Summary 2009, only 32% software was released successfully and others were neglected for many reasons [2]. Among the reasons for the negligence is the existence of defects in the software and thus, it is unfit to be deployed.

However, in the last few years, the industry is swamped with the term security testing which is claimed to find vulnerabilities in software [3]. The vulnerabilities are affirmed as any flaw that may be exploited by a threat. A research by US-Computer Emergency Readiness Team (CERT) shows total of 6000 vulnerabilities was cataloged in the first three quarters of 2008 [2].

Therefore, to control the situation, many researchers develop their own models or frameworks to reduce the numbers of bugs and vulnerabilities as in [4] and [5]. This

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 16–30, 2011.

paper analyzed two perspectives of testing (software testing and software security testing) to analyze and propose the possible incorporation between the two in software development process or also known as software development life cycle.

The rest of the paper is organizes as follows: section 2 present related work of testing and security testing. Section 3 discusses the analysis approach used to evaluate the software testing and software security testing frameworks and the last two sections present discussion and future work.

2 Related Work

Testing is a task in software development process. However, there are various software development process style such as waterfall, V-shape, Rational Unified Process (RUP), agile, and secure software development process as discussed in [6], [7] and [8]. Therefore, the way testing is conducted within each style may vary. A comparison analysis of software development process style is conducted to determine the testing roles.

2.1 An Overview of Software Development Process

Software development process is a life cycle of developing software. A traditional lifecycle consist of processes, namely as: requirement, analysis, design, coding, testing and operation [6]. A modified lifecycle consist of business modeling process (as in RUP), timeline and iteration (as in V-shape, spiral, RUP and agile). In this paper, the process is also referred as stage to emphasis the element of sequence. A fundamental stage of software development process was grouped based on similar activities on both traditional and modified lifecycle as represented in Table 1.

Table 1. Summary of software development model (√ = the respective stage exist is the software development style, Not stated = the respective stage is not described or not exist in the software development style)

Stage	SOFTWARE DEVELOPMENT					
	Waterfall	Spiral	V-shape	RUP	Agile	Secure software development
Feasibility	Not stated	√	Not stated	√	Not stated	Not stated
Requirement	√	√	√	√	√	√
Analysis	√	√	√	√	√	√
Design	√	√	√	√	√	√
Coding	√	√	√	√	√	√
Testing	√	√	√	√	√	√
Operation	√	√	√	√	√	√

As shown in Table 1, the Waterfall, V-shape, Agile and Secure Software Development consists of six stages starting from the requirement stage. On the contrary, the Spiral and RUP consist of seven stages including the feasibility stage. Each of the stage has unique roles with their own set of activities. Feasibility is a stage of understanding the business concept; determine the objectives, alternatives and constraints; and estimate monetary resources. Requirement is a stage of elicit the stakeholders requirements within the project scope for what the system must do. Analysis is a stage of breaking down the requirement into workable process. Design is a stage of transforming the analysis into its architectural view, which may include high level design and low level design. Coding is a stage of writing the programming code into modules, test the individual developed modules (unit test), and integrate with other author modules as a functional system. Testing is a stage of checking the code implementation which covers various levels such as unit test, integration test and acceptance test with the objective to find defects, fixed it, and later verify it with the requirement. Operation is a stage of deploying the user accepted system into real environment, provide training, get it exploited, receive feedback and perform maintenance. The activities in software development stages are summarized as in Fig. 1.

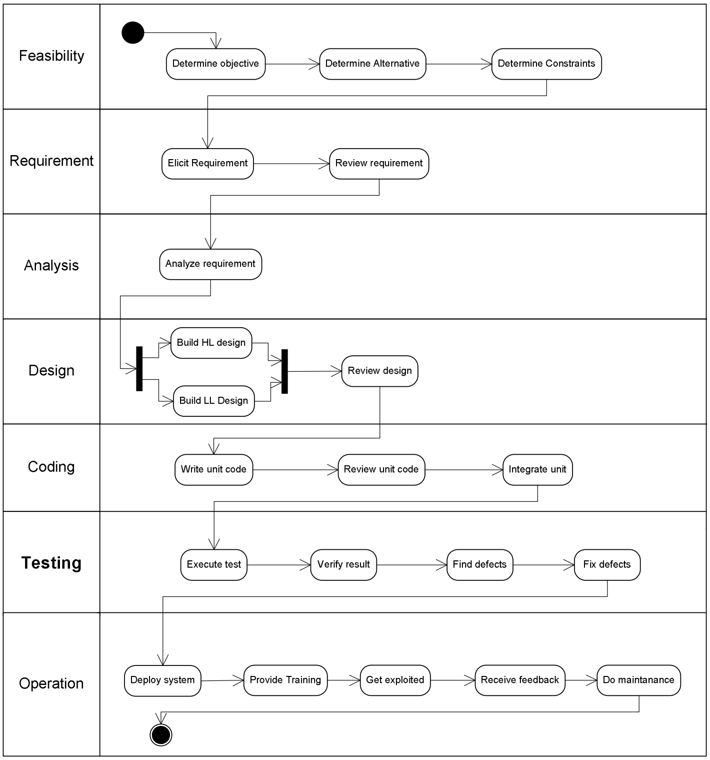


Fig. 1. Activities in software development stages. Testing is a stage of checking the code implementation which covers various levels such as unit test, integration test and acceptance test with the objective to find defects, fixed it, and later verify it with the requirement.

Thus, given a typical or a secure software development process model, the testing is a stage between coding and operation as shown in Fig. 1, and any dissatisfaction of testing stage require an assessment either at coding stage, design stage or requirement stage. To further understand the testing stage, an overview of software testing component is discussed in the next section.

2.2 An Overview of Software Testing

Software testing (ST) is a process of executing a program to find the defects. Testing involves searching for runtime failures and recording information about runtime faults [9].

There are two basic components in ST. First is the test strategy which explains the way a test is conducted; involve the reveal of the code. If the code is reveal, it is white-box testing, else, it is black-box testing. If the code is used to design a test for black-box testing, it is called grey-box testing. Second is the test level which address the entire software development life cycle and each test level is a foundation of the higher level. The three levels are unit testing, integration testing and acceptance testing. Each level is viewed by programmers, designers, and business users respectively [10]. Both, the test levels and test strategy are summarized as shown in Fig. 2.

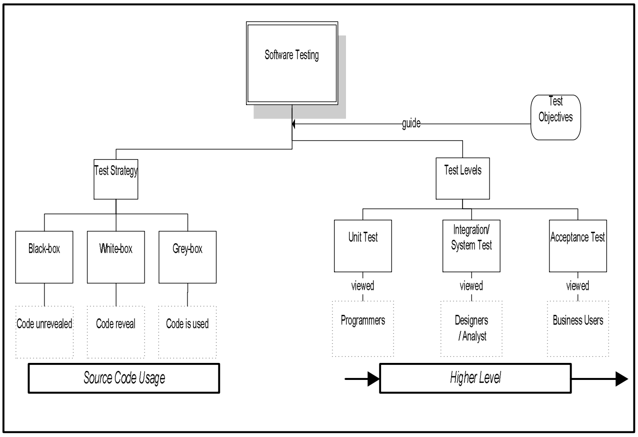


Fig. 2. The basic components in software testing: a) test strategy consist of black-box, white-box and grey-box, b) test level consist of unit test, integration or system test and c) acceptance test

The test strategy assists in accomplishing the test levels objectives. A unit test use white-box testing strategy, integration test use grey-box testing strategy, and both integration and acceptance test use black-box testing strategy [4]. The test strategy and test levels chosen are guided by the test objectives, i.e. to answer why testers do the test. Test objectives are the factors that classify a test into a specific test type as described in [10] and [11]. The test level is viewed either by (involved) programmer, designer or business developers. In short, basic testing components are test level, test strategy and users.

In order to show the importance of the testing stage, we highlight the expectation of test: a) to find defect, b) fixed it and c) verify with requirement. Therefore, a comparison of actual test result and expected test result must be conducted. Any discrepancy of the two shall trigger the tester to trace the origin of error or fault.

The issue was pointed out as early as in 1970, when [6] proposed the manageable waterfall software development by introducing skip-over the code and analysis stages to design stage if the testing stage fail to satisfy the external constraints. In 1986, [12] imposed the risk, prototype, review and evolutionary elements in spiral software development. In their model, testing is a notion in more composite manner; known as the test level. The test level describes the incremental of test consists of unit test, integration test, and acceptance test. However, there is no clear guideline of where to point back in case the testing result is not satisfied. It give the impression that coding and design, which are the nearest adjacent stages to unit testing, are good candidates for the purpose.

The same notion of *test level* is followed in V-shape model with illustration that each test level shall correspond to the development process stages respectively [13]. RUP, the IBM popular software development, denote testing as the verification process of all objects, integration and application requirements. Therefore, any dissatisfaction in testing caused a revisit to the code and requirements as concluded in [10] and [13]. In the iterative, incremental and user-centered agile software development, automated acceptance testing is emphasized to keep the prototype evolve in the next iteration [8]. The relatively new secure software development, proposed a secure development process within four stages, namely as: requirement, design, code and feedback [11]. The design stage consists of analysis and design, where as the feedback stage is the deployment stage. Major test is done in code stage such as risk-based security test and pent-test. Another secure software development by Microsoft, combined both spiral and waterfall approach, outlined an extended stage of testing named as verification stage prior to release and response stage [14]. Again, it is not clearly mentioned here, at which stage shall the developers return to in the event of test discrepancy occurs. Based on the stage sequence and the artifacts produced, the assumption is to check the code, design or requirement stage.

Currently, the aim of ST is to find errors or to verify test result with requirements [15]. However, the complexity of software had emerged and invites unintended users to penetrate the system. Hence, testing for boundary values only is insufficient to guarantee system functionality and minimize potential vulnerabilities. As a result, security testing is required to overcome this current issue.

2.3 Software Security Testing

The current practice of software security testing (SST) is to detect any defects that contribute to the flaws exhibit in an application and always considered as an afterthought concerns [11]. SST is concerns with two objectives: first, the test is executes to find what should not happened within a system and second, is to disclose any attack from intruders. The attack can breach the security by exploiting the vulnerabilities (weakness) exist in the application. Therefore, the software security testing scope requires a tester to be equipped not only with testing expertise but also with software security knowledge [11].

The distinction between ST and SST come in twofold. Firstly is on the second objective of the SST, i.e. the existence of intruder or attacker element as discovered by [1] and [16]. Secondly is on the existing research of ST which focuses on how to combine the testing components to produce an effective testing result, where as SST focus on how to eliminate the afterthought concern by incorporate the security aspect into the whole software development from the beginning stage. It is debatable that ST utilizes all the test components effectively and how SST could utilizes ST components to achieve its objectives [16]. To the best of our knowledge, the position of this SST within the current view on software testing diagram or vice versa is yet to be determined. Thus, an analysis on software security testing frameworks needs to be conducted in order to establish a comprehensive understanding of the SST within the software development.

3 Analysis Approach

In order to obtain the perspective, an analysis of software testing and software security testing frameworks is conducted

3.1 Software Testing Framework

Software Testing Framework (henceforth, STF) is an approach used to perform testing as effective as it can [5]and [10],. As the testing phase of a software life cycle is extremely cost intensive (40% of the whole budget) many researchers look into it with various perspectives. A framework could consist of an approach, a technique or a model. Generally, a testing framework execute using five steps: a) identify the test objective, b) generate testing input using application's specification, c) produce expected output results, d) execute and validate the test cases, and e) amend the application following with regression test [17] and [18]. The objectives of the test shall consider the type of application (software) under test. For example, during a test in web application software, the number of distributed users is enormous, an aspect that need to be concerned.

In [4], they introduced meta programming in testing framework in order to reduce the test preparation load by overcome the three challenges in testing i.e. poorly design test cases, manual works and discrepant tools. They implemented the framework in unit test and system test level. The tool was developed using Java. Meanwhile, [18] and [19], proposed a multi-agent system architecture for automated testing framework in distributed environment to deal with different type of users. The framework in [19] focused on four properties during integration testing level, namely: interoperability, compatibility, function, and performance. [20] proposed a new algorithm to verify completeness and consistency property in web services . In an extended work of unit testing framework, Diffut, run on Java to test the difference between two same inputs using Rostra and Symtra [21]. A slight different domain, in electronic health records environment, introduce the Archetype Definition Language (ADL) in their testing framework [22]. Next, in [23], an improve framework of regression test is demonstrated in database application. The framework used DOT-select as it test tool which is part of a larger Data-Oriented-Testing framework that is under development at the University of Manchester. In another work, [24] explained the testing framework for model transformation, that is to test the changes in graphical model. The framework used an extended declarative language, Embedded Constraint Language (ECL) for describing rules (applied in UML). [25] introduced a web services test framework by mapping the Web Service Definition Language (WSDL) to Testing and Test Control Notation Version 3 (TTCN-3) using a third party test tool, TTworkbench. TTworkbench is the full-featured integrated test development and execution environment (IDE) test automation.

The frameworks are selected based on as view by (involve) developer, such as unit test or integration test or regression test. In order to review the framework, the common fundamental aspects in software development are extracted. This research discloses that the aspects used to determine the perspective in software testing framework are programming language, tools, domain, standard, test level and test strategy as summarized in Table 2.

Table 2. Common aspect extracted from software testing framework. (Not stated = the common aspect is not found in the reviewed framework*)*

Author	**Common Aspect**					
	Language	**Tools**	**Domain**	**Standard**	**Test level**	**Test strategy**
[4]	Java	JUnit, CUnit, CPPUnit	Distributed	Not stated	Unit test	White-box
[18]	Object-oriented	Not stated	Web-based application	Not stated	Unit test Integration test System test	White-box, Grey-box
[19]	Java	JADE	Distributed	Unified interface standard	Integration test	White-box, Grey-box
[20]	OWL-S	Not stated	Web-based application	SOAP	Unit test, Integration test	White-box, Black-box
[21]	Java	Rostra, Symtra	Not stated	Not stated	Unit test	White-box
[22]	Archetype Definition Language	Not stated	Distributed Electronic Health Record	Not stated	Unit test	Black-box
[23]	XML	JDBC (DOT-select)	Database	Not stated	Regression test	Black-box
[24]	Embedded Constraint Language	Generic Modeling Environment	Model transformation	Not stated	Unit test, Integration test	White-box, Grey-box
[25]	WSDL	TTWorkbench	Distributed, Web-based application	TTCN-3, ETSI, ITU, SOAP	Unit test	Black-box

JADE=Java Agent DEvelopment Framework, OWL-S=Ontology Web Language for Service, WDSL=Web Service Definition Language, TTCN-3=Testing and Test Control Notation Version 3, ETSI=European Telecommunications Standards Institute, ITU=International Telecommunication Union , SOAP=Simple Object Access Protocol

Based on Table 2, the programming language aspect explains any specific language used in the framework, either as object-oriented or specific name such as Java, eXtended Markup Language (XML), ADL, ECL or WSDL. The tools aspect is any assistant used to demonstrate the author's chosen algorithm or techniques. The domain aspect focuses the platform of the software development such as in general (distributed or web-based environment) or specific (model transform, database or others). The standard aspect composed of any policy adhere and enforced by the author, such as access protocol in Simple Object Access Protocol (SOAP) TTCN-3or unified interface standard for open source. For example, most of web-based or distributed testing frameworks follow the standard used specifically in communication. Regression test, a re-test after fixing a bug is noted significantly in database application to maintain the database state whenever changes occurred. The grey-box test is relevant in two scenarios, firstly is when the code was not directly revealed such as in agent-based testing framework [18], [19] and model testing framework [24], but rather is invoked by other metadata. Interestingly, the methodology used such as waterfall or agile is not a major concern in these reviewed STF.

The findings of the analysis is then summarized and mapped into the software development stages as illustrated in Fig.3. Fig. 3 shows that test strategy is used to support the unit test and integration test. It is noted that the test components summarized in the testing framework (domain, language, tools and standards) limits the test level conducted by only developer and designer (based on unit test, integration test or regression test). These internal testers are claimed to assess the design stage to compare the test result [10]. Consequently, this situation allowed the review process to go as further as to design stage only.

Business user which is claimed to prefer review the requirement stage (as have minimum knowledge in software design or programming language) in finding out the discrepancy of test [13], is not yet engaged. The lack of business user involvement in the reviewed frameworks had limit utilization of test level to system test. Another test level, the acceptance test, only could be utilized in the existence of business users. The diamond shapes in Fig.3 denote the defect-by-developer is defect found as test conducted by developer and defect-by-designer is defect found as test conducted by designer.

Hence, the aspects discussed in this software testing framework are addition to existing testing components discussed in Section 2.2. Therefore, apart from test strategy and test level, a testing framework also depends on language and domain aspects assist by tools and benchmark by standards.

Based on the findings, this research proposed the issues tackled in the STF framework are simplified into five objectives : a) to achieve the objective of test (checking for completeness, consistency, interoperability as in [20], [22] and [24], b) to optimize the test as in database state [23], c) to reduce the test load as in [4] and [21], and d) to adhere the required test guideline (standard) as in [19], [20] and [25]. In order to add the essential security issue, an analysis of software security testing framework is discussed in the next section.

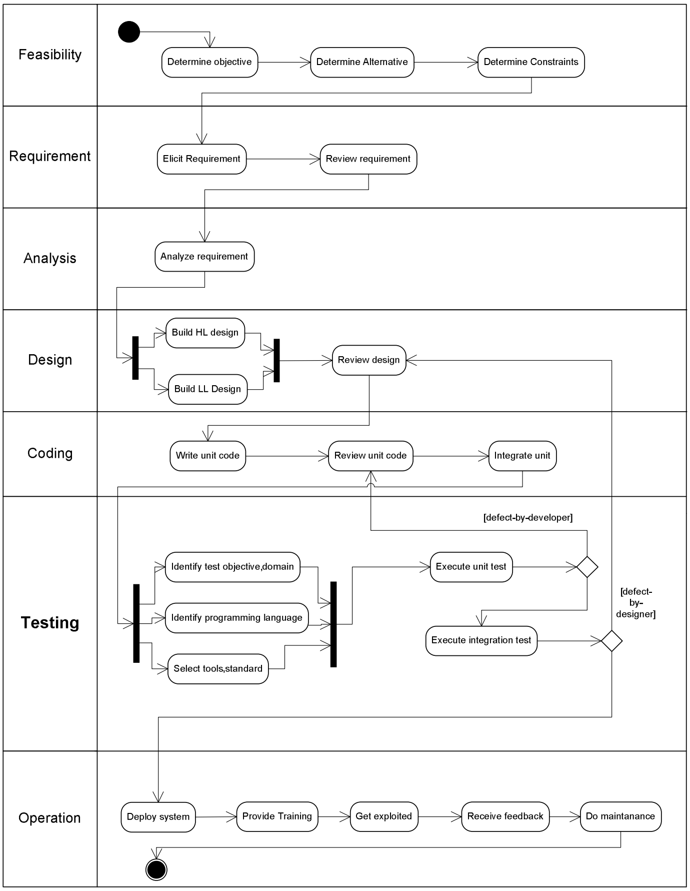


Fig. 3. The mapping of software development stages. The test components summarized in the testing framework (domain, language, tools and standards) limits the *test level* conducted by only developer and designer (based on unit test and integration test).

3.2 Software Security Testing Framework

Security testing framework (SSTF) is an approach used to test the security aspects of a product [16]. Always, this framework is considered as an afterthought concerns, i.e. the process only begin once the product is ready.

A comparison of security framework had been made by [26]. They compared eleven frameworks from year range 1996 to 2004. They highlighted the needs for a standardized methodological approach that taking into account security aspects from the earliest stages of development till the completion. Another work was conducted by [27] to summarize the security dimensions, such as cause, impact, and location, encountered in security frameworks. However, there is still lack of research done of how to integrate testing operation in software development process [28].

For the purpose of this paper, we examined eight SSTF frameworks to disclose the attributes involved. The frameworks are selected based on the security aspects integrate or enforced within. The selected SSTF are *Knowledge Acquisition Automatic Specification (KAOS), Model Driven Security (MDS), i*, Secure Tropos,* Security *Quality Requirements Engineering (SQUARE)* methodology, *Security Requirement Engineering Process (SREP), Security Requirement Engineering (SRE)* and *Threat Modeling*. *KAOS* started from cooperation between the University of Oregon and the University of Louvain (Belgium) in 1990. *KAOS* is a goal-oriented software requirements capturing approach in requirements engineering which has been extended to capture security threat using it anti-goals [29]. *MDS* is a framework that automatically constructing secure, complex, distributed, applications with the UML integration [30]. The *i** framework was developed for modeling and reasoning about organizational environments and their information systems which later embed the trust model [31]. *Secure Tropos* is an extended approach from *Tropos*. It explains a formal framework to model and analyze security requirements that focus on ownership, trust and delegation [32]. *SQUARE*, is a nine-step approach to elicit, categorize and priority security requirement [33]. *SREP* which is similar to *SQUARE* imposed the standards and policy enforcement (Common Criteria) within the framework [34]. *SRE* is a framework to determine how adequate is a security requirements [35]. There are particular frameworks that had been adapted into tools such as *Threat Modeling* by Microsoft. The findings revealed that those frameworks start their security consideration as early as requirement stage as depicted in Table 3.

Table 3. The stages focused in software secrurity testing framework. *(*√ = the software development stage that the model focus on, *Not stated* = the model is not clearly stated focus in this stage*)*

Software security testing framework	**Stage**	
	Requirement	**Design**
KAOS (Security Extension	✓	✓
MDS (Model Driven Security)	*Not stated*	✓
i* **framework**	✓	*Not stated*
ST (Secure Tropos)	✓	*Not stated*
SQUARE	✓	✓
SREP	✓	✓
SRE	✓	✓
TM (Threat Modeling)	✓	✓

As shown in Table 3, we scope our analysis into both requirement and design stages for two reasons. First, according on a number of research [11], to have an efficient cost, testing should start as early as possible in product development. Second, to produce a general framework and suitable in any domain, an early stage is prominence. Table 3 illustrates that in SSFT, security is a concern as early as during the requirement stage. However, the requirement elicitation activities need a guideline to present the software security vulnerabilities effectively. Software security vulnerabilities are caused by defective specification, design, and implementation. Unfortunately, common development practices leave software with much vulnerability. In order to have a secure cyber infrastructure, the supporting software must contain few, if any, vulnerabilities. This requires that software be built to sound security requirements. Therefore, we propose the utilization of ST activities into SSTF to provide the attributes needed in constructing security requirements. The details are discussed in the next section.

4 Discussion and Findings

This research shows that current analysis focuses three factors. First, the affected stage - the STF concerns with on internal users and has tendency to review the test discrepancy at coding or design stage via unit or integration test. On the other hand, the SSTF required knowledge of security attribute to elicit its requirement as early as at requirement stage. Second, the collection of security attribute - the knowledge of security attribute is acquired based on SSTF objectives; to find what should not happen in a system and to reveal any attack from intruders. These attributes sources are best collected during the test level activities (unit test, integration test and acceptance test). However, the reviewed STF limits the test conducted during unit and integration or system test only. Consequently, the actual result does not reflect the whole test levels carried out within a system. Third, the issues of complex system - the issue of emerged complex system suggest that a system to be secure. On the other hand, the current STF cover at least five objectives except for security (see section 3.1). Hence, there is a need for software security testing within the software development stage. As a result, in this paper, we recommend to utilize the acceptance test within the software development stage by integrating it with relevance unit and integration test result. This integration shall provide a comprehensive test result to support the testing process as early as from the requirement stage.

4.1 The Proposed Framework

Based on the findings, the three factors discussed previously are incorporate into the software development stages to formulate a generic SSTF as illustrated in Fig. 4. Fig. 4 describes a testing process is guided by a test objectives, which denoted by attributes such as completeness, interoperability and compatibility. During the testing, all type of test strategy (black box and white box or both) is conducted depending on domain and language used in code. The testing process is executing manually or automatable assisted by tools. The actual output is compared with the expected output derived from design specification [10]. Any comparison discrepancy is returned during the feedback stage as review process. It is noted that the derived expected output is bound to the design process; hence, any inappropriateness shall lead to an inappropriate comparison at the feedback stage. Referred in Fig.2, ST activities involve programmer, designer and business user. Any test levels that involve the authors as the tester, the comparison is done alike – tracing the expected output from design stage. The acceptance test which involves business user and actual environment reviewed its test discrepancy between requirement stage and testing stage.

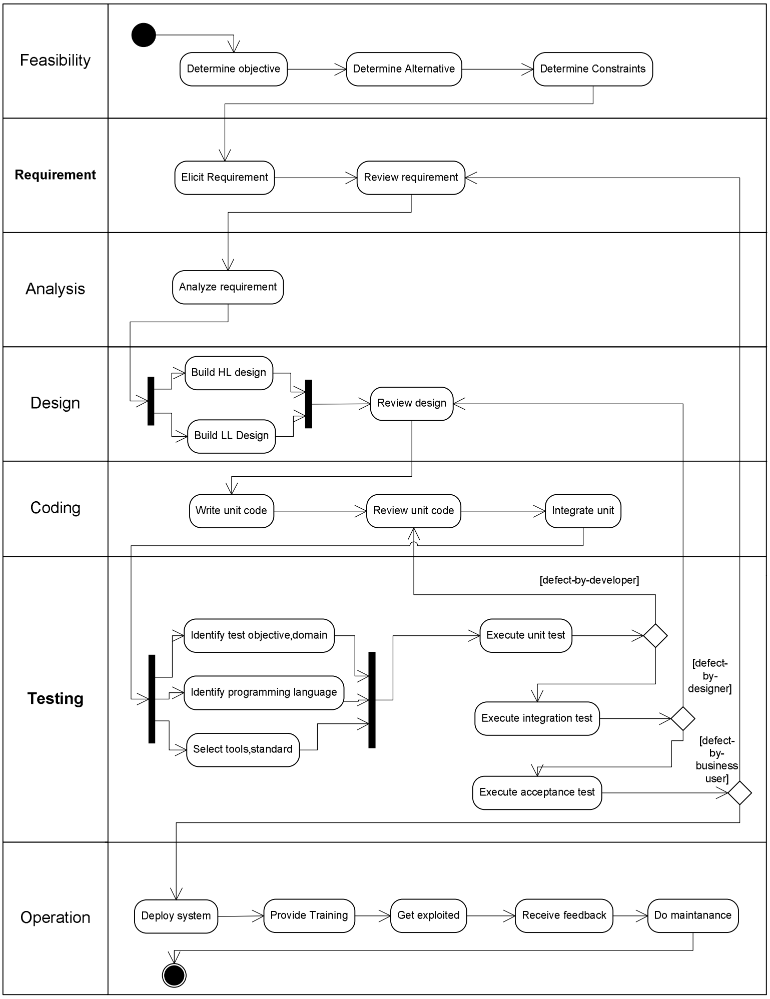


Fig. 4. Software Security Testing Framework. Any comparison discrepancy is returned during the feedback stage as review process. It is noted that the derived expected output is bound to the design process; hence, any inappropriateness shall lead to an inappropriate comparison at the feedback stage.

This research proposes the incorporation of the software security testing framework in software development by utilizing the software testing activities. The analysis proves that the software testing activities which focus on internal user (developer and designer) as tester has specific limit to overcome the test result discrepancy. Furthermore, the actual environment contributes to the existence of possible intruders which is important in security testing. As a result, the acceptance test which utilizes the business user and the actual environment is recommended to assist the test result discrepancy assessment. In other words, to preserve the test result consistency and completeness, the test result is best compared with the requirement stage.

In addition, the proposed framework consists of all stages in software development that are, feasibility, requirement, analysis, design, coding, testing and operation as summarized in Table 2. Therefore, the proposed framework can be used as a generic framework for security testing in any software development life cycle.

5 Conclusion and Future Work

In this paper, the software testing frameworks and software security testing frameworks are reviewed and analyzed. Based on the findings, this research proposed to fully utilize the acceptance testing in STF by incorporating it in SSTF. This incorporation is able to improve the security attribute needed during requirement stage of secure software development process. The acceptance test which exposed the system to real situation, including vulnerability, risk, impacts and the intruders shall provide a various set of security attribute to the requirement stage. Further improvement should be done in identifying the security attributes during acceptance test to generate a test pattern. This test pattern will further assist as a possible source in eliciting the security requirement during the requirement stage of software development. In addition, the proposed framework is expected to assist in tracing the web security attacks via a specific case study to generate a relevance testing pattern. The traceability process can be adapted in any other domain, such as in digital forensic investigation during collection of previous incidents data.

Acknowledgments. This research was kindly supported by Universiti Teknikal Malaysia Melaka under Short Grant Funding (PJP / 2009 / FTMK / (7E) S556) and Ministry of Higher Education Malaysia.

References

1. Thompson, H.H.: Why Security Testing Is Hard. J. Security & Privacy 1(4), 83–86 (2003)
2. Venter, H.S., Eloff, J.H.P., Li, Y.L.: Standardising Vulnerability Categories. J. Computers & Security 27(3-4), 71–83 (2008)
3. Jiwnani, K., Zelkowitz, M.: Maintaining Software With A Security Perspective. In: International Conference on Software Maintenance, pp. 194–203 (2002)
4. Cho, H.: Using Metaprogramming to Implement a Testing Framework. In: ACM SouthEast Regional Conference. ACM, USA (2009)
5. Misra, S.: An Empirical Framework For Choosing An Effective Testing Technique For Software Test Process Management. J. Information Technology Management 16(4), 19–26 (2005)
6. Royce, W.W.: Managing The Development of Large Software Systems. In: IEEE Western Electronic Show and Convention, pp. 1–9 (1970)
7. Rational Unified Process: Best Practices for Software Development Teams. Rational Software White Paper (2001)
8. Boehm, B., Brown, W., Turner, R.: Spiral Development Of Software-Intensive Systems Of Systems. In: 27th International Conference of Software Engineering (2005)
9. Ko, A.J., Myers, B.A.: A Framework And Methodology For Studying The Causes Of Software Errors In Programming Systems. J. Visual Languages & Computing 16(1-2), 41–84 (2005)

10. Mustafa, K., Khan, R.A.: Software Testing: Concepts and Practices. Alpha Science (2007)
11. Potter, B., McGraw, G.: Software Security Testing. J. Security & Privacy 2(5), 81–85 (2004)
12. Boehm, B.: A Spiral Model of Software Development and Enhancement. ACM SIGSOFT Software Engineering Notes 11(4), 14–24 (1986)
13. Craig, R.D., Jaskiel, S.P.: Systematic Software Testing. Artech House Publishers, Boston (2002)
14. Microsoft Security Development Lifecycle (SDL) Version 5.0, M. Library, Microsoft, `http://msdn.microsoft.com/en-us/library/cc307748.aspx`
15. Myers, G.J.: The Art of Software Testing. Wiley, New York (1979)
16. Tondel, I.A., Jaatun, M.G., Jensen, J.: Learning from Software Security Testing. In: 8th IEEE International Conference on Software Testing Verification and Validation Workshop, pp. 286–294. IEEE Computer Society, Washington (2008)
17. Pu-Lin, Y., Jin-Cherng, L.: Toward Precise Measurements Using Software Normalization. In: Proceedings of the 21st International Conference on Software Engineering, pp. 736–737. ACM, Los Angeles (1999)
18. Xu, L., Xu, B.: A Framework for Web Application Testing. In: International Conference on Cyberworlds, pp. 300–305. IEEE Computer Society, Washington (2004)
19. Jing, G., Yuqing, L.: Agent-based Distributed Automated Testing Executing Framework. In: International Conference on Computational Intelligence and Software Engineering, pp. 1–5. IEEE Press, Wuhan (2009)
20. Tsai, W.T., Wei, X., Chen, Y., Paul, R.: A Robust Testing Framework for Verifying Web Services by Completeness and Consistency Analysis. In: Proceedings of the IEEE International Workshop, pp. 159–166. IEEE Computer Society, Washington (2005)
21. Xie, T., Taneja, K., Kale, S., Marinov, D.: Towards a Framework for Differential Unit Testing of Object-Oriented Programs. In: 2nd International Workshop on Automation of Software Test. IEEE Computer Society, Minneapolis (2007)
22. Chen, R., Garde, S., Beale, T., Nystrom, M., Karlsson, D., Klein, G.O., Ahlfedlt, H.: An Archetype-based Testing Framework. J. Studies in Health Technology and Informatic 136, 401–406 (2008)
23. Tang, J., Lo, E.: A Lightweight Framework For Testing Database Applications. In: Symposium on Applied Computing. ACM, New Zealand (2010)
24. Lin, Y., Zhang, J., Gray, J.: A Testing Framework for Model Transformations. In: Model-Driven Software Development - Research and Practice in Software Engineering, pp. 219–236. Springer, Heidelberg (2005)
25. Werner, E., Grabowski, J., Troschutz, S., Zeiss, B.: A TTCN-3-based Web Service Test Framework. In: Software Engineering Workshops, pp. 375–382 (2008)
26. Villarroel, R., Fernández-Medina, E., Piattini, M.: Secure Information Systems Development - A Survey And Comparison. J. Computers & Security 24(4), 308–321 (2005)
27. Igure, V.M., Williams, R.D.: Taxonomies of Attacks and Vulnerabilities in Computer Systems. J. IEEE Communication Surveys & Tutorials 10(1), 6–19 (2008)
28. Maatta, J., Harkonen, J., Jokinen, T., Mottonen, M., Belt, P., Muhos, M., Haapasalo, H.: Managing Testing Activities In Telecommunications: A Case Study. J. Eng. Technol. Manage. 26, 73–96 (2009)
29. Lamsweerde, A.v., Brohez, S., Landtsheer, R.D., Janssens, D.: From System Goals to Intruder Anti-Goals: Attack Generation and Resolution for Security Requirements Engineering. In: Requirements for High Assurance Systems, pp. 49–56 (2003)

30. Basin, D., Doser, J., Lodderstedt, T.: Model Driven Security: From UML Models To Access Control Infrastructures. ACM Transactions on Software Engineering and Methodology (TOSEM) 15(1), 39–91 (2006)
31. Yu, E., Liu, L.: Modelling Trust In The i* Strategic Actors Framework. In: Proceedings of the 3rd Workshop on Deception, Fraud and Trust in Agent Societies. LNCS, pp. 175–194. Springer, London (2001)
32. Giorgini, P., Massacci, F., Mylopoulus, J., Zannone, N.: Modeling Security Requirements Through Ownership, Permission And Delegation. In: 13th IEEE International Conference on Requirements Engineering Proceedings, pp. 167–176. IEEE Computer Society, USA (2005)
33. Mead, N.R., Stehney, T.: Security Quality Requirements Engineering (SQUARE) Methodology. In: Proceedings of the 2005 Workshop On Software Engineering For Secure Systems- Building Trustworthy Applications, pp. 1–7. ACM, New York (2005)
34. Mellado, D., Fernández-Medina, E., Piattini, M.: A Common Criteria Based Security Requirements Engineering Process For The Development Of Secure Information Systems. Computer Standards & Interfaces 29(2), 244–253 (2007)
35. Haley, C.B., Laney, R., Moffett, J.D.: Security Requirements Engineering: A Framework for Representation and Analysis. IEEE Transactions on Software Engineering 34(1), 133–155 (2008)

A Performance Modeling Framework Incorporating Cost Efficient Deployment of Multiple Collaborating Instances

Razib Hayat Khan and Poul E. Heegaard

Norwegian University of Science & Technology
7491, Trondheim, Norway
{rkhan,poul.heegaard}@item.ntnu.no

Abstract. Performance evaluation of distributed system is always an intricate undertaking where system behavior is distributed among several components those are physically distributed. Bearing this concept in mind, we delineate a performance modeling framework for a distributed system that proposes a transformation process from high level UML notation to SRN model and solves the model for relevant performance metrics. To capture the system dynamics through our proposed framework we outline a specification style that focuses on UML collaboration and activity as reusable specification building blocks, while deployment diagram identify the physical components of the system and the assignment of software artifacts to identified system components. Optimal deployment mapping of software artifacts on the available physical resources of the system is investigated by deriving the cost function. Way to deal with parallel thread processing of the network nodes by defining the upper bound is precisely mentioned to generate the SRN model.

Keywords: UML, SRN, Performance attributes.

1 Introduction

Modeling phase plays an important role in the whole design process of the distributed system for qualitative and quantitative analysis. However in a distributed system, system behavior is normally distributed among several objects. The overall behavior of the system is composed of the partial behavior of the distributed objects of the system. So it is obvious to capture the behavior of the distributed objects for appropriate analysis to evaluate the performance related factors of the overall system. We therefore adopt UML collaboration and activity oriented approach as UML is the most widely used modeling language which models both the system requirements and qualitative behavior through different notations [2]. Collaboration and activity diagram are utilized to demonstrate the overall system behavior by defining both the structure of the partial object behavior as well as the interaction between them as reusable specification building blocks and later this UML specification style is applied to generate the SRN model by our proposed performance modeling framework. UML collaboration and activity provides a tremendous modeling framework containing several interesting properties. Firstly collaborations and

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 31–45, 2011.

activity model the concept of service provided by the system very nicely. They define structure of partial object behaviors, the collaboration roles and enable a precise definition of the overall system behavior. They also delineate the way to compose the services by means of collaboration uses and role bindings [1].

The proposed modeling framework considers system execution architecture to realize the deployment of the service components. Considering the system architecture to generate the performance model resolves the bottleneck of system performance by finding a better allocation of service components to the physical nodes. This needs for an efficient approach to deploy the service components on the available hosts of distributed environment to achieve preferably high performance and low cost levels. The most basic example in this regard is to choose better deployment architectures by considering only the latency of the service. The easiest way to satisfy the latency requirements is to indentify and deploy the service components that require the highest volume of interaction onto the same resource or to choose resources that are connected by links with sufficiently high capacity [3].

It is indispensable to extend the UML model to incorporate the performance-related quality of service (QoS) information to allow modeling and evaluating the properties of a system like throughput, utilization, and mean response time. So the UML models are annotated according to the *UML profile for MARTE: Modeling & Analysis of Real-Time Embedded Systems* to include quantitative system parameters [4].

We will focus on the stochastic reward net [5] as the performance model generated by our proposed framework due to its increasingly popular formalism for describing and analyzing systems, its modeling generality, its ability to capture complex system behavior concisely, its ability to preserve the original architecture of the system, to allow marking dependency firing rates & reward rates defined at the net level, to facilitate any modification according to the feedback from performance evaluation and the existence of analysis tools.

Several approaches have been followed to generate the performance model from system design specification. However, most existing approaches [6] [7] [8] do not highlight more on the issue that how to optimally conduct the system modeling and performance evaluation. The framework presented here is the first known approach that introduces a new specification style utilizing UML behavioral diagrams as reusable specification building block which is later used for generating performance model to produce performance prediction result at early stage of the system development process. Building blocks describe the local behavior of several components and the interaction between them. This provides the advantage of reusability of building blocks, since solution that requires the cooperation of several components may be reused within one self-contained, encapsulated building block. In addition the resulting deployment mapping provided by our framework has great impact with respect to QoS provided by the system. Our aim here is to deal with vector of QoS properties rather than restricting it in one dimension. Our presented deployment logic is surely able to handle any properties of the service, as long as we can provide a cost function for the specific property. The cost function defined here is flexible enough to keep pace with the changing size of search space of available host in the execution environment to ensure an efficient deployment of service components. Furthermore we aim to be able to aid the deployment of several different services at the same time using the same proposed framework. The novelty of our approach also reflected in showing the optimality of our solution with respect to both deployment logic and evaluation of performance metrics.

The objective of the paper is to provide an extensive performance modeling framework that provides a translation process to generate SRN performance model from system design specification captured by the UML behavioral diagram and later solves the model for relevant performance metrics. To incorporate the cost function to draw relation between service component and available physical resources permit us to identify an efficient deployment mapping in a fully distributed manner. The way to deal with parallel thread processing of the network node by defining the upper bound is precisely mentioned while generating the SRN model through the proposed framework. The work presented here is the extension of our previous work described in [9] [10] [14] where we presented our proposed framework with respect to the execution of single and multiple collaborative sessions and considered alternatives system architecture candidate to describe the system behavior and evaluate the performance factors. The paper is organized as follows: section 2 introduces our proposed performance modeling framework, section 3 demonstrates the application example to show the applicability of our modeling framework, section 4 delineates conclusion.

2 Proposed Performance Modeling Framework

The proposed framework is composed of 6 steps shown in Fig.1 where steps 1 and 2 are the parts of Arctis tool suite [11].

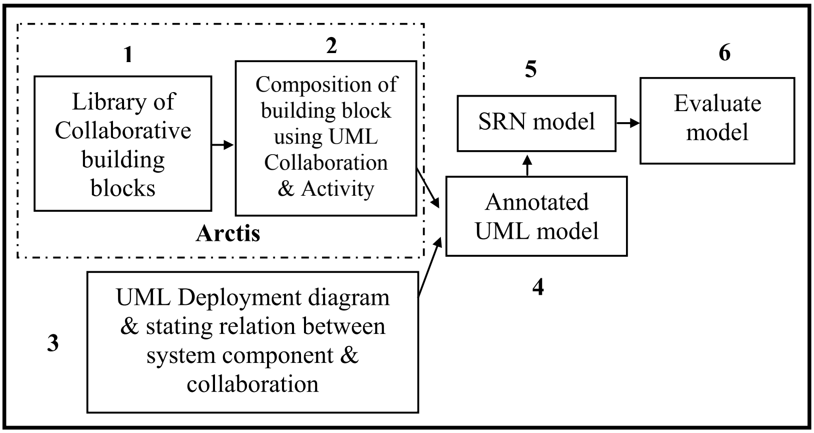

1
Library of Collaborative building blocks
2
Composition of building block using UML Collaboration & Activity
Arctis
5
SRN model
6
Evaluate model
Annotated UML model
4
3
UML Deployment diagram & stating relation between system component & collaboration

Fig. 1. Proposed performance modeling framework

Arctis focuses on the abstract, reusable service specifications that are composed form UML 2.2 collaborations and activities. It uses collaborative building blocks as reusable specification units to create comprehensive services through composition. To support the construction of building block consisting of collaborations and activities, Arctis offers special actions and wizards. In addition a number of inspections ensure the syntactic consistency of building blocks. A developer first consults a library to check if an already existing collaboration block or a collaboration of several blocks solves a certain task. Missing blocks can also be created from scratch and stored in the library for later reuse. The building blocks are expressed as UML models. The structural aspect, for example the service component and their multiplicity, is expressed by means of UML 2.2 collaborations. For the detailed internal behavior, UML 2.2 activities have been used. They express the local behavior of each of the service components as well as their necessary interactions in a compact and self-contained way using explicit control flows [11]. Moreover the building blocks are combined into more comprehensive service by composition. For this composition, Arctis uses UML 2.2 collaborations and activities as well. While collaborations provide a good overview of the structural aspect of the composition, i.e., which sub-services are reused and how their collaboration roles are bound, activities express the detailed coupling of their respective behaviors [11]. The steps are illustrated below:

Step 1: Construction of collaborative building block: The proposed framework utilizes collaboration as main specification units. The specifications for collaborations are given as coherent, self-contained reusable building blocks. The structure of the building block is described by UML 2.2 collaboration. The building block declares the participants (as collaboration roles) and connection between them. The internal behavior of building block is described by UML activity. It is declared as the classifier behavior of the collaboration and has one activity partition for each collaboration role in the structural description. For each collaboration use, the activity declares a corresponding call behavior action refereeing to the activities of the employed building blocks. For example, the general structure of the building block *t* is given in Fig. 2(a) where it only declares the participants A and B as collaboration roles and the connection between them is defined as collaboration use t_x (x=1…n$_{\mathrm{AB}}$ (number of collaborations between collaboration roles A & B)). The internal behavior of the same building block is shown in Fig. 2(b). The activity *transfer*$_{ij}$ (*where ij = AB*) describes the behavior of the corresponding collaboration. It has one activity partition for each collaboration role: A and B. Activities base their semantics on token flow [1]. The activity starts by placing a token when there arrives a response (indicated by the streaming pin *res*) to transfer by either participant A or B.

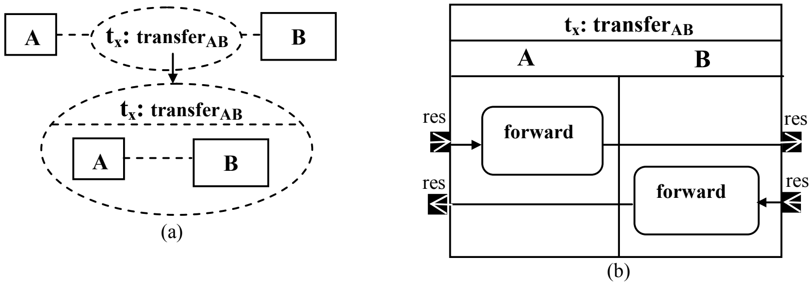


Fig. 2. (a) Structure of the building block using collaboration diagram (b) behavior of the building block using activity diagram

After completion of the processing by the collaboration role A and B the token is transferred from the participant A to participant B and from participant B to Participant A which is represented by the call behavior action *forward.*

Step 2: Composition of building block using UML collaboration & activity: To generate the performance model, the structural information about how the collaborations are composed is not sufficient. It is necessary to specify the detailed behavior of how the different events of collaborations are composed so that the desired overall system behavior can be obtained. For the composition, UML collaborations and activities are used complementary to each other; UML collaborations focus on the role binding and structural aspect, while UML activities complement this by covering also the behavioral aspect for composition. For this purpose, call behavior actions are used. Each sub-service is represented by call behavior action referring the respective activity of building blocks. Each call behavior action represents an instance of a building block. For each activity parameter node of the referred activity, a call behavior action declares a corresponding pin. Pins have the same symbol as activity parameter nodes to represent them on the frame of a call behavior action. Arbitrary logic between pins may be used to synchronize the building block events and transfer data between them. By connecting the individual input and output pins of the call behavior actions, the events occurring in different collaborations can be coupled with each other. Semantics of the different kinds of pins are given in more detailed in [1]. For example the detailed behavior and composition of the collaboration is given in following Fig. 3. The initial node(●) indicates the starting of the activity. The activity is started at the same time from each participant. After being activated, each participant starts its processing of the request which is mentioned by call behavior action P_i (Processing$_i$, where i = A, B & C). Completions of the processing by the participants are mentioned by the call behavior action d_i (Processing_done$_i$, i = A, B & C). After completion of the processing, the responses are delivered to the corresponding participants indicated by the streaming pin *res*. When the processing of the execution of the task by the participant B completes the result is passed through a decision node *k* and only one flow is activated at the certain time instance. The response of the collaboration role A and C are forwarded to B and the response of collaboration role B is forwarded to either A or C which is mentioned by collaboration *t: transfer$_{ij}$* (where ij = AB or BC). The completion of the activity of each participant is shown by the ending node (◉)

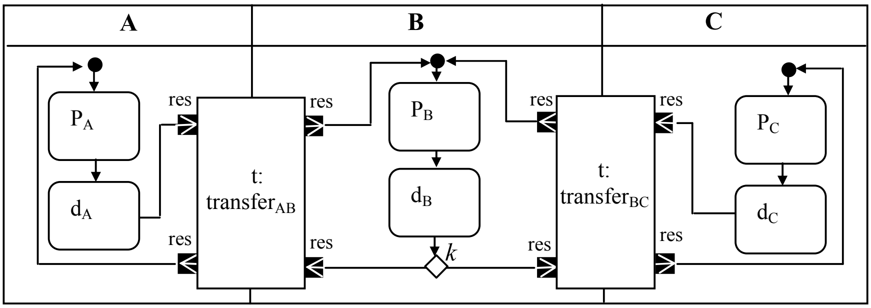


Fig. 3. System activity to couple the collaboration

Step 3: Designing UML deployment diagram & stating relation between systemcomponents & collaborations: We model the system as collection of N interconnected nodes. Our objective is to find a deployment mapping for execution environment for a set of service components C available for deployment that comprises service. Deployment mapping can be defined as M: C → N between a numbers of service components instances c, onto nodes n. Components can communicate via a set of collaborations. We consider four types of requirements in the deployment problem. Components have execution costs, collaborations have communication costs and costs for running of background process and some of the components can be restricted in the deployment mapping to specific nodes which are called bound components. Furthermore, we consider identical nodes that are interconnected in a full-mesh and are capable of hosting components with unlimited processing demand. We observe the processing load that nodes impose while host the components and also the target balancing of load between the nodes available in the network. By balancing the load the deviation from the global average per node execution cost will be minimized. Communication costs are considered if collaboration between two components happens remotely, i.e. it happens between two nodes [3]. In other words, if two components are placed onto the same node the communication cost between them will not be considered. The cost for executing the background process for conducting the communication between the collaboration roles is always considerable no matter whether the collaboration roles deploy on the same or different nodes. Using the above specified input, the deployment logic provides an optimal deployment architecture taking into account the QoS requirements for the components providing the specified services. We then define the objective of the deployment logic as obtaining an efficient (low-cost, if possible optimum) mapping of component onto the nodes that satisfies the requirements in reasonable time. The deployment logic providing optimal deployment architecture is guided by the cost function F (M). The evaluation of cost function F (M) is mainly influenced by our way of service definition. Service is defined in our approach as a collaboration of total E components labeled as $\mathrm{c_i}$ (where i = 1…. E) to be deployed and total K collaboration between them labeled as $\mathrm{k_j}$, (where j = 1 … K). The execution cost of each service component can be labeled as $f\mathrm{c_i}$; the communication cost between the service components is labeled as f_{kj} and the cost for executing the background process for conducting the communication between the service components is labeled as f_{Bj}. Accordingly we only observe the total load ($\overset{\frown}{l}_N$, n = 1…N) of a given deployment mapping at each node. We will strive for an optimal solution of equally distributed load among the processing nodes and the lowest cost possible, while taking into account the execution cost $f\mathrm{c_i}$, i = 1….E, communication cost f_{kj}, j = 1….K and cost for executing the background process f_{Bj}, j = 1….k. $f\mathrm{c_i}$, f_{kj} and f_{Bj} are derived from the service specification, thus the offered execution load can be calculated as $\sum_{i=1}^{E} f\mathrm{c_i}$ .This way, the logic can be aware of the target load [3]:

$$\mathrm{T} = \frac{\sum_{i=1}^{E} f\mathrm{c_i}}{\mathrm{N}} \tag{1}$$

Given a mapping M = {$\mathrm{m_n}$} (where m_n is the set of components at node n & $n \in {}_{\mathrm{N}}$) the total load can be obtained as $\widehat{l}_n = \sum_{\mathrm{c_i} \in \mathrm{m_n}} f_{\mathrm{c_i}}$. Furthermore the overall cost function F (M) becomes (where $\mathrm{I_j}$ = 1, if $\mathrm{k_j}$ external or 0 if $\mathrm{k_j}$ internal to a node):

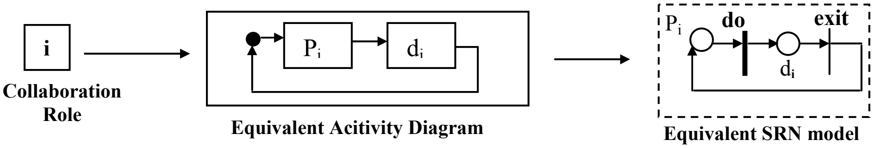
$$\mathrm{F\ (M)} = \sum_{n=1}^{N} \left| \widehat{l}_n - \mathrm{T} \right| + \sum_{j=1}^{K} \left(\mathrm{I_j} f_{\mathrm{kj}} + f_{\mathrm{Bj}} \right) \qquad (2)$$

Step 4: Annotating the UML model: Performance information is incorporated into the UML activity diagram and deployment diagram according to *UML profile for MARTE: Modeling & Analysis of Real-Time Embedded Systems* [4] for evaluating system performance by performance model solver.

Step 5: Deriving the SRN model: To generate the SRN model of the system, first we generate the SRN model of the individual system components and later compose them together to generate the system level SRN model. The rules are based on decomposition of UML collaboration, activity and deployment diagram into basic elements of SRN model like states as places, timed transition and immediate transition. In addition the rules are based on the rendezvous synchronization that means when communication between two processes of two interconnected nodes occur it follows the rendezvous synchronization [12]. The rules are following:

SRN model of the collaboration role of a reusable building block is mentioned by the 6-tuple {**Φ, T, A, K, N, m₀**} in the following way [5]: **Φ** = Finite set of the places (drawn as circles), derived from the call behavior action of the collaboration role; **T** = Finite set of the transition (drawn as bars), derived from the annotated UML activity diagram that denotes system's behavior; **A** $\subseteq$ {Φ × T} $\cup$ {T × Φ} is a set of arcs connecting Φ and T; K: T → {Timed (time>0, drawn as solid bar), Immediate (time = 0, drawn as thin bar)} specifies the type of the each transition, derived from the annotated UML activity diagram that denotes system's behavior; **N**: A→ {1, 2, 3…} is the multiplicity associated with the arcs in A; **m**: Φ → {0, 1, 2...} is the marking that denotes the number of tokens for each place in Φ. The initial marking is denoted as $\mathbf{m_0}$.

Rule1: The SRN model of the collaboration role of a reusable building block is represented by the 6-tuple in the following way: $\mathbf{\Phi_i}$ = {$\mathrm{P_{i,}}$ $\mathrm{d_i}$}; **T** = {do, exit}; **A** = {{($\mathrm{P_i}$ × do) $\cup$ (do × $\mathrm{d_i}$)}, {($\mathrm{d_i}$ × exit) $\cup$ (exit × $\mathrm{P_i}$)}}; **K** = (do → Timed, exit → Immediate); **N** = {($\mathrm{P_i}$ × do) →1, (do × $\mathrm{d_i}$) →1, ($\mathrm{d_i}$ × exit) →1, (exit × $\mathrm{P_i}$)→1}; $\mathbf{m_o}$ = {($\mathrm{P_i}$→1}, ($\mathrm{d_i}$ →0)}. Here places are represented by $\mathrm{P_i}$ and $\mathrm{d_i}$. Transitions are represented by *do* and *exit* where *do* is a timed transition and *exit* is an immediate transition. Initially place $\mathrm{P_i}$ contains one token and place $\mathrm{d_i}$ contains no token. SRN model of the collaboration role is graphically represented by the following way:

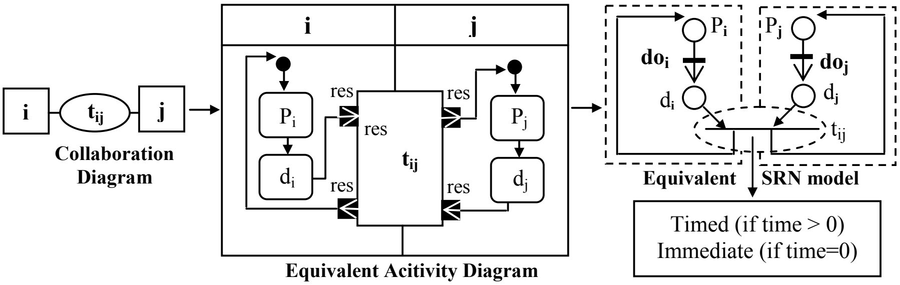


Collaboration Role

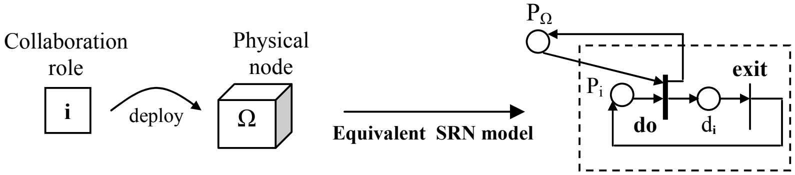
Equivalent Acitivity Diagram

Equivalent SRN model

Rule2: The SRN model of a collaboration where collaboration connects only two collaboration roles is represented by the 6-tuple in the following way: $\mathbf{\Phi} = \{\Phi_{i,} \Phi_j\} = \{P_i, d_i, P_j, d_j\}$; $\mathbf{T} = \{do_i, do_j, t_{ij}\}$; $\mathbf{A} = \{\{(P_i \times do_i) \cup (do_i \times d_i)\}, \{(d_i \times t_{ij}) \cup (t_{ij} \times P_i)\}, \{(P_j \times do_j) \cup (do_j \times d_j)\} \{(d_j \times t_{ij}) \cup (t_{ij} \times P_j)\}\}$; $\mathbf{K} = (do_i \rightarrow \text{Timed}, do_j \rightarrow \text{Timed}, t_{ij} \rightarrow \text{Timed} \mid \text{Immediate})$; $\mathbf{N} = \{(P_i \times do_i) \rightarrow 1, (do_i \times d_i) \rightarrow 1, (d_i \times t_{ij}) \rightarrow 1, (t_{ij} \times P_i) \rightarrow 1, \{\{(P_j \times do_j) \rightarrow 1, (do_j \times d_j) \rightarrow 1, (d_j \times t_{ij}) \rightarrow 1, (t_{ij} \times P_j) \rightarrow 1\}$; $\mathbf{m_o} = \{(P_i \rightarrow 1, d_i \rightarrow 0, P_j \rightarrow 1, d_j \rightarrow 0\}$. Here places are represented by P_i, d_i, P_j, d_j, transitions are represented by do_i, do_j and t_{ij} where do_i and do_j are timed transition and t_{ij} is a timed transition if the two collaboration roles deploy on the different physical node (communication time > 0) or immediate transition if the two collaboration roles deploy on the same physical node (communication time = 0). Initially place P_i and P_j contains one token and place d_i and d_j contains no token. SRN model of the collaboration is graphically represented in the following way:



Rule3: When the collaboration role of a reusable building block deploys onto a physical node the equivalent SRN model is represented by 6-tuple in following way: $\mathbf{\Phi_i} = \{P_{i,} d_i, P_\Omega\}$; $\mathbf{T} = \{do, exit\}$; $\mathbf{A} = \{\{(P_i \times do) \cup (do \times d_i)\}, \{(P_\Omega \times do) \cup (do \times P_\Omega)\}, \{(d_i \times exit) \cup (exit \times P_i)\}\}$; $\mathbf{K} = (do \rightarrow \text{Timed}, exit \rightarrow \text{Immediate})$; $\mathbf{N} = \{(P_i \times do) \rightarrow 1, (do \times d_i) \rightarrow 1, (P_\Omega \times do) \rightarrow 1, (do \times P_\Omega) \rightarrow 1 (d_i \times exit) \rightarrow 1, (exit \times P_i) \rightarrow 1\}$; $\mathbf{m_o} = \{(P_i \rightarrow 1\}, (d_i \rightarrow 0), (P_\Omega \rightarrow q)\}$. Here places are represented by P_i, d_i and P_Ω. Transitions are represented by *do* and *exit* where *do* is a timed transition and *exit* is an immediate transition. Initially place P_i contains one token, place d_i contains no token and place P_Ω contains **q** tokens which define the upper bound of the execution of the threads in parallel by the physical node Ω and the timed transition *do* will fire only when there is a token available in both the place Pi and P_Ω. The place P_Ω will again get back it's token after firing of the timed transition *do* indicating that the node is ready to execute incoming threads. SRN model of the collaboration role is graphically represented by the following way:



Step 6: Evaluate the model: We focus on measuring the throughput of the system from the developed SRN model. Before deriving formula for throughput estimation we consider several assumptions. Firstly if more than one service component deploy on a network node the processing of all the components will be done in parallel at the same time and the processing power of the network node will be utilized among the multiple threads to complete the parallel processing of that node. There must be an upper bound of the execution of parallel threads by a network node. Secondly when communication between two processes of two interconnected nodes occur it follows the rendezvous synchronization. Moreover all the communications among the interconnected nodes occur in parallel. Finally the communications between interconnected nodes will be started following the completion of all the processing inside each physical node. By considering the all the assumption we define the throughput as function of total expected number of jobs, E (N) and cost of the network, C_Net. The value of E (N) is calculated by solving the SRN model using SHARPE [15]. The value of C_Net is evaluated by considering a subnet which is performance limiting factor of the whole network i.e., which posses maximum cost with respect to its own execution cost, communication cost with other subnet and cost for running background processes. Assume cost of the network, C_Net is defined as follows (where $f c_m$ = execution cost of the mth component of subnet$_{i;}$ c_subnet$_i$ = cost of the ith subnet where i = 1…n that comprises the whole network and $I_j = 0$ in this case as k_j internal to a node):

$$c\_subnet_i = \max\{f c_m + I_j f k_j + f B_j\} = \max\{f c_m + f B_j\}; \qquad (3)$$

Now we evaluate the cost between each pair of subnet (sbunet$_i$ & subnet$_j$; where $i \neq j$) with respect to the subnet's own processing cost, cost for running background process and the cost associated with the communication with other subnet in the network. Cost between subnet$_i$ and subnet$_j$, C_subnet$_{i,j}$ is defined as (where $f k_{i,j}$ = communication cost between subnet$_i$ & subnet$_j$ and $I_{i,j} = 1$ as $k_{i,j}$ external to a node):

$$c\_subnet_{i,j} = \max\{\max\{c\_subnet_i, c\_subnet_j\} + I_{i,j} f k_{i,j} + f B_{i,j}\}; \qquad (4)$$

$$C\_Net = \max\{c\_subnet_{i,j}\}; \qquad (5)$$

$$\therefore \text{Throughput} = \frac{\mathbf{E(N)}}{\mathbf{C\_Net}} \qquad (6)$$

3 Application Example

As a representative example, we consider the scenario originally from Efe dealing with heuristically clustering of modules and assignment of clusters to nodes [13]. This scenario is sufficiently complex to show the applicability of our proposed framework. The problem is defined in our approach as a service of collaboration of E = 10 components or collaboration role (labeled C_1 . . . C_{10}) to be deployed and K = 14 collaborations between them depicted in Fig. 4. We consider four types of requirements in this specification. Besides the execution cost, communication costs and cost for running background process, we have a restriction on components C_2, C_7, C_9 regarding their location. They must be bound to nodes $\boldsymbol{n_2}$, $\boldsymbol{n_1}$, $\boldsymbol{n_3}$, respectively.

The internal behavior of the collaboration *K* of our example scenario is realized by the call behavior action through UML activity like structure already mentioned in Fig. 2(b). The composition of the collaboration role *C* is realized through UML activity diagram shown in Fig. 5. The initial node (●) indicates the starting of the activity. The activity is started at the same time from the entire participants C_1 to C_{10}. After being activated, each participant starts its processing of request which is mentioned by call behavior action P_i (Processing of the i^th^ service component). Completions of the processing by the participants are mentioned by the call behavior action d_i*(*Processing done of the i^th^ service component)*).* After completion of the processing, the responses are delivered to the corresponding participants indicated by the streaming pin *res*. When any participant is associated with more than one participant through collaborations the result of the processing of that participant is passed through a decision node and only one flow is activated at the certain time instance. For example after completion of the processing of participant C_2 the response will be passed through the decision node X_2 and only one flow (flow towards C_1 or C_3 or C_5) will be activated. The completion of the processing of the each participant is shown by ending node (◉.)

In this example, the target environment consists only of N = 3 identical, interconnected nodes with a single provided property, namely processing power and

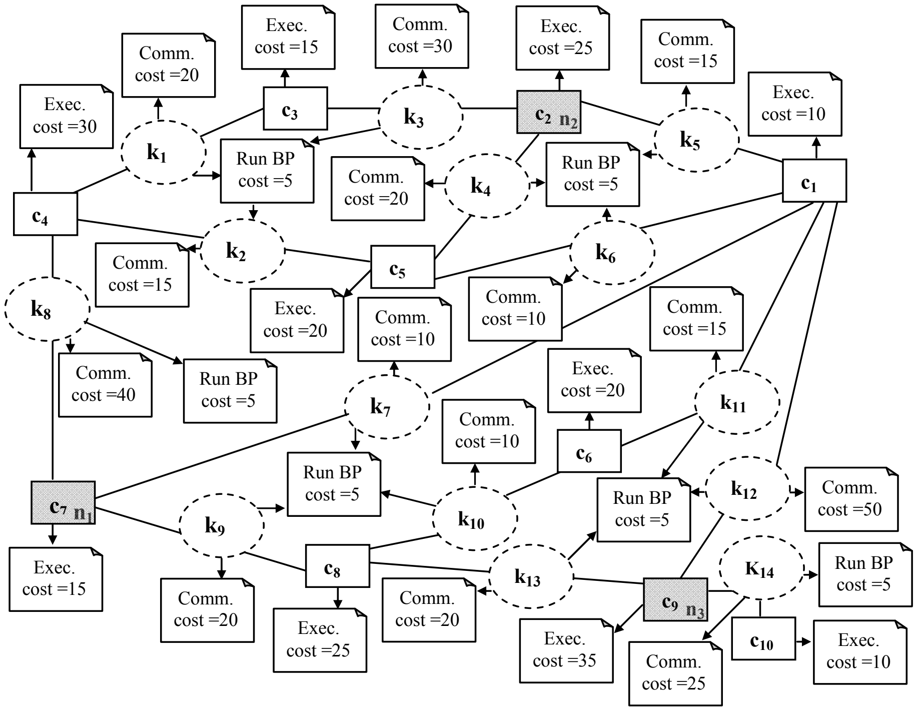


Fig. 4. Collaborations and components in the example scenario

with infinite communication capacities depicted in Fig. 6 (a). The optimal deployment mapping can be observed in Table. 1. The lowest possible deployment cost, according to (2) is 17 + (270 − 100) = 187.

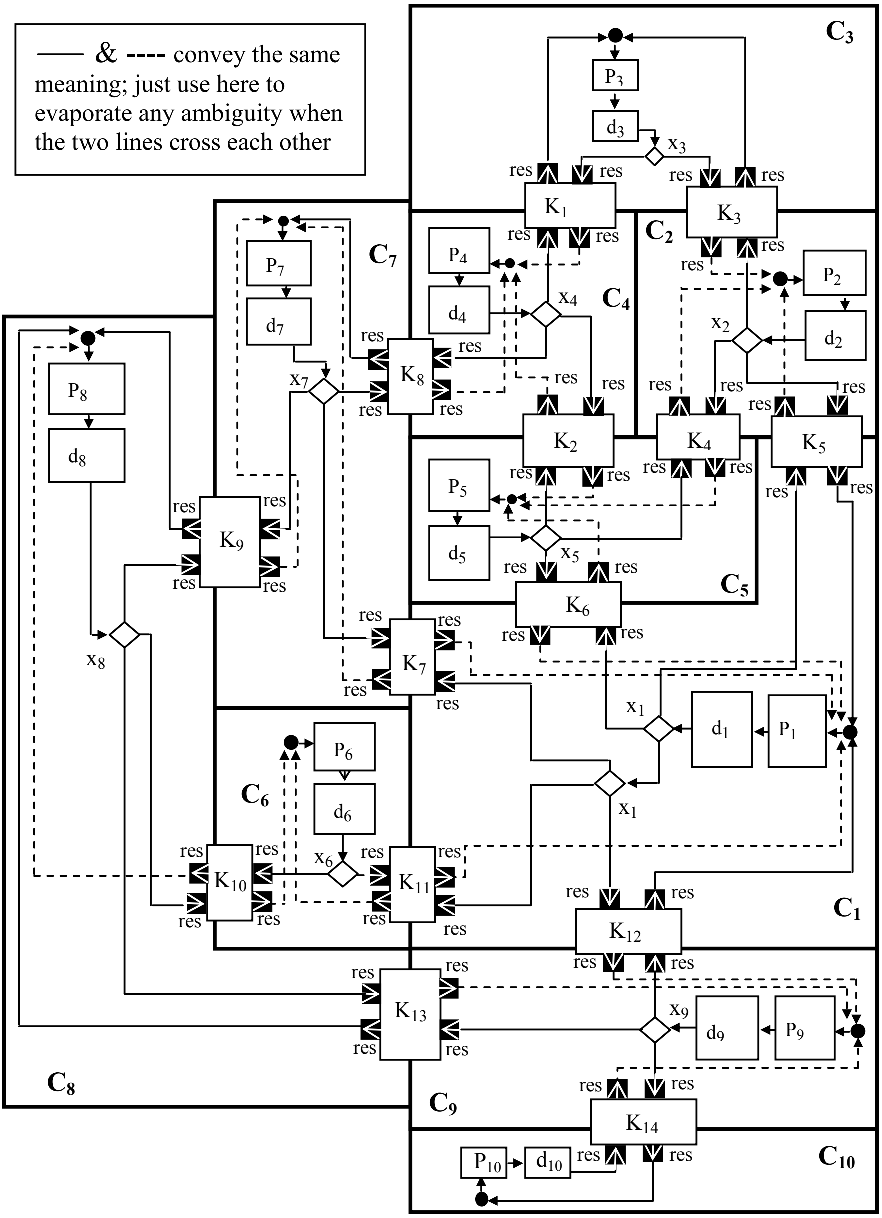


Fig. 5. Detail behavior of the event of the collaboration using activity for our example scenario

To annotate the UML diagram in Fig. 5 & 6(a) we use the stereotype *saStep computingResource, scheduler* and the tag value *execTime, deadline* and *schedPolicy* [4]. *saStep* is a kind of step that begins and ends when decisions about the allocation of system resources are made. The duration of the execution time is mentioned by the tag value *execTime* which is the average time in our case. *deadline* defines the maximum time bound on the completion of the particular execution segment that must me met. A Scheduler is defined as a kind of ResourceBroker that brings access to its brokered ProcessingResource or resources following a certain scheduling policy tagged by *schedPolicy.* Collaboration K_i is associated with two instances of *deadline* (Fig. 6(b)) as collaborations in example scenario are associated with two kinds of cost: communication cost & cost for running background process.

By considering the above deployment mapping and the transformation rule the analogous SRN model of our example scenario is depicted in Fig. 7. The states of the SRN model are derived from the call behavior action of the corresponding collaboration role and collaboration among them. While generating the SRN model of the system if more than one service component deploy on a network node the processing of all the components will be done in parallel at the same time and the processing power of the network node will be utilized among the multiple threads to complete the parallel processing of that node. This can be achieved through marking dependency firing rate defined as the following way in SRN model:

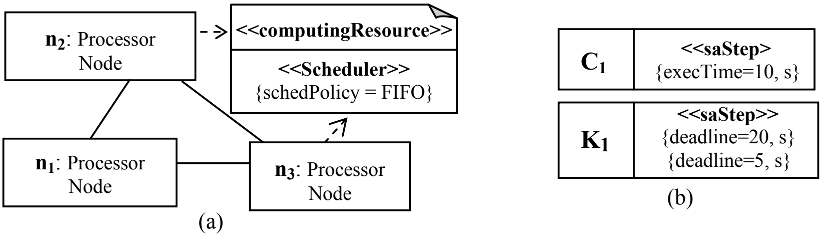


Fig. 6. (a)The target network of hosts (b) annotated UML model using MARTE profile

Table 1. Optimal deployment mapping in the example scenario

Node	Components	$\widehat{l}_n$	$\vert \widehat{l}_n - T \vert$	Internal collaborations
n_1	c_4, c_7, c_8	70	2	k_8, k_9
n_2	c_2, c_3, c_5	60	8	k_3, k_4
n_3	c_1, c_6, c_9, c_{10}	75	7	k_{11}, k_{12}, k_{14}
$\sum$ cost			17	100

$$\lambda_i / \sum_{i=1}^{n} (\# (P_i)) \qquad (7)$$

Where λ_i = processing rate of the ith service component deploys in a network node and i=1…n that means total n service components deploy on a network node. (# (P_i)) returns the number of tokens in the place P_i.

Initially there will be a token in the place p_1 to p_{10}. For generating the SRN model firstly we will consider the collaboration roles deploy on the processor node $\boldsymbol{n_1}$ which are C_4, C_7 & C_8. Here components C_7 are connected with C_4 and C_8. The communication cost between the components is zero but there is still some cost for execution of the background process. So according to rule 2, after the completion of the state transition from p_7 to d_7 (states of component C_7), from p_4 to d_4 (states of component C_4) and from p_8 to d_8 (states of component C_8) the states d_7, d_4 and d_7, d_8 are connected by the timed transition k_8 and k_9 to generate the SRN model. Collaboration roles C_2, C_3 & C_5 deploy on the processor node $\boldsymbol{n_2}$. Likewise after the completion of the state transition from p_2 to d_2 (states of component C_2), from p_3 to d_3 (states of component C_3) and from p_5 to d_5 (states of component C_5) the states d_2, d_3 and d_2, d_5 are connected by the timed transition k_3 and k_4 to generate the SRN model according to rule 2. Collaboration roles C_6, C_1, C_9 & C_{10} deploy on the processor node $\boldsymbol{n_3}$. In the same way after the completion of the state transition from p_1 to d_1 (states of component C_1), from p_6 to d_6 (states of component C_6), p_9 to d_9 (states of component C_9) and from p_{10} to d_{10} (states of component C_{10}) the states d_1, d_6; d_1, d_9 and d_9, d_{10} are connected by the timed transition k_{11}, k_{12} and K_{14} to generate the SRN model following rule 2. To generate the system level SRN model we need to combine the entire three SRN model generated for three processor nodes by considering the interconnection among them. To compose the SRN models of processor node $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$, states d_4 and d_3 connect by the timed transition k_1 and states d_4 and d_5 connect by the timed transition k_2 according to rule 2. Likewise to compose the SRN models of processor node $\boldsymbol{n_2}$ and $\boldsymbol{n_3}$, states d_2 and d_1 connect by the timed transition k_5 and states d_5 and d_1 connect by the timed transition k_6 according to rule 2.

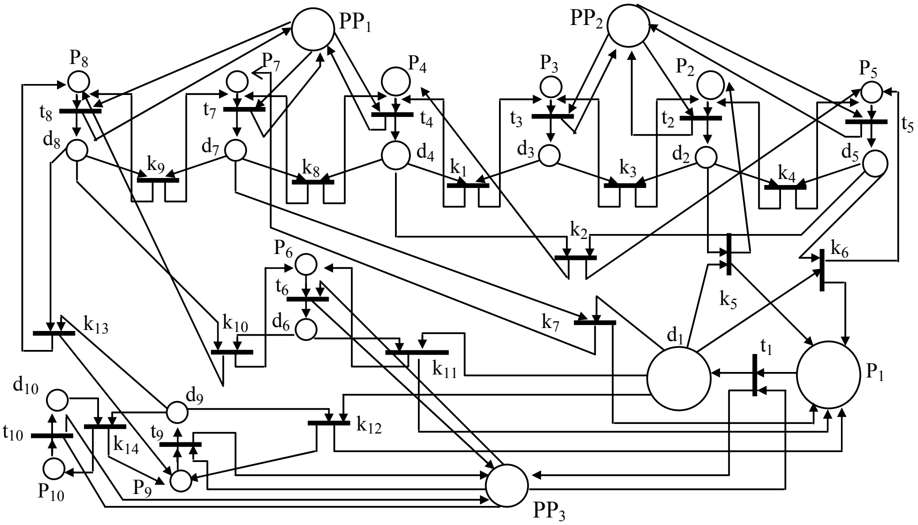


Fig. 7. SRN model of our example scenario

To compose the SRN models of processor node $\boldsymbol{n_1}$ and $\boldsymbol{n_3}$, states d_7 and d_1 connect by the timed transition k_7, states d_8 and d_6 connect by the timed transition k_{10} and states d_8 and d_9 connect by the timed transition k_{13} according to rule 2. By the above way the system level SRN model is derived. According to rule 3, To define the upper bound of the execution of parallel threads by a network node we introduce three places PP_1, PP_2 and PP_3 in the SRN model for the three network nodes and initially these three places will contain $\boldsymbol{q}$ ($\boldsymbol{q}$ = 1, 2, 3,…….) tokens where $\boldsymbol{q}$ will define the maximum number of the threads that will be handled by a network node at the same time. To ensure the upper bound of the parallel processing of a network node $\boldsymbol{n_1}$ we introduce arcs from place PP_1 to transition t_4, t_7 and t_8. That means components C_4, C_7 and C_8 can start their processing if there is token available in place PP_1 as the firing of transitions t_4, t_7 and t_8 not only depend on the availability of the token in the place p_4, p_7 and p_8 but also depend on the availability of the token in the place PP_1. Likewise to ensure the upper bound of the parallel processing of a network node $\boldsymbol{n_2}$ and $\mathbf{n_3}$ we introduce arcs from place PP_2 to transition t_2, t_3 and t_5 and from place PP_3 to transition t_1, t_6, t_9, t_{10}.

The throughput calculation according to (6) for the different deployment mapping including the optimal deployment mapping is shown in Table. 2. The throughput is 0.107s^{-1} while considers the optimal deployment mapping where E (N) = 6.96 (calculated using SHARPE [15]) and C_Net = 65s. The optimal deployment mapping presented in Table 1 also ensures the optimality in case of throughput calculation. We present here the throughput calculation of some of the deployment mappings of the software artifacts but obviously the approach presented here confirms the efficiency in both deployment mapping and throughput calculation for all the cases.

Table 2. Optimal deployment mapping in the example scenario

Node	Components	Possible cost	Throughput
$\{n_1, n_2, n_3\}$	$\{\{c_4, c_7, c_8\}, \{c_2, c_3, c_5\}, \{c_1, c_6, c_9, c_{10}\}\}$	187	0.107
$\{n_1, n_2, n_3\}$	$\{\{c_4, c_6, c_7, c_8\}, \{c_2, c_3, c_5\}, \{c_1, c_9, c_{10}\}\}$	218	0.106
$\{n_1, n_2, n_3\}$	$\{\{c_4, c_7\}, \{c_2, c_3, c_5, c_6,\}, \{c_1, c_8, c_9, c_{10}\}\}$	232	0.102
$\{n_1, n_2, n_3\}$	$\{\{c_5, c_7, c_8\}, \{c_2, c_3, c_4\}, \{c_1, c_6, c_9, c_{10}\}\}$	227	0.086
$\{n_1, n_2, n_3\}$	$\{\{c_3, c_7, c_8\}, \{c_2, c_4, c_5\}, \{c_1, c_6, c_9, c_{10}\}\}$	252	0.084
$\{n_1, n_2, n_3\}$	$\{\{c_1, c_6, c_7, c_8\}, \{c_2, c_3, c_5\}, \{c_4, c_9, c_{10}\}\}$	257	0.083
$\{n_1, n_2, n_3\}$	$\{\{c_1, c_6, c_7, c_8\}, \{c_2, c_3, c_4\}, \{c_5, c_9, c_{10}\}\}$	247	0.075
$\{n_1, n_2, n_3\}$	$\{\{c_4, c_7, c_8\}, \{c_1, c_2, c_3, c_5\}, \{c_6, c_9, c_{10}\}\}$	217	0.073
$\{n_1, n_2, n_3\}$	$\{\{c_3, c_6, c_7, c_8\}, \{c_1, c_2, c_4, c_5\}, \{c_9, c_{10}\}\}$	302	0.072
$\{n_1, n_2, n_3\}$	$\{\{c_6, c_7, c_8\}, \{c_1, c_2, c_4, c_5\}, \{c_3, c_9, c_{10}\}\}$	288	0.071

4 Conclusion

We present a novel approach for model based performance evaluation of distributed system which spans from capturing the system dynamics through UML diagram as reusable building block to efficient deployment of service components in a distributed manner by capturing the QoS requirements. System dynamics is captured through UML collaboration and activity oriented approach. Furthermore, quantitative analysis of the system is achieved by generating SRN performance model from the UML specification style. The transformation from UML diagram to corresponding SRN elements like states, different pseudostates and transitions is proposed. Performance related QoS information is taken into account and included in the SRN model with equivalent timing and probabilistic assumption for enabling the evaluation of performance prediction result of the system at the early stage of the system development process. In addition, the logic, as it is presented here, is applied to provide the optimal, initial mapping of components to hosts, i.e. the network is considered rather static. However, our eventual goal is to develop support for run-time redeployment of components, this way keeping the service within an allowed region of parameters defined by the requirements. As the results with our proposed framework show our logic will be a prominent candidate for a robust and adaptive service execution platform.

References

1. Kramer, F.A., Bræk, R., Herrmann, P.: Synthesizing components with sessions from collaboration-oriented service specifications. In: Gaudin, E., Najm, E., Reed, R. (eds.) SDL 2007. LNCS, vol. 4745, pp. 166–185. Springer, Heidelberg (2007)
2. OMG UML Superstructure, Version-2.2
3. Csorba, M., Heegaard, P., Herrmann, P.: Cost-Efficient Deployment of Collaborating Components. In: Meier, R., Terzis, S. (eds.) DAIS 2008. LNCS, vol. 5053, pp. 253–268. Springer, Heidelberg (2008)
4. OMG 2009, UML Profile for MARTE: Modeling & Analysis of Real-Time Embedded Systems, V – 1.0 (2009)
5. Trivedi, K.S.: Probability and Statistics with Reliability, Queuing and Computer Science application. Wiley-Interscience Publication, Hoboken, ISBN 0-471-33341-7
6. Lopez, J.P., Merseguer, J., Campos, J.: From UML activity diagrams to SPN: application to software performance engineering. ACM SIGSOFT Software Engineering Notes, NY (2004)
7. Distefano, S., Scarpa, M., Puliafito, A.: Software Performance Analysis in UML Models. FIRB-PERF (2005)
8. D'Ambrogio, A.: A Model Transformation Framework for the Automated Building of Performance Models from UML Models. In: WOSP (2005)
9. Khan, R.H., Heegaard, P.E.: Translation from UML to SPN model: A performance modeling framework. In: Aagesen, F.A., Knapskog, S.J. (eds.) EUNICE 2010. LNCS, vol. 6164, pp. 270–271. Springer, Heidelberg (2010)
10. Khan, R.H., Heegaard, P.: Translation from UML to SPN model: Performance modeling framework for managing behavior of multiple session & instance. In: ICCDA 2010 (2010)
11. Kramer, F.A.: ARCTIS, Department of Telematics, NTNU, http://arctis.item.ntnu.no
12. Rendezvous synchronization, http://book.opensourceproject.org.cn/embedded/cmprealtime/opensource/5107final/lib0091.html (retrieved June, 2010)
13. Efe, K.: Heuristic models of task assignment scheduling in distributed systems. Computer (June 1982)
14. Khan, R.H., Heegaard, P.: A Performance modeling framework incorporating cost efficient deployment of collaborating components. In: ICSTE (2010)
15. Trivedi, K.S., Sahner, R.: Symbolic Hierarchical Automated Reliability / Performance Evaluator (SHARPE). Duke University, Durham

Software Quality Models: A Comparative Study

Anas Bassam AL-Badareen, Mohd Hasan Selamat, Marzanah A. Jabar,
Jamilah Din, and Sherzod Turaev

Faculty of Computer Science and Information Technology
University Putra Malaysia
Anas_badareen@hotmail.com,
{hasan,marzanah,jamilah,sherzod}@fsktm.upm.edu.my

Abstract. In last decade, researchers have often tried to improve the usability, portability, integrity and other aspects of software in order for it to be more users friendly and gain user trust. Several approaches and techniques have been proposed to reduce the negative effects of software size and complexity. Moreover, several software quality models were proposed to evaluate general and specific type of software products. These models were proposed to evaluate general or specific scopes of software products. The proposed models were developed based on comparisons between the well-known models, in order to customize the closed model to the intended scope. These comparisons are leak of criteria that is conducted based on different perspectives and understanding. Therefore, a formal method of comparison between software quality models is proposed. The proposed method is applied on a comprehensive comparison between well-known software quality models. The result of the proposed method shows the strength and weaknesses of those models.

Keywords: Quality Model, Model Comparison, Model Development.

1 Introduction

US Air force Electronic System Division (ESD), the Rome Air Development Centre (RADC) and General Electric [1] intends to improve the quality of the software products and to make it measurable. therefore, McCall [2] model was developed in 1976-7, which is one of the oldest software quality models. This model started with a volume of 55 quality characteristics which have an important influence on quality, and called them "factors". The quality factors were compressed into eleven main factors in order to simplify the model. The quality of software products was defined according to three major perspectives, product revision (ability to undergo changes), product transition (adaptability to new environments) and product operations (its operation characteristics).

Since McCall model was proposed, new factors have been added to the original and some of them are redefined [3]. Second model was defined is Boehm model[4], the model was based on McCall model, he defined the second set of quality factors. SPARDAT is a commercial quality model was developed in the banking environment. The model classified three significant factors: applicability, maintainability, and adaptability.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 46–55, 2011.

Nowadays, several software quality models were proposed in order to evaluate general and specific software quality products, they were developed based on well-known models, such as McCall, Boehm, FURPS, Dromey, and ISO. The method of develop a software quality models is started based on comparisons between selected well-known models in order to customize the closed model to the intended scope, such as [5-9]. The comparisons were conducted based on different perspectives and understanding, and at the factors level. Therefore, a contradiction of the software quality factors definition is occurred.

Table 1. Sample of Models Comparisons

Quality Factors	Hamada, 2008 [3]			Rawashdeh, 2006 [4]			Haiguang, 2008 [7]		
	McCall	Boehm	ISO	McCall	Boehm	ISO	McCall	Boehm	ISO
Integrity	×	×	×	×		×	×	×	
Efficiency	×	×	×	×	×	×		×	Maintenance
Reusability	×	×		×			×	×	
Changeability		×	Maintenance		×			×	Maintenance
Testability	×		Maintenance	×	×	×	×		

Table 1 shows sample of comparisons were conducted between same models and same factors, but with a different results. It shows that the researchers were conducted the comparisons based on different points of views. For example, Hamada [5] shows that the integrity is included in McCall, Boehm, and ISO models. Rawashdeh [6] shows that the integrity in included under McCall and ISO. Haiguang [9] shows that the integrity is included in McCall and Boehm. Hamada [5] and Rawashdeh [6] show that the efficiency is included in all of the selected models, whereas Haiguang [9] shows that it is included in Boehm as a main factor and in ISO a sub factor under the maintainability.

However, the inconsistency in the definitions of software quality factors results contradictions in the developed models. Therefore, different software quality models are developed intends to achieve same goals of software quality evaluation. Thus, in this study we propose a method of comparison based on the analysis and discussion of four well-known software quality models. Moreover, we analyzed and compared several comparisons were conducted between these models and we identified the main differences between them.

This study intends to develop a formal method that can be used to compare and differentiate between software quality models mathematically. That will help to avoid any contradictions that may occur during development. Moreover, it helps to define a standard basic for developing a software quality model.

In section two, we present the well-known software quality models. In section three, we discuss and describe the proposed method. In section four, we describe the case study and how the weight is assigned. In section five, we present and discuss the result. Section seven present the conclusion and future work.

2 Quality Models Background

Whereas, several software quality models are proposed, in order to evaluates different types of software products. This section presents the most popular software quality models, were considered in different studies.

2.1 McCall Model

McCall's model [2] was developed by the US air-force electronic system decision (ESD), the Rome air development center (RADC), and general electric [1], to improve the quality of software products. This model was developed to assess the relationships between external factors and product quality criteria. Therefore, the quality characteristics were classified in three major types, 11 factors which describe the external view of the software (user view), 23 quality criteria which describe the internal view of the software (developer view), and metrics which defined and used to provide a scale and method for measurement. The number of the factors was reduced to eleven in order to simplify it. These factors are Correctness, Reliability, Efficiency, Integrity, Usability, Maintainability, Testability, Flexibility, Portability, Reusability, and Interoperability. The major contribution of this model is the relationship between the quality characteristics and metrics. However, the model not consider directly on the functionality of software products.

2.2 Boehm Model

Boehm [4] added new factors to McCall's model and emphasis on the maintainability of software product. The aim of this model is to address the contemporary shortcomings of models that automatically and quantitatively evaluate the quality of software. Therefore, Boehm model represents the characteristics of the software product hierarchically in order to get contribute in the total quality. Furthermore, the software product evaluation considered with respect to the utility of the program. However, this model contains only a diagram without any suggestion about measuring the quality characteristics.

2.3 FURPS Model

FURPS model was proposed by Robert Grady and Hewlett-Packard Co. The characteristics were classified into two categories according to the user's requirements, functional and non-functional requirements. Functional requirements (F): Defined by input and expected output. Non-functional requirements (URPS): Usability, reliability, performance, supportability. And then, the model was extended by IBM Rational Software – into FURPS+. Therefore, this model considered only the user's requirements and disregards the developer consideration. However, the model fails to take into account the software some of the product characteristics, such as portability and maintainability.

2.4 Dromey Model

Dromey (1995) [10] states that the evaluation is different for each product, hence a dynamic idea for process modeling is required. Therefore, the main idea of the proposed model was to obtain a model broad enough to work for different systems. The model seeks to increase understanding of the relationship between the attributes (characteristics) and the sub-attributes (sub-characteristics) of quality. This model defined two layers, high-level attributes and subordinate attributes. Therefore, this model suffers from lack of criteria for measurement of software quality.

2.5 ISO IEC 9126 Model

Since, the number of the software quality models were proposed, the confusion happened and new standard model was required. Therefore, ISO/IEC JTC1 began to develop the required consensus and encourage standardization world-wide. The ISO 9126 is part of the ISO 9000 standard, which is the most important standard for quality assurance. First considerations originated in 1978, and in 1985 the development of ISO/IEC 9126 was started.

In this model, the totality of software product quality attributes were classified in a hierarchical tree structure of characteristics and sub characteristics. The highest level of this structure consists of the quality characteristics and the lowest level consists of the software quality criteria. The model specified six characteristics including Functionality, Reliability, Usability, Efficiency, Maintainability and Portability; which are further divided into 21 sub characteristics. These sub characteristics are manifested externally when the software is used as part of a computer system, and are the result of internal software attributes. The defined characteristics are applicable to every kind of software, including computer programs and data contained in firmware and provide consistent terminology for software product quality. They also provide a framework for making trade-offs between software product capabilities.

3 The Comparison Method

In order to show the clear differences between software quality models, mathematical comparison method is proposed. The method aims to show the clear and accurate differences between quality models, which consider the sub factors in addition to the factors. It consists of four main tasks: model selection, assigning values, factors comparison, and models comparison.

Models Selection: The process of models selection is depends on the scope intended to be evaluated, usually the well-known software quality models are considered in developing a new model.

Factors Selection: The factors of the selected models are collected and combined in one structural tree (Fa, Fb…Fn) (*See figure 1*).

Different sub-factors are considered in each factor from different model, the sub factors are combined under their factors (S1, S2, Sn). According to the aim and definitions of each factor and sub factor, the repeated are excluded.

Factors Weighting: After analyze the scope that needed to be evaluated, the experts in this field are required to assign the weight of these factors (W1, W2…….Wn) and sub factor (Wa, Wb……Wm) are assigned.

Factors Values: the value of each factor in the original models is calculated, based on the weight that assigned in the previous step.

- First, the value of the same factor within the selected models is calculated (*Formula 1*).
- Second, the total value of each model is calculated (*Formula 2*), based on the calculated values of their factors.

The Comparison: the total value for each factor is compared between the selected models. That shows the comprehensiveness differences between these factors in different models.

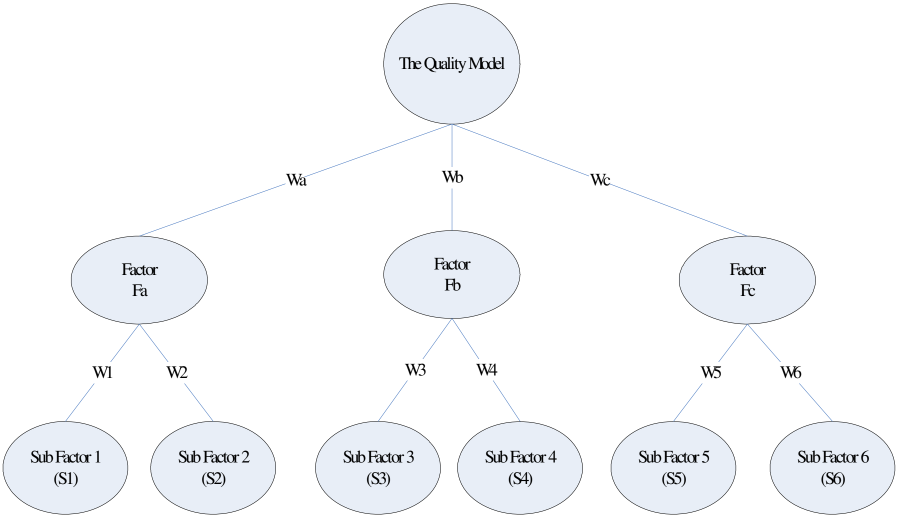


Fig. 1. The Model Weighting

$$F = \sum_{1}^{n} Wn \tag{1}$$

$$Qm = \sum_{1}^{m} Fm \times Wm \tag{2}$$

$$Qm = \sum_{1}^{m} Wm \times \sum_{1}^{n} Sn. \tag{3}$$

4 Case Study

The case study considered a general comparison between most popular software quality models. The comparison shows the main differences between these models. The following steps are followed in order to perform the task:

Step 1: combine the factors of the selected models and remove the repeated
Step 2: combine the sub-factors for each factor
Step 3: assign the weight for each factor
Step 4: assign the weight for each sub factor
Step 5: calculate the weight for each factor in every model independently
Step 6: compare the values of same factors in all of the selected models

Whereas, the comparison is a general term comparison, the weight of the software characteristics is considered equivalently. While, in order to compare the models for specific scope, the expertise weighting is very important to show the most model enclose to the scope and which model emphasize on the characteristics of the intended scope.

A simple formula is used to assign the values for these attributes. In order to present whether the characteristic is considered as a factor, 50% was given to the factor and 25% to the sub factor. For the rest of the percentage it was equivalently divided between the numbers of the sub factors included in the comparison.

5 Result and Discussion

The first step of this study was to collect the factors that included by selected models and remove the repeated according to the definition of each of them. The second step is combining the sub-factors from all of the models for specific factor. The repeated sub characteristics were removed according to the definition of each of them.

The values were seated equivalently which gave 50% of the value to present whether the factor is included in the model, whereas 25% was given if the characteristic is included as a sub factor. Because of the generality of this comparison which not considered any type of software or any specific software domain, the value of the factors are same. Therefore, the second 50% was divided between the sub factors that included by the selected factor equivalently, where the 50 is divided into the number of the sub factors were collected for the factor from those models.

Finally, in order to calculate the value for each model, the values for each factor within the same model are calculated according to the same formula that was used to calculate the values of the factors. Table 2 presents the total value for each model, whereas figure 2 shows the graphical presentation of these values. The total value was decomposed between those factors equivalently, where each factor was given 7.14% from the total quality of the model. Figure 3 shows the difference between those models per factor.

McCall model is a hierarchical model with two levels, the models has many to many relationships. This model considered the most of the software product characteristics, except the functionality of the software and the human engineering, whereas software understandability is covered implicitly through the sub-characteristics that required for other factors such as maintainability, flexibility, and reusability. Furthermore, there is no any mention about the human characteristics, which is important to identify the usability characteristic for the software product. Moreover, the model is the highest model which cover the most of the software product characteristics, whereas lake to the relationships between the factors, which cause overlapping in its relations.

Table 2. The total value of the software quality models

Factor/Model	*ISO*	*McCall*	*Boehm*	*FURPS*	*Dromey*
Efficiency	80%	70%	55%	25%	50%
Integrity	25%	100%	25%	25%	0%
Reliability	70%	65%	50%	65%	50%
Usability	63%	60%	60%	73%	0%
Correctness	0%	100%	0%	0%	50%
Maintainability	73%	68%	64%	0%	50%
Testability	25%	78%	53%	0%	0%
Changeability	25%	83%	42%	0%	0%
Interoperability	25%	100%	25%	0%	0%
Reusability	0%	100%	0%	0%	50%
Portability	78%	67%	61%	0%	50%
Functionality	86%	0%	0%	71%	50%
Understandability	0%	0%	25%	0%	0%
Human Engineering	0%	0%	75%	25%	0%
Total	**39.27%**	**63.67%**	**38.19%**	**20.34%**	**25.00%**

Boehm model is similar to McCall model, which is hierarchical structure. This model focused on the structure of the quality characteristics, whereas several characteristics are dropped from this model such as software correctness, reusability, in addition to the functionality. Furthermore, this model considered new quality characteristics there were not included in McCall model such as Human Engineering and developer understandability. The model is very success in software maintainability relations, which reduce the overlapping and represent perfectly the meaning of the software maintainability. At the same time, the does not consider the software from the end user perspective and just be content with the developer perspective.

FURPS model is a two level hierarchical model with one to many relationships. This model evaluates the software products only from the user's viewpoint, whereas no any mention from the developer viewpoint such as maintainability and reusability. However, this model succeed to cover the user concerned in the software product, where is missed the software portability. Furthermore, the evaluation on this model considered also the supportability facilities that required for operating the system properly, which are not considered any model else.

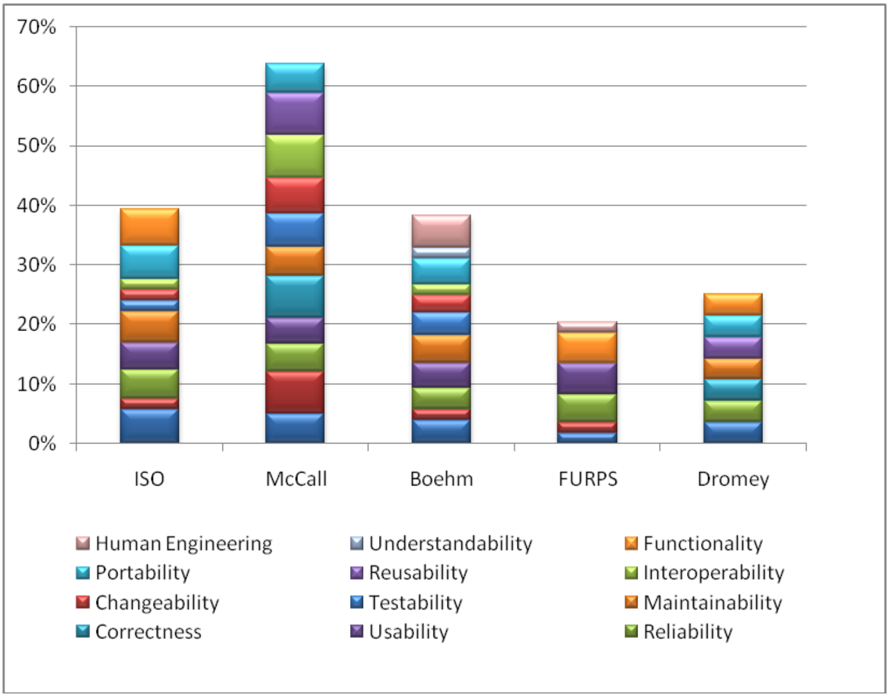


Fig. 2. Total Values Comparison

Dromey model is a dynamic software quality model which aims to evaluate different type of software products. The model considered four levels of software products, Correctness, Internal, Contextual, and Descriptive. Moreover, with each level the software quality attributes are considered. Whereas, the model suffer from the lack of criteria for measure the software quality. Furthermore, the model does not consider the usability of the software.

ISO model which proposed to overcome the confusing of the software quality models, considered six main quality factors related to twenty one criteria. However, this model in addition to the generality the model lakes to several factors were considered in previous models such as software correctness, reusability, and human engineering.

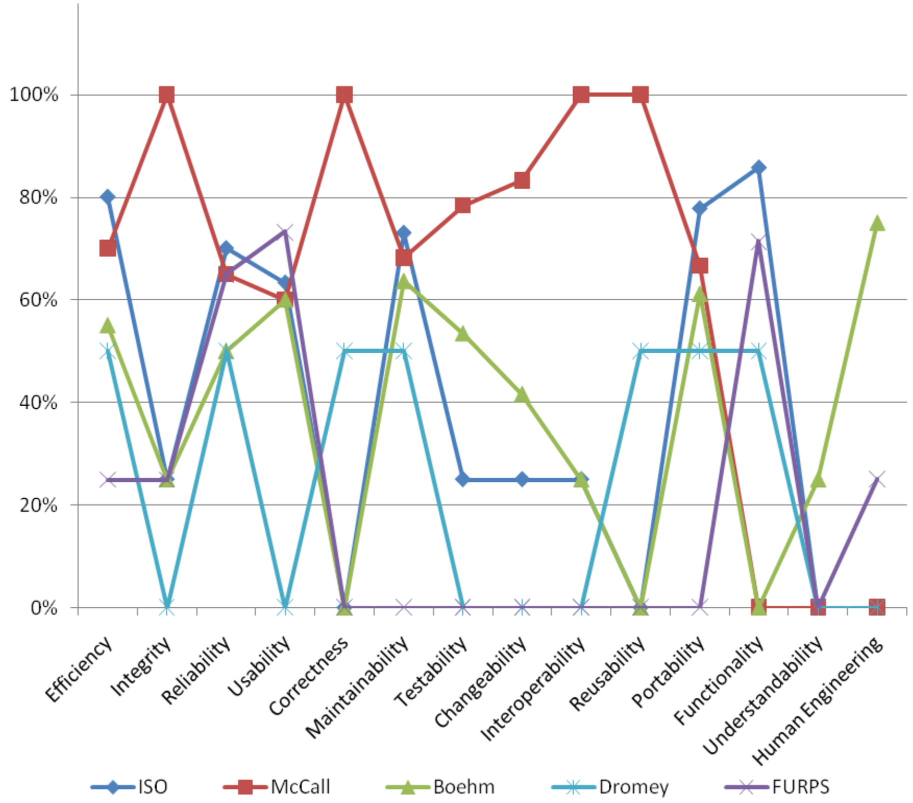


Fig. 3. Models Comparison

6 Conclusion

In this paper, well-known software quality models were presented. Hence, each model was discussed in details, the advantages and disadvantages were expressed. Finally, comprehensive comparison between the selected models was presented. The comparison goes behind the definitions of the software quality factors into sub factors and criteria.

Furthermore, new comparison method was proposed, in order to get clear and accurate differences between software quality models. The comparison was basic on mathematical formula, in order to show graphically the differences between those models. This method requires assign values for the sub factors moreover the main factors. Which is gave a clear picture of the differences between the models.

The values in this study were given equivalently between the factors and between the sub factors that is because this comparison was generally. In specific domain, the values for each factor and sub factor have to be defined according to the selected domain. For future work, the proposed method has to be verified by apply it in specific domain.

References

1. Ravichandran, T., Rothenberger, M.A.: Software reuse strategies and component markets. Commun. ACM 46, 109–114 (2003)
2. McCall, J.A., Richards, P.K., Walters, G.F.: Factors in Software Quality. Griffiths Air Force Base, N.Y. Rome Air Development Center Air Force Systems Command (1977)
3. AL-Badareen, A.B., Selamat, M.H., Jabar, M.A., Din, J., Turaev, S., Malaysia, S.: Users' Perspective of Software Quality. In: The 10th WSEAS International Conference on Software Engineering, Parallel And Distributed Systems (SEPADS 2011), pp. 84–89. World Scientific and Engineering Academy and Society (WSEAS), Cambridge (2011)
4. Boehm, B.: Characteristics of software quality (1978)
5. Hamada, A.A., Moustafa, M.N., Shaheen, H.I.: Software Quality model Analysis Program. In: International Conference on Computer Engineering & Systems, pp. 296–300 (2008)
6. Rawashdeh, A., Matalkah, B.: A new software quality model for evaluating cots components. Journal of Computer Science 2, 373–381 (2006)
7. Behkamal, B., Kahani, M., Akbari, M.K.: Customizing ISO 9126 quality model for evaluation of B2B applications. Information and Software Technology 51, 599–609 (2009)
8. Kumar, A., Grover, P.S., Kumar, R.: A quantitative evaluation of aspect-oriented software quality model (AOSQUAMO). SIGSOFT Softw. Eng. Notes 34, 1–9 (2009)
9. Haiguang, F.: Modeling and Analysis for Educational Software Quality Hierarchy Triangle. In: Seventh International Conference on Web-based Learning, pp. 14–18 (2008)
10. Dromey, G.: A Model for Software Product Quality. IEEE Transactions on Software Engineering 146, 21 (1995)

QTCP: An Optimized and Improved Congestion Control Algorithm of High-Speed TCP Networks

Barkatullah Qureshi, Mohamed Othman, Shamala Sabraminiam,
and Nor Asila Wati

Department of Communication Technology and Networks
Faculty of Computer Science and Information Technology
University Putra Malaysia, 43400 UPM, Serdang, Selangor, DE, Malaysia
barkatupm@yahoo.com,
{mothman,shamala,asila}@fsktm.upm.edu.my

Abstract. TCP researchers evaluated the performance and fairness of different TCP protocols on the basis of new algorithms. The new High-Speed Transport Control Protocols (HS-TCP) were developed but there are still many problems regarding to bandwidth utilization, throughput and packet loss rate. To overcome these problems Quick Transport Control Protocol (QTCP) algorithm based on optimizations of HS-TCP slow start algorithm and Additive Increase and Multiplicative Decrease (AIMD) algorithm have been proposed. A modified algorithm has been developed by using an additive increase approach to grow window with normal speed and to increase scalability by putting constant value of stability of timeline in congestion avoidance phase. This constant timeline gives long stability time; it provides many benefits as compared to other high-speed TCP protocols. The improved algorithm increased throughput and decreased packet loss rate and fairly share link utilization. In this regards several experiment of simulations were observed the fairness. The results show best bandwidth utilization, improved throughput and less packet loss rate as compared to other high speed TCP variants.

Keywords: Fairness, QTCP, AIMD, Congestion Avoidance, Throughput.

1 Introduction

Transport Control Protocols (TCP) is one of the most widely used transport agent which is being extensively used since last couple of decades. There are several versions of TCP that the researchers compared and evaluated on the basis of algorithms [1]. On the other hand new high-speed TCP protocols have been expanded and designed for solving the problem of bandwidth limitation, particularly when the data is attempted to send by a single connection at very high-speed it is very difficult to maintaining the efficiency and fairness to the standard TCP flows [2]. The most imperative protocols are High-Speed Transport Control Protocol (HS-TCP), Scalable-TCP (STCP), Hamilton TCP (HTCP), Binary Increase Control (BIC), and CUBIC. On high bandwidth-delay product (BDP) networks, the main optimizations consist of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 56–67, 2011.

adding more efficient mechanisms for acquiring bandwidth faster. The comprehensive analysis of different highspeed TCP versions carried out [3] and found dynamic sensitive fairness metric for high bandwidth delay product networks. The fluid flow model of highspeed TCP/RED network proposed [4], to examine the performances of highspeed delay product networks with RED active queue management at the router.

HS-TCP [5] modifies the standard TCP response function to acquire very fast available bandwidth (more efficiency) and quickly recovers packet losses in the network. The drawback of such a behavior is that fairness between TCP and HS-TCP, and even between HS-TCP flows, is affected since HS-TCP is much slower to give back bandwidth.

In high-speed TCP networks few problems are faced; the congestion which produces other problems like loss rate, RTT fairness, and link/band width utilization. Many researchers have worked on high-speed TCP and suggested enhanced algorithms for optimizing performance of TCP. Several algorithms are designed to reduce loss rate and other parameters but still there is a need for smarter optimization techniques in High-speed TCP. Congestion in High-speed TCP can be optimized by improvement in AIMD which increases throughput and decrease loss rate in slow start, and fairly share bandwidth link utilization. The purpose of this study is to create an adaptive algorithm named Quick Transport Control Protocol (QTCP) based on HS-TCP. QTCP changes decrease factor like CUBIC which will grows the window size in slow start with a speed that increases link utilization, throughput, decreases packet loss rate and maintain inter and intra protocol fairness.

2 Related Work

Numerous works has been done on congestion control algorithm, an adaptive window algorithm HS-TCP has been discovered [6] and [7]. It has been reported that this algorithm has capacity to operate on a very large Bandwidth Delay Product (BDP) which is 10^4 packets or more, in networks. In the congestion window the increment and decrement of window size is dependent upon reply to an acknowledgment or packet loss. Mostly the preceding research findings [8], [9], [10] and [11] were based on the two issues which are about the links in the same time scenario. First, how the TCP implementations perform individually, second how fairly share the bandwidth link utilization.

2.1 HS-TCP

HSTCP Protocol [5], [6] and [7] modules increase and reduce congestion window parameters according to the current value of window cwnd. It uses an AIMD model with a logarithmic modulation of the parameters according to the value of the congestion window as following:

```
if ACK then cwnd ← cwnd + $\alpha$ (cwnd)
    else cwnd ← $\beta$ (cwnd) * cwnd
```

2.2 STCP

One of the most basic concepts of STCP proposed [12] is to make the recovery time after a congestion event independent of window size. Particularly TCP cwnd modernized by, STCP as follows:

$$\begin{array}{ll} Ack: & \mathrm{cwnd} \leftarrow \mathrm{cwnd} + \alpha \\ \mathrm{Loss}: & cwnd \leftarrow \beta \times cwnd \end{array}$$

2.3 HTCP

The elapsed time Δ used for HTCP until last congestion event occurs. It specifies the bandwidth-delay product and the parameter increased AIMD as a function of **Δ** [13]. The path of RTT is also scaled as increment in AIMD to alleviate unfairness between competing flows with different round-trip times. The AIMD decreasing factor is adjusted to improve link utilization based on an estimate of the queue provisioning on a path. In more details, HTCP proposes that cwnd be updated as follows:

$$Ack: \qquad cwnd \leftarrow cwnd + \frac{2(1-\beta)f_a(\Delta)}{cwnd} \tag{1}$$

$$Loss: \qquad cwnd \leftarrow g_{\beta}\,(B) \times cwnd \tag{2}$$

with

$$\begin{array}{l} f_{\alpha}(\Delta) = \begin{Bmatrix} 1 & \Delta \le \Delta_L \\ \max\left(\bar{f}_{\alpha}(\Delta)\,T_{mim}, 1\right) & \Delta > \Delta_L \end{Bmatrix} \\ g_{\beta}(B) = \begin{Bmatrix} 0.5 & \left|\frac{B(K+1)-B(k)}{B(k)}\right| > \Delta_B \\ \min\left(\frac{T_{\min}}{T_{\max}}, 0.8\right) & otherwise \end{Bmatrix} \end{array} \tag{3}$$

2.4 BIC

The study on BIC [14] employed a form of binary search algorithm to update cwnd. Briefly, a variable w_1 is maintained that holds a value halfway between the values of cwnd just before and just after the last loss event. The revised rule of cwnd search swiftly increases cwnd when it is beyond a particular distance Smax from w_1, and slowly revises cwnd when its value is close to w_1. In this protocol for packet losses detection, utilize Multiplicative backoff, with a recommended backoff factor of 0.8. In more detail, the BIC revise algorithm is as follows:

$$\mathrm{Ack}\ : \begin{Bmatrix} \delta = \dfrac{(w_1 - cwnd\ )}{\mathrm{B}} \\ cwnd\ \leftarrow\ cwnd\ + \dfrac{f_{\alpha}\,(\delta - cwnd\ )}{cwnd} \end{Bmatrix} \tag{4}$$

$$\text{Loss} : \left\{ \begin{array}{l} w_1 = \begin{cases} \dfrac{1+\beta}{2} \times cwnd & cwnd < w_1 \\ cwnd & otherwise \end{cases} \\ w_2 = cwnd \\ \\ cwnd \leftarrow \beta \times cwnd \end{array} \right\} \tag{5}$$

Where

$$f_\alpha(\delta, cwnd) = \left\{ \begin{array}{ll} B/\delta & (\delta \le 1, \quad cwnd \le w_1) \\ & or\,(w_1 \le cwnd < w_1 + B) \\ \delta & 1 < \delta \le S_{\max}, cwnd < w_1 \\ w_1/(B-1) & B \le cwnd - w_1 < S_{\max}(B-1) \\ S_{\max} & otherwise \end{array} \right\} \tag{6}$$

2.5 CUBIC

It has been reported that the modified and improved version of BIC [15], expressed as CUBIC, to achieve more BIC's fairness. The study verified a cubic function to increase the window size. $cwnd = \mathrm{C(t - K)}^3 + \mathrm{W}max$, In this cubic function the constant used for scaling is C, time for the window was last reduced is t, the size of the window just before the window was last reduced is *Wmax*, and $\mathrm{K} = Wmax.\beta / C)^{1/3}$, where the constant decrease factor is β . When a loss occurs, the window is reduced to β. *Wmax*, with β = 0.8. CUBIC maintains inter and intra protocol fairness.

3 Congestion Control Algorithm QTCP: In High-Speed Protocols

Numerous studies on HS-TCP [5], [6] and [7] to develop a new algorithm QTCP show that the two most significant parameters are cwnd and time. They have examined the protocols HS-TCP, CUBIC and BIC graphically and observed that in slow start if cwnd grows fast it increases bandwidth utilization, throughput and packet loss rate. If cwnd grows slow then it will decrease all three quality parameters; for instance we need intermediate solutions that will increases throughput and utilization, and decreases loss rate in slow start.

In congestion avoidance phase if cwnd is growing fast it will increase loss rate and remaining parameters will be positive. However, packet losses is not good for TCP improvement, from previous algorithms studies it is evident that in congestion avoidance phase if cwnd grows with slow speed it will increase its stability and will take time to reach its saturation point and as a result packet loss will occur after long time to stability. If cwnd grows fast then it will reduce its stability and will reach to its saturation point very fast. As a result of less stability and swiftness packet loss rate increased. In prior studies, HS-TCP specified as more aggressive than other protocols and due to this violent characteristic packet loss rate increased rapidly. For resolving this problem when AIMD phase was modified, the length of timeline of cwnd remained stable in congestion avoidance phase as shown in Fig. 1 and following algorithm.

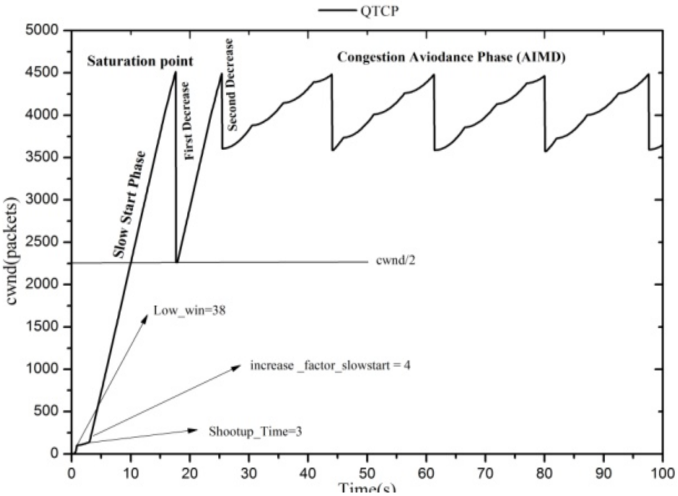


Fig. 1. QTCP congestion window behavior

```
  // Initialization:
  max_inc_sst = 30 // maximum increment in slow start
  inc_f_sst = 4   // increase factor in slow start
  sht_sst = 3 // shoot up time in slow start
  dec_f= 0.8   // decrease factor
  inc_update _delay = 0.5 // increment update delay
  inc_f_ca = 1//increase factor in congestion avoidance
  max_inc_ca=10//maximum increase in congestion avoidance
  sst = 1 //Slow start
  l_win = 38   //Low window
  // On Each Acknowledgement:
      // cwnd is the congestion window size
      if (cwnd <= l_win )
      {
            cwnd = cwnd + 1/cwnd
      }
      else
      {
            if (sst)
               q_slow_start ();//QTCP slow start function
            else
q_congestion_avoidance();//QTCP Congestion avoidance function
      }

  q_slow_start (): // QTCP slow start function

      if (elapsed_time <= sht_sst)
            increment = elapsed_time
      else
            if increment < max_inc_sst)
                  increment += inc_f_sst

      cwnd = cwnd + increment;

      last_update_time = elapsed_time

  q_congestion_avoidance (): //QTCP Congestion avoidance function

      if ((elapsed_time - last_update_time) >= inc_update_delay)
      {
            increment = increment + inc_f_ca
            last_update_time = elapsed_time
      }

      if (increment > max_inc_ca)
            increment = 1

      cwnd = cwnd + increment

  OnPacketLoss:

      sst = 0
      increment = 1;
      ssthresh_ = cwnd = dec_f * cwnd
```

4 Simulation Topology

In this simulation, flow 1 started from 0 seconds and the flow 2 started after the 50 seconds. The running time for each simulation is 500 seconds. Synchronization loss occurs when same RTT used for both flows. This research emphasizes to factoring out the effects of RTT on the simulation results. Mostly high-speed protocols are not RTT fair [16], [17] and [18], thus there is a main difference between the flows. In this regard the flow with shorter RTT attained higher throughput as compared to flows with longer RTTs. This research is focused on the competing flows therefore we select balanced RTTs and tested five high-speed loss based protocols (HS-TCP, HTCP, Scalable TCP, BIC and CUBIC). In this testing for every protocol and maximum router buffer size, six sets of experiments are run. The recommended protocol parameters are used for each protocol.

All experiments were done in ns-2 version 2.35 network simulator [19] using the topology shown in Fig.2. Two senders are on the left side and two receivers are on the right side of the network respectively. Each end node is connected to a router by 1Gbps link with a propagation delay of 1ms. Bottleneck link capacity between two routers is connected by a 622 Mbps and propagation delay 48ms. Round Trip Time

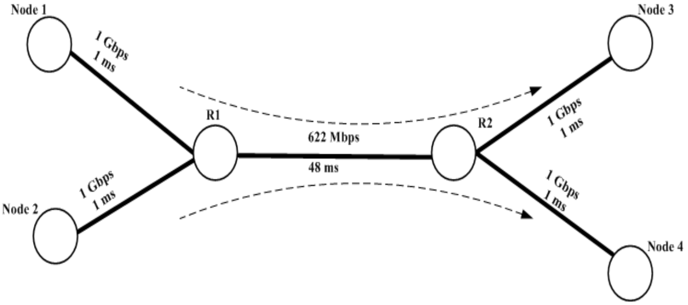


Fig. 2. Network Topology

(RTT) for each sender is 100ms. In this network the bandwidth delay product (BDP) is 7775, 1000 bytes segments, and the both routers used drop-tail queues. Three different routers queue buffer length 100%, 20% BDP and 40 segments are used in the full set of experiment. Maximum window size of 67,000 segments approximately 64 MB is used to conformed that TCP is not a linking factor. There are two connections started in each simulation one from node 1 and one from node 2.

4.1 Performance Evaluation Criteria

In this section we emphasize on the evaluation performance of QTCP using the above simulation. The main focus is on bandwidth utilization, throughput, fairness and friendliness. The QTCP protocol has very low packet loss rate therefore our protocols has the characteristic of TCP friendliness. Table 1 illustrates the effect of QTCP and other high-speed TCP protocols and evaluates the bottleneck link utilization, throughput and packet loss rate of flow1 and flow2.

Table 1. Evaluation of QTCP algorithm

High-speed TCP (Flow1-Flow2)	**Bottleneck link Utilization (%)**	**Throughput (%) (Flow1-Flow2)**	**Packet Loss rate (Flow1-Flow2)**
CUBIC-CUBIC	97.89	97.88	0.0016
HS-HS	97.44	97.44	0.0075
QTCP-QTCP	98.07	98.19	0.0011
QTCP-BIC	98.21	98.20	0.0094
QTCP-CUBIC	98.22	98.22	0.0095
QTCP-HS	97.65	97.65	0.0104
QTCP-HTCP	84.31	84.31	0.0159
QTCP-STCP	98.32	98.31	0.0661

We examined the congestion window of the two pairs QTCP-HTCP and CUBIC-HTCP, as shown in Fig. 3, which correspond to the fairest pair then other protocols.

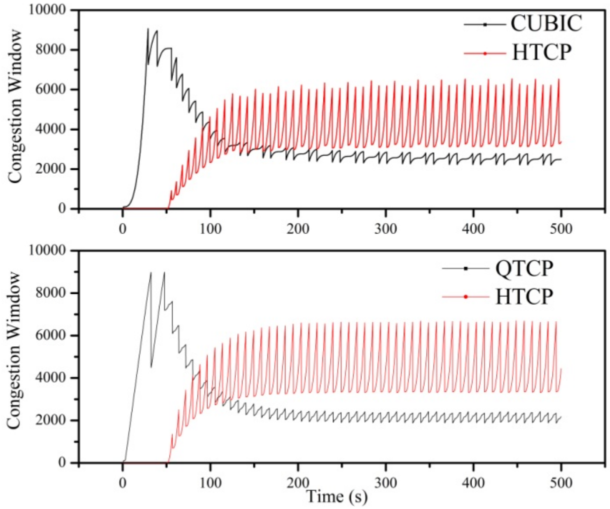


Fig. 3. Congestion window for the behavior of CUBIC- HTCP and QTCP- HTCP

The congestion window of two pairs QTCP-QTCP and HTCP-HTCP as shown in Fig. 4 is the fairer among the other pairs.

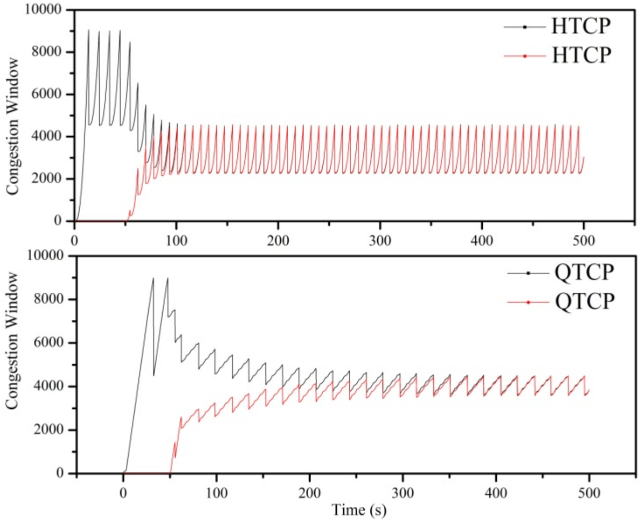


Fig. 4. Congestion window for the performance of HTCP and QTCP shows experiments are fairer

The congestion windows of pairs HS-HS and CUBIC- QTCP, as shown in Fig. 5, observed that these two protocols are slightly fair than BIC-BIC, HS-HTCP, BIC-CUBIC, BIC-HS and HS-BIC.

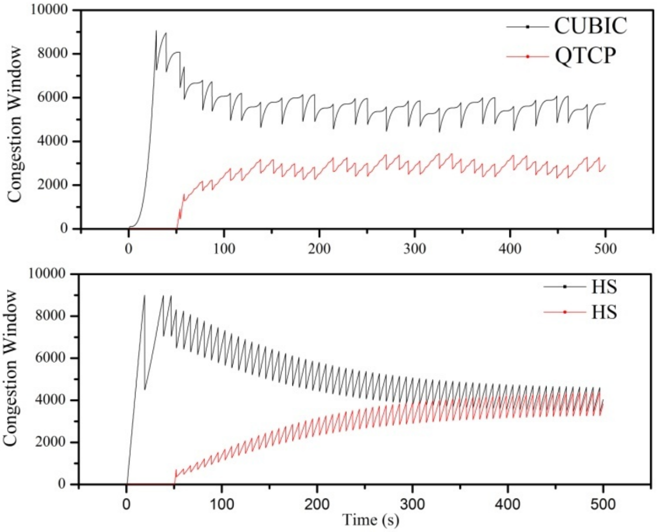


Fig. 5. Congestion window for the behavior of CUBIC - QTCP and HS-HS

The congestion windows of pairs HS-QTCP and BIC-QTCP as shown in Fig. 6 are less unfair as compared to other pairs.

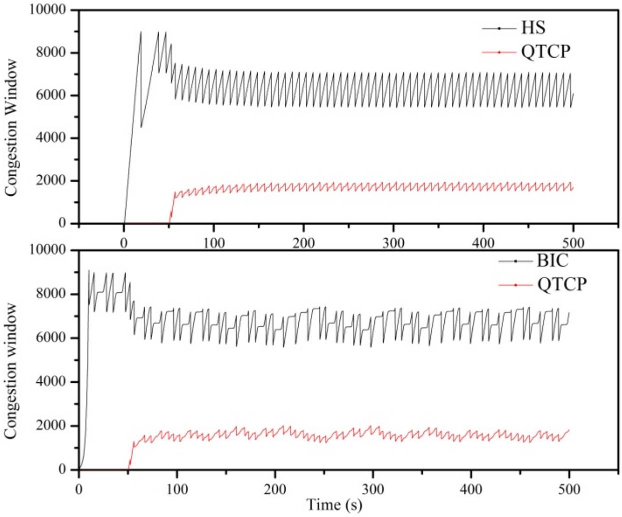


Fig. 6. Congestion window shows HS and BIC more aggressive than QTCP

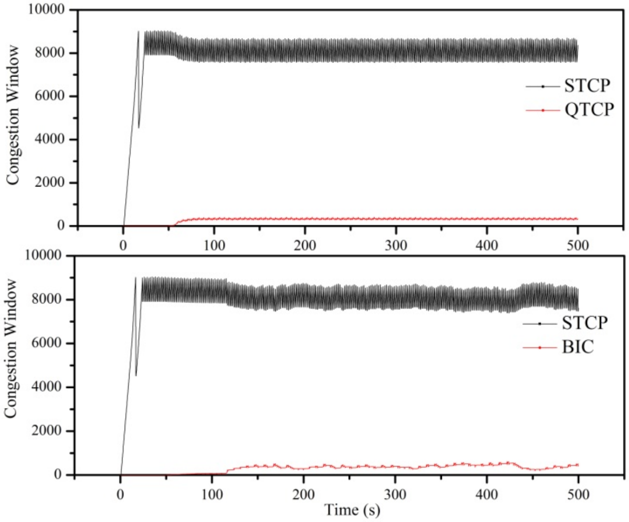


Fig. 7. Congestion window STCP-QTCP and STCP-BIC shows unfairness

The congestion window of two pairs STCP-QTCP and STCP-BIC shown in Fig. 7 s lightly less unfair then other pairs.

4.2 Jain's Fairness Index Evaluation

It is the fairness between two flows of the same protocol but the sending and receiving hosts are different. Wherein different performances are evaluated [9], to examine how the different protocols fairly behave towards each other. In this research the fair share link metric considered and computed the fairness index [20].

$$\mathrm{F} = \frac{(\sum_{i=1}^{n} \bar{x}_i)^2}{n \sum_{i=1}^{n} \bar{x}_i^{\,2}} \tag{7}$$

Where n is number of flows, 1/n Capacity of bottleneck link, x_i average bandwidth of each source i. The perfect value of fairness index of throughput for all protocols is 1. We evaluate same and different pairs of high-speed TCP and run several simulations on the basis of Jain fairness index and observed their fairness. The evaluation result is then sorted out and divided into pairs of five groups G-A, G-B, G-C, G-D and G-E as shown in Fig. 8.

In group (G-A) we assessed that the efficiency and fairness of all high-speed TCPs pairs are more fair then other groups. We observed that group G-B, group G-C, group G-D and group G- E are gradually less fair then group G-A.

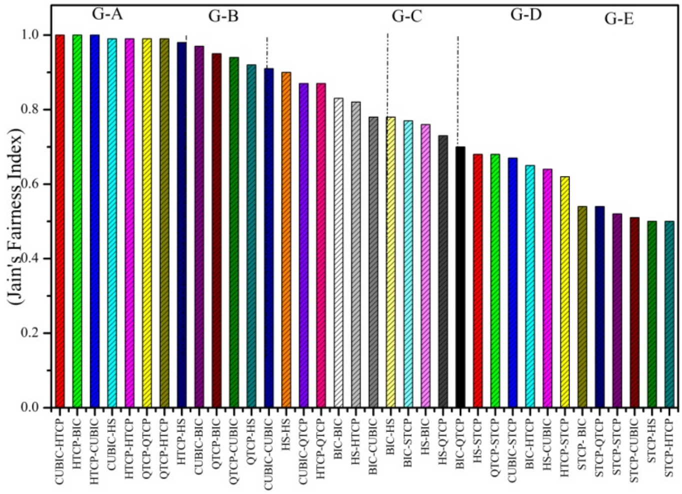


Fig. 8. Jain's fairness Index

5 Conclusions

We propose QTCP as an optimized improved algorithm for fairness of different high-speed protocols. In this algorithm scalability can be increased in congestion avoidance phase by putting constant value of stability timeline. It makes constant timeline between decrease event and saturation point and gives long stability time. The simulation results showed that proposed algorithms has better fairness and friendliness as compared to other high-speed transport control protocols. We studied the performance metrics of high-speed TCP protocols and evaluated that HTCP and QTCP are fairer than other protocols on the basis of Jain fairness index. However there are limitations in HTCP such as average throughput and link utilization is lower, and packet loss rate is higher than QTCP.

Acknowledgments

This work was supported by the Research University Grant Scheme (RUGS Number: 05/01/07/0180RU).

References

1. Qureshi, B., Othman, M., Hamid, N.A.W.: Progress in Various TCP Variants: Issues, Enhancements and Solutions. Mausaum Journal of Computing 1(14), 493–499 (2009)
2. Weigle, M.C., Sharma, P., Freeman, J.: Performance of Competing High-Speed TCP Fows. In: IFIP Networking, Coimbra, Portugal (2006)
3. Sonkoly, B., Trinh, T.A., Molnár, S.: Benchmarking High Speed TCP Fairness. Technical Report, Budapest University of Technology and Economics, BME (2007)

4. Sonkoly, B., Trinh, T.A., Molnár, S.: Understanding Highspeed TCP: A Control-theoretic Perspective. In: Third IASTED International Conference on Communications and Computer Networks, Marina del Rey, CA, USA, pp. 24–26 (2005)
5. Floyd, S., Ratnasamy, S., Shenker, S.: Modifying TCP's Congestion Control for High Speeds. Technical note (2002)
6. Floyd, S.: High Speed TCP for Large Congestion Windows. RFC, 3649 Experimental (2003), http://www.icir.org/floyd/hstcp.html
7. Floyd, S.: Limited Slow-Start for TCP with Large Congestion Windows. RFC, 3742 (2004), http://www.ietf.org/rfc/rfc3742.txt
8. Antony, A., Blom, J., de Laat, C., Lee, J., Sjouw, W.: Microscopic Examination of TCP Flows over Transatlantic Links. Future Generation Systems 19, 1017–1029 (2003)
9. Bullot, H., Cottrell, R.L., Hughes-Jones, R.: Evaluation of Advanced TCP Stacks on Fast Long-distance Production Networks. Journal of Grid Computing, 345–359 (2003)
10. Souza, E., Agarwal, D.A.: A Highspeed TCP Study: Characteristics and Deployment Issues. Technical report, LBNL-53215 (2003)
11. Tokuda, K., Hasegawa, G., Murata, M.: Performance Analysis of Highspeed TCP and its Improvements for High Throughput and Fairness against TCP Reno Connections. In: Highspeed Networking Workshop (2003)
12. Kelly, T.: Scalable TCP: Improving Performance in High-Speed Wide Area Networks. Computer Communication Review, 83–91 (2003)
13. Shorten, R.N., Leith, D.J.: H-TCP: TCP for High-speed and Long-distance Networks. In: Proceedings of PFLDnet, Argonne, Illinois (2004)
14. Xu, L., Harfoush, K., Rhee, I.: Binary Increase Congestion Control for Fast, Long Distance Networks. In: Proceeding of IEEE, INFOCOM (2004)
15. Rhee, I., Xu, L.: CUBIC: A New TCP-friendly High-Speed TCP Variant. In: Proceedings of PFLDnet, Lyon, France (2005)
16. Mbarek, R., Tahar bin Othman, M., Salem, N.: Performance Evaluation of Competing High-Speed TCP Protocol. International Journal of Computer Science and Networking Security, 99–105 (2003)
17. Pan, X.-z., Su, F.-j., Lu, Y., Ping, L.-d.: CW-HSTCP: Fair TCP in High-Speed Networks. Journal of Zhejiang University Science, 172–178 (2006)
18. Su, F.-j., Pan, X.-z., Wang, j.-b., Wan, Z.: An Algorithm for Reducing Loss Rate of High-Speed TCP. Journal of Zhejiang University Science, 245–251 (2006)
19. McCanne, S., Floyd, S.: Ns-2 Network Simulator, http://www.isi.edu/nsnam/ns/
20. Chiu, D., Jain, R.: Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks. Compute Networks and ISDN Systems, 1–14 (1989)

A Study on the Forecast of Meridian Energy and Biochemical Test by Using Bio-inspired NeuroMolecular Computing Model

Yo-Hsien Lin* and Hao-En Chueh

Department of Information Management,
Yuanpei University, Hsinchu, Taiwan, R.O.C.
{yohsien,haoen.chueh}@gmail.com

Abstract. There were differences between diagnosis methods of Chinese and Western medicine. The development of Meridian Energy devices has provided Chinese medicine doctors with a scientific way of testing and diagnosis. In this research, we tried to see if the test results of Meridian Energy could match the ones of Biochemical Test. The NeuroMolecular Computing Net (NMCN) model was used in this study. NMCN model is a bio-inspired operation mechanism. It has the ability of making message integration and collaboration in the neuron groups. NMCN model possesses an adaptive learning ability, and it can do self-organized search for solutions. The clinical data of Meridian Energy and Biochemical Test were processed through the NMCN model. The experimental results showed a significant correlation between Meridian Energy and Biochemical Test. This indicated that though there were different diagnosis ways for Chinese and Western medicine, highly similar results could be obtained from the same patient.

Keywords: Artificial NeuroMolecular Computing, Evolutionary Learning, Meridian Energy, Biochemical Detection, Ryodoraku.

1 Introduction

Chinese medicine has accumulated thousands years of experiences and developed its unique diagnosis methods. Traditionally, Chinese medicine had four ways for diagnosis, and they were inspection, auscultation and olfaction, inquiring, and palpation. Comparatively, Western medicine (modern medicine) emphasized scientific methods for diagnosis. It would use various medical devices, along with data-based biochemical tests, to verify the diseases. There have been many differences between Chinese and Western medicine for so long.

The Meridian Theory is an important realm of knowledge handed down in Chinese medicine. Chinese medicine doctors traditionally would understand a patient's physical state through pulse-taking and palpation, thus inferring the situation of his or her twelve meridians and the possible diseases [1,2]. In recent

* Corresponding author. Tel.:+88635381183x8291; fax:+88636102362.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 68–77, 2011.

decades, the study of Meridian Energy has been rapidly developed. Actually, it started from the Ryodoraku treatment in 1950s [3,4,5]. Ryodoraku indicates the meridian response points that are very conductive on human skin, and these response points have their specific distribution ways. After cautious investigation, we discovered that the distribution of Ryodoraku is almost the same as the distribution of human meridians. Therefore, the motive of studying more on meridian energy was aroused.

In recent years, the development of meridian energy devices has provided Chinese medicine doctors with another way of diagnosis. Due to the objective measurement of scientific devices, the meridian energy diagnosis is considered to be a more reliable reference. However, from the Western medicine perspective, this kind of measurement and test is still doubtful. Western medicine commonly depends on the data of biochemical tests in order for diagnosis. The biochemical test items would usually include the data of hemoglobin (Hb), total cholesterol (TC), blood urine nitrogen (BUN), creatinine (Cr), Albumin (Alb), and Glucose (Glu) etc. Nevertheless, these data are not totally dependable. Researchers have shown that the correlation between the data of fetoprotein (AFP) and accurate cancer prediction was about 60%. Although Western medicine would also use other examination items such as ultrasound to assist in cancer diagnosis, it still mostly relies on the result of AFP.

Although there were many differences between the Chinese and Western medicine diagnoses, the motives of both sides were the same, hoping to correctly diagnose a patient's diseases. Since both of meridian energy tests and biochemical tests could serve as the diagnosis reference, to the same patient there might be certain correlation between the two diagnosis methods. If the corresponding relations of the two could be found, it would contribute to the Western medicine doctors' understanding of meridian energy while providing Western medicine doctors with another diagnosis reference.

The purpose of this research was to see if there were any corresponding relations or prediction relations between meridian energy tests and biochemical tests. In other words, since Chinese medicine doctors could diagnose the abnormality of a patient's internal organs through the energy value of twelve meridians, there should be similar results of organ abnormality found in the biochemical tests. The research was hoping to find a certain corresponding relationship between the two.

2 Architecture of the NMCN Model

This research used the NeuroMolecular Computing Net (NMCN) model which is a bio-inspired NeuroMolecular operation mechanism. It includes the cytoskeletal molecular mechanism of nerve cells and the brain neuron memory mechanism. The NMCN model has good adaptivity and fault-tolerance ability, and it can self-organizedly search for solutions.

2.1 The Operation Hypotheses of the NMCN Model

- hypothesis 1: There are some brain neurons being in charge of the time-space information transition. This kind of neuron is called Cytoskeletal Neurons (CN). CN is based on the operation hypothesis between the nerve cell cytoskeletons and molecules, producing a time-space input signal and transducing it into a series of time outputs [6,7,8].
- hypothesis 2: There are some brain neurons being in charge of memory control and neuron group organization. This kind of neuron is called Reference Neurons (RN). The purpose of RN is to form a common-goal information processing group from CN. By the memory screening of RN, each workgroup would have neurons of different internal structures, thus being able to finish the group task [6,7,8].
- hypothesis 3: Regarding CN and RN, the brain would form a Neuron Computing Net for a specific task. This computing net would consist of CN and RN. CN is in charge of the main information processing while RN is controlling CN to form workgroups. This is called a NeuroMolecular Computing Net, or "Net" for short. A Net would be considered an independent problem-solving unit.

2.2 The Reference Neurons

Reference Neuron is motivated by some brain neurons' memory operation, and it has the ability to organize the cytoskeletal neurons to be a common-goal task group. The mechanism was from the firing-rekindling associative memory of the neurons. In our definition, when a RN synapse connected a CN, we called this a selective connection. The connection itself is a preparation process, and the purpose is for forming the next-step CN workgroup. When a RN firing happened, all the RN-connected-and-selected CN would be fired up as well. This process was called the rekindling. The firing CN would be in a workable state, waiting to cope with the input pattern information. As to the unselected CN, they would be in a resting state. Fig. 1 is the illustration of RN.

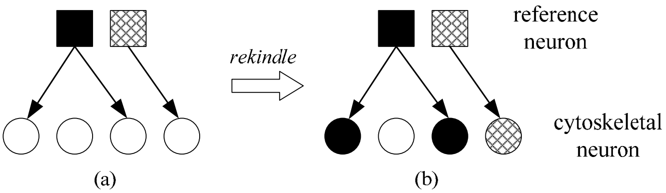


Fig. 1. Schematic illustration of reference neuron. (a) The synaptic connections between reference neuron and cytoskeletal neuron. (b) The firing of reference neuron will rekindle the cytoskeletal neurons.

A CN set chosen by RN was called a "Reference Working Group (RWG)" in this research. There could be several RWGs in a Net according to its task.

The CNs in the same RWG will work together for the group goal, and the CNs in different RWGs will compete with each other. Fig. 2 shows the structure of NMCN while explaining the relationship between CN, RN, and RWG. By the application of evolution technology, the arrangement of CNs within each RWG would be changed in each generation, and the purpose was to find suitable CN sets to achieve RWGs' goals.

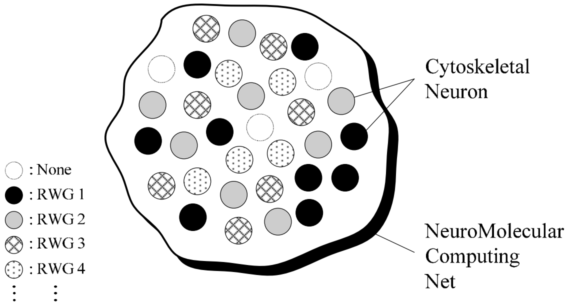


Fig. 2. Scheme of a neuromolecular computing net

2.3 The Cytoskeletal Neurons

When a neuron firing occurs, the internal dynamics of CN could transform the external input signals into another form of energy. This process was called ”transduction”; therefore, a CN could be considered a transducer with a specific structure. CNs are platforms of message processing, and they are inspired by the signal integration and memory function of the cytoskeleton.

This research utilized 2D Cellular Automata (CA) to conduct the experiment of CN, and the wraparound fashion links were adopted for the CA arrangement. A cytoskeleton has multiple molecule networks of microtubules, microfilaments, and neurofilaments. In order to simulate these networks, we defined three kinds of fibers to make a cytoskeleton-type (C-type), and the fibers were named C1, C2, and C3.

Each of the cytoskeletal elements will have its own shape, thus forming the cytoskeletal molecule networks. The conformation of each cytoskeletal element is variable; therefore, molecule-mass-like groups may possibly be formed. Different types of cytoskeletal elements have different signal transmission features. C1 has the strongest signal bearing capacity, but it has the slowest transmission speed. C3 has the weakest signal bearing capacity, but it has the fastest speed. C2's performance is between C1 and C3. The illustration of CN structure is shown in Fig. 3. Each CN has its unique cytoskeletal fiber structure. The types of signal flows depend on the different structures and different transmission characteristics. Some signal flow would execute the transduction tasks with a diffusion-like method, sometimes fast and sometimes slow.

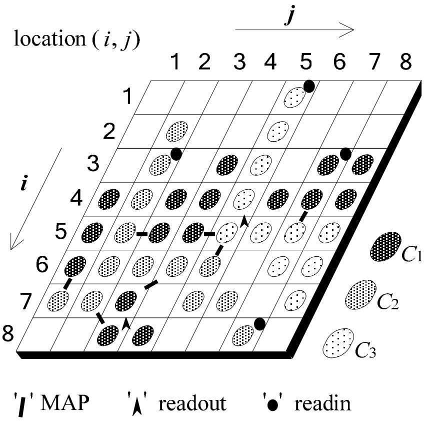


Fig. 3. Structure of cytoskeletal neuron

When an external stimulus hits a CN membrane, it will activate the readin enzyme at that location. The activation will cause a signal flow to transmit along the route of the same cytoskeletal elements. For example, after the on-location (3,2) readin received the external input, it will transmit the signals to its eight neighbors that have the same cytoskeletal element locations. The illustration shows that it can transmit the signal to C2 at locations (2,2) and (4,2). Any cytoskeletal element that receives this kind of signal will do the same, thus forming the phenomenon of a signal flow. In order to ensure it is a one-way transmission, meaning there will not be any signal backflow or loop formed, the cytoskeletal element will enter a temporal resting state after the transmission. This is called a refractory state.

The additional remark is that after a signal was transmitted by a cytoskeletal element, the signal did not disappear immediately within the element. Instead, the signal would decrease progressively until it finally disappeared. The decreasing signal and the new-coming signals would cause a time-space integration reaction, and that is a very important mechanism that decides when a firing will occur.

There could be some interactions among different cytoskeletal fibers. Microtubule associated proteins (MAPs) have the ability to connect different cytoskeletal fibers, thus causing cross-fiber signal flow channels. This will help the flow of micro-substances within neurons. For instance, when the input signal originated from location (3,2) goes along the C2 elements of the second column, it will meet a MAP-linked C1 element at location (5,2). The C2 signal will be transmitted to C1 through MAP, and another signal flow will be formed in C1.

However, due to different types of cytoskeletal fibers and different transmission features, there might be some energy transition problems when signals going through different mediums. Hence, regarding the cross-fiber signals, this research defined the signal bearing capacity of C1, C2, and C3 as S, I, and W, meaning Stong, Intermediate, and Weak. Because the linking function provided by

MAP allows the signals to flow among different molecule elements, there exist information processing behaviors within the neurons.

When a time-space integrated cytoskeletal signal arrives at a location of a readout enzyme, the activation will lead to a neuron firing. For example, the signal flows started at locations (1,5) and (8,7) may be integrated at location (5,5), and the readout enzyme at that location would be activated, thus causing a neuron firing. Because the integrated cytoskeletal signals may continuously appear, the firing outputs become a series of signals happened in different time points. This research collected these signals to serve as the reference for transduction efficiency assessments.

2.4 Evolutionary Learning

"Net" is the problem-solving core of the NMCN model, and the purpose of evolution learning was to find the right Net. A Net could be regarded as a computing unit, and there were 31 Nets for the population in this research. In evolution learning, there would be an initial population for the first generation. Each Net of the initial population would be produced randomly. Afterwards, each Net would learn and discriminate the patterns of the realm. The Net that could correctly discriminate the most patterns has the optimum fitness value, and that Net would be regarded as the new population to produce the next generation.

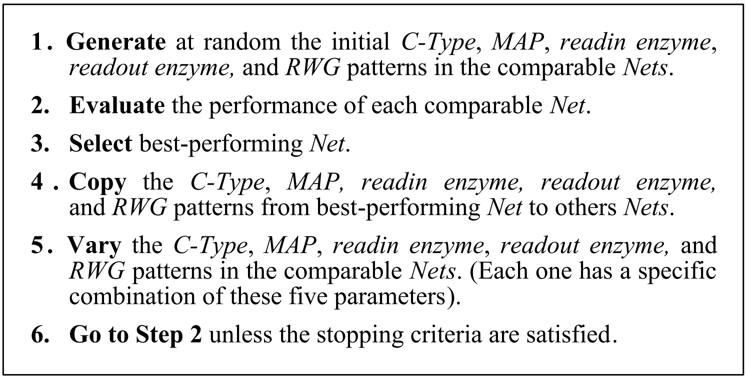
The new population would duplicate the same genes for the 31 Nets of the next generation. In order for variation, each Net would be mutated. According to the permutation and combination of the evolution parameters (C-type, MAP, Readout, Readin) and RWG, each Net would have a corresponding mutation strategy.

The variation Net would continuously learn and discriminate the patterns of the realm. If the fitness value of a certain Net is higher than the current Best Net, then the current one will be replaced. This process will be repeated until the system matches the learning termination conditions. Then, the Best Net at that time will be the solution that we wanted. Figure 4 is the algorithm of evolutionary learning.

3 Experimental Results

The clinical dataset used in this research was provided by a hospital in central Taiwan, and it included 364 instances. The dataset attributes could be divided into two kinds: input attributes and prediction attributes. The input attributes included two demographic attributes (gender and age) and 24 meridian energy attributes. The prediction attributes included 19 biochemical test attributes.

The NMCN model used in this research can only analyze binary data; therefore, the dataset must be coded into binary forms. In order to decrease the coding work, different input attributes had different digit coding methods. For example, two-digit coding was used in the gender item, and five-digit coding was used in



1. **Generate** at random the initial *C-Type*, *MAP*, *readin enzyme*, *readout enzyme,* and *RWG* patterns in the comparable *Nets*.
2. **Evaluate** the performance of each comparable *Net.*
3. **Select** best-performing *Net.*
4. **Copy** the *C-Type*, *MAP, readin enzyme, readout enzyme,* and *RWG* patterns from best-performing *Net* to others *Nets.*
5. **Vary** the *C-Type*, *MAP*, *readin enzyme*, *readout enzyme,* and *RWG* patterns in the comparable *Nets*. (Each one has a specific combination of these five parameters).
6. **Go to Step 2** unless the stopping criteria are satisfied.

Fig. 4. Algorithm of the Evolutionary Learning

each meridian energy value. There were 122 bits of coded input attributes. The prediction attributes were mainly for learning and prediction purposes, and each prediction attribute would tell if the participant were normal on the specific test. The normal items were marked as "0" while the abnormal ones were marked as "1".

3.1 Discrimination Learning

The NMCN model utilized the evolution learning method which needed more computing resources. Hence, this research only aimed at five prediction attributes to conduct experiments, and the attributes were GOT, GPT, Hb, TC, and UA. The purpose was to establish the discrimination model of these five attributes through the known instances, so we can understand the system's discrimination ability and its efficiency in the entire evolution learning progression.

The following is the experimental design methodology of this research: 364 instances of data were divided into the training group (243 instances) and the test group (121 instances). The discrimination model was based on the training and learning of the training group. The experiment aimed at five attributes and used the NMCN model to conduct the evolution learning. The individual learning curves of these five attributes are shown in Fig. 5.

The illustration shows the discrimination progression of each prediction attribute. Taking the GOT attribute as an example, the discrimination accuracy achieved more than 80% in the initial evolution stage. Along with the development of evolution learning, the discrimination accuracy was very close to 95% after 500 evolutionary generations. In order to control the experimental schedule of each attribute, the limit of evolutionary generation was set at 2,000. Once there were more than 2,000 evolutionary generations, the experiment would be terminated. When the GOT experiment was terminated after 2,000 generations, the accuracy achieved 95.1%.

After the experiments ended, the discrimination accuracy were GPT 90.5%, Hb 82.3%, TC 87.7%, and UA 85.6%. The results revealed that different

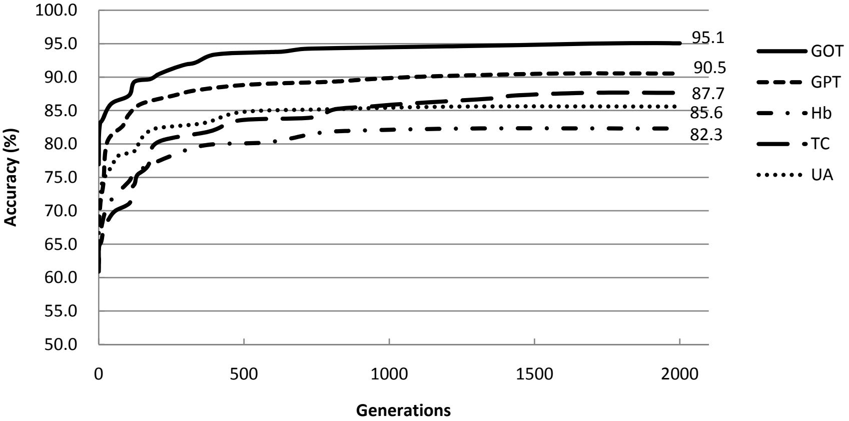


Fig. 5. Progression of evolutionary learning

prediction attributes would influence their corresponding relations with the input attributes, thus affecting the final discrimination rates. Besides the dataset factor, the NMCN evolutionary parameters might need to be adjusted in order to match the data properties.

In addition, if time were permitted, prolonging the learning of evolutionary generations would certainly increase the entire discrimination accuracy. Although there would be room for improvement in the system, the discrimination accuracy still achieved more than 88%, and it meant that the NMCN model had a certain degree of discrimination ability. The results also indicated that the NMCN model had the potential of continuous learning. It would adapt to the realm properties through the gradual structure-function changes, thus finding the optimum route.

3.2 The Forecast Experiment

This experiment mainly used the aforementioned discrimination model and predicted the corresponding relations between meridian energy and biochemical tests. The discrimination model aimed at the training group (243 instances of data) and conducted the training and learning program. The experiment put the test group (121 instances of data) into the discrimination model in order to test its prediction accuracy. Table 1 shows the comparison of discrimination accuracy and predictive accuracy of the five attributes.

The results showed that GOT had the highest prediction rate (92.6%) while Hb had the lowest rate (71.1%). From the data we could infer that increasing the discrimination accuracy will help increase the prediction rate. This inference could be applied to the five attributes as well. Namely, when the discrimination accuracy increased, the prediction rate would increase as well. Although this kind of relationship is not absolute, a certain correlation does exist.

In addition, as to the individual results, the GOT and GPT attributes had higher prediction rates while Hb, TC, and UA comparatively having lower rates.

There could be two reasons for the phenomenon. One was that the dataset itself had influenced the learning rate and prediction rate. The other reason was that there was not enough training and learning time or there were not enough evolutionary generations. From the past experiences, the more evolutionary generations there were, the higher efficiency the system would get. However, the experiment had to cease to a certain generation due to limited time.

From the results as a whole, the average discrimination accuracy could achieve more than 88%, and the prediction rate could achieve more than 80%. This kind of result has achieved a certain satisfactory degree. The purpose of this research was to discuss whether there was a corresponding or prediction relationship between Chinese meridian energy and Western biochemical tests. The experimental results showed that both of Chinese meridian energy and Western biochemical tests had an average of 80% prediction rate. This revealed that there existed significant correlations between the two. If the crucial factors of each individual attribute could be further researched, decisive and useful data should be found.

Table 1. Comparison of discrimination accuracy and predictive accuracy

	GOT	GPT	Hb	TC	UA
discrimination accuracy	95.1%	90.5%	82.3%	87.7%	85.6%
prediction accuracy	92.6%	87.6%	71.1%	75.2%	73.6%

4 Conclusions

Because there were differences between diagnosis methods of Chinese and Western medicine, both sides had different understandings on diagnosis results. This research hoped to use scientific methods to shorten the distance of both sides, thus helping each other to play a collaborative role of information exchange and complementary medicine.

The research method was to use our developed NMCN model, and it was a bio-computing mechanism that could self-organizedly find the solution for a problem. The model has been applied to others domain. This research selected five prediction attributes of biochemical tests from the known clinical dataset, along with the input attributes such as meridian energy values and demographic variables. By the NMCN model, discrimination and prediction experiments were conducted. The results revealed that the average discrimination accuracy could achieve more than 88%, and the GOT attribute even achieved more than 95% due to the individual attribute properties. If the evolutionary generations could be continued, the learning rate could still be increased.

As to the prediction experiments, the results showed that there existed significant correlations between Chinese meridian energy and Western biochemical tests. The prediction rates of both sides could achieve 80%, indicating that although they had different diagnosis methods, there could be a high degree of similarity regarding diagnosis results on the same patient.

Because the Meridian Energy Test is a non-invasion method, if a dependable relationship between meridian energy and biochemical tests could be proven in the future, this research could contribute to Western medicine as an initial reference before conducting the biochemical tests, thus reducing unnecessary waste of medical resources.

Acknowledgments. This research was supported in part by the Taiwan National Science Council (Grant NSC 98-2221-E-264-012)

References

1. Unschuld, P.U.: Huang Di Nei Jing Su Wen: Nature, Knowledge, Imagery in an Ancient Chinese Medical Text. University of California Press, Berkeley (2003)
2. Wiseman, N., Ellis, A.: Fundamentals of Chinese Medicine. Paradigm Publications, MA (1997)
3. Nakatani, Y.: Skin Electric Resistance and Ryodoraku. J. Auton. Nerve 6, 52 (1956)
4. Nakatani, Y.: A Guide for Application of Ryodoraku Autonomous Nerve Regulatory Therapy. Chans Books and Products, Alhambra (1972)
5. Nakatani, Y.: Ryodoraku acupuncture: A Guide for the Application of Ryodoraku Therapy: Electrical Acupuncture, a New Autonomic Nerve Regulating Therapy. Ryodoraku Research Institute, Japan (1977)
6. Conrad, M.: Complementary Molecular Models of Learning and Memory. BioSystems 8, 119–138 (1976)
7. Conrad, M.: Physics and Biology: toward a Unified Model. Appl. Math. Comput. 32(2), 75–102 (1989)
8. Chen, J.-C.: Computer Experiment on Evolutionary Learning in a Multilevel Neuromolecular Architecture. Ph.D. Dissertation, Department of Computer Science, Wayne State University, Detroit, U.S.A. (1993)

A Model of Knowledge Management System and Early Warning System (KMS@EWS) for Clinical Diagnostic Environment

Norzaliha Mohamad Noor, Rusli Abdullah, and Mohd Hasan Selamat

Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, Serdang, Selangor
Norzalihaster@gmail.com
{Rusli,Hasan}@fsktm.upm.edu.my

Abstract. Early warning system (EWS) is a technology to mitigate risk in multi disciplinary areas. Issues on timeliness for timely reporting are still regarded as a main challenge in EWS. Therefore in this paper, we suggest the model of the integration between knowledge management system (KMS) and EWS known as KMS@EWS for clinical diagnostics (CD) environment. The integration model is to combine the advantage of KM system with the EWS functionalities and its components. The proposed model is based on empirical study by using literatures on KMS and EWS. We synthesize the findings of KMS and EWS model of integration. To demonstrate the application of this model, we propose into CD environment as a platform KMS@EWS system implementation. Our propose model can provide early warning when any abnormalities or peculiar pattern of disease and symptoms arise and detected during the interaction between physicians and patients. Thus, this model can support the managing of data and information for timely reporting and detection by providing decision facilitation thru early warning during the CD processes.

Keywords: Knowledge management, Knowledge management system, Early warning, Early warning system, Clinical diagnostic environment.

1 Introduction

The field of KM has some significant process that can be integrated into and strengthen EWS in order to enable the knowledge flow in CD environment. CD environment which comprises main activities for physicians to approach patient to obtain a diagnosis of the disease and to select therapy [1] is used for implementation model of integrated KMS and EWS known as KMS@EWS. The idea of this environment as a platform for KMS@EWS system implementation is due to the activity during CD process. Each activity during CD process can provide early warning when any abnormalities or peculiar pattern of disease and symptoms arise and detected during the interaction between physicians and patients.

Reporting on disease outbreaks such as severe acute respiratory (SARS), influenza, influenza like illness (ILI), malaria and dengue have indicated the importance of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 78–91, 2011.

information and knowledge for controlling, management and mitigation of disease outbreaks. Early warning of disease outbreak is essential for preparation, tracking and monitoring an outbreak and minimizes associated morbidity and mortality [2]. Rapidly detection of disease outbreak on a timely scale is crucial in preventive and controlling the spread of infectious disease.

Even though extensive research and study has been carried out into providing effective detection, controlling and monitoring of disease outbreak, but still timeliness is a crucial factor as it will impact the decision for managing the outbreaks [3]. Several research studies have been made on EWS to detect disease outbreak rapidly [4] [5] [6] [7] [8] [9]. However, issues on providing a timely detection is still a main concern as it can impact any decision made regarding the prevention and preparedness pertaining the outbreak [6].

Therefore, we propose a model of KMS@EWS that can support the managing of data and information for timely reporting and detection by providing decision facilitation thru early warning during the CD process. In CD environment, both KMS and EWS functionalities are interconnected and affect each other. So, KM process, technologies and techniques are use to support the integration model between KMS and EWS by combining the advantage of KMS lifecycle and EWS functionalities. The motivation of this study is therefore to provide a model of integrated KMS and EWS in CD environment to proactively provide timely detection and response of decision facilitation during CD process. By providing early warning during CD process, practitioners, hospitals, health administration can be notified of outbreaks and potential outbreak situations.

This paper is organized as follows; Section 2 is a literature review on KM process, technologies and techniques, EWS functionalities and CD processes. Section 3 discusses methodology used to derive the model of KM@EWS. Subsequently, Section 4 will explain the integration model of KMS@EWS in clinical diagnostic environment and shows the mapping of KM process and technique to EWS functionalities. Finally Section 5 is the conclusion and discusses the future works of the study.

2 Literature Review

2.1 Knowledge Management System

There are many definition of KM and one of the definitions is by Chen, 2001 [10]: Knowledge management is the system and managerial approach to the gathering, management, use, analysis, sharing, and discovery of knowledge in an organization or a community in order to maximize performance. Thus, Abidi, 2001 [11] defined KM from healthcare perspective as the systematic creation, modeling, sharing, operationalization and translation of healthcare knowledge to improve the quality of patient care. Additionally, Satyadas, 2001 [12] viewed KM as a discipline that provides strategy, process and technology to share and leverage information and expertise that will increase level of understanding to move effectively solve problems and make decisions. Furthermore, KM is viewed as a process of turning data into information and forming information into knowledge that can lead to decision making. In which this process is subdivided into creating internal knowledge, acquiring external knowledge, storing knowledge in documents versus storing in routines, as well as updating the knowledge and sharing knowledge internally and externally [13].

Mean while, KMS is an IT-based system that supports the activities of creating, codifying and distributing knowledge within organization or communities [13]. There are four common activities of KM process namely knowledge acquisition, knowledge codification, knowledge dissemination and knowledge application [13] [14] [15] [16] [17]. Table 1 depicted the example of KM processes from the related studies which are based on four common activities as mentioned earlier. Knowledge acquisition relates to the creation and capturing of knowledge either from internal or external source and generally involve tacit knowledge. Knowledge codification or storage relates the converting knowledge into a tangible, explicit in order to be understood, maintained and improved. Knowledge dissemination means knowledge created and codified is ready for distribution via multiple delivery channels. Knowledge application refers to taking the shared knowledge and internalizing it.

Table 1. Example of KM Processes

Choo	Alavi	Zack	Bukowitz	McElroy
(1996)	(1999)	(1999)	(2000)	(2003)
Sense making	Acquisition	Acquisition	Get	Individual and group learning
Knowledge creation	Indexing	Refinement	Use	Knowledge claim validation
Decision making	Linking	Store/retrieve	Learn	Information acquisition
	Distributing	Distribution	Contribute	Knowledge validation
	Application	Presentation	Access	Knowledge integration
			Build/sustain	
			Divest	

KM is a concept on how to manage knowledge in terms of knowledge creation, codification, dissemination and application. So, from the technical perspective, KM can be viewed in terms of how this concept can be applied by using IT as a tool to facilitate knowledge sharing collaboratively. KM tools can be define as tools or technologies to support and facilitate knowledge sharing of the communities for the performance of application, activities or actions such as knowledge generation, knowledge codification and knowledge transfer [18].

Chua, 2004 [19] proposed a three-tiered KMS architecture that identified three distinct services supported by KM technologies: i) presentation services, ii) knowledge services and iii) infrastructures service. Kerschberg, 2001 [20] also pointed out the KM architecture and KM process model that could be used for knowledge capture, creation, distribution and sharing consist of three-tiered or layer of KM architecture namely i) knowledge presentation layer, ii) knowledge management layer and iii) data source layer.

2.2 Early Warning System

The early warning is referring to the term of information on an emerging dangerous circumstance which can enable action in advance to reduce risks involved. EWS can be regarded as an approach to mitigate risk that can be associated over many disciplinary areas for natural geophysical and biological hazards, complex socio-political emergencies, industrial hazards, public health risks and many other related risks. Additionally, Grasso, 2007 [21] looked EWS as an effective tool to disseminate timely information in order to mitigate risk of potentially catastrophic hazards for preventive action to be initiated. On the other hand, Austin, 2004 [22] argues EWS as a mean to obtain knowledge and to use that knowledge to assist in the mitigation of conflict. He emphasized that response to any conflict situation require knowledge to facilitate a common awareness regarding problems or risks and thus accelerate decision making with appropriate implementation action to deals with the risks.

The Hyogo conference in Japan, 2005 had emphasized that the effective implementation of EWS must constitutes four interacting components namely i) risk knowledge, ii) monitoring and warning service, iii) disseminate and communication and iv) response capability [23].

Several research studies have been made on EWS that are associated to natural hazard, aquaculture and disease outbreak. Disasters such as the Indian Ocean tsunami in 2004 and Katrina hurricane in 2005, highlighted inadequacies in early warning process towards disaster mitigation lead to the development of alert system known as GEAS [24]. Then, the risk of misdiagnosis, incorrectly treatment or over-treatment that exists after disease outbreak is a motivational factor to develop a knowledge-based EWS for fish disease [25]. While TeCoMed is the EWS telecommunication on medical events development was inspired due to the inadequacies of information system to confront the healthcare worker timely on communicable disease [26].

Generally, the successful implementation of EWS either in natural environmental hazard or in public health is depending on the four components as defined by ISDR. First component is risk knowledge which is identified as a risk assessment that relate to knowledge acquisition, analysis storage and manipulation. In second component, the monitoring and warning service deals with the technical capacity to monitor and forecast that can provide timely estimation of the potential risk faced by the communities. The dissemination and communication means delivering and distribute warning messages to alert the communities with a reliable, synthetic and simple message for preparedness action. Last component is the response capability which involves coordination, good governance for timely and appropriate action plan by authorities.

Meanwhile, from the public health perspective, Ebi and Schmier, 2005 [27] identified main components of EWS should include identification and forecasting of the event, prediction of the possible health outcomes, an effective and timely response plan and an ongoing evaluation of the systems and components. They highlighted that the EWS should be developed in collaboration with all relevant stakeholders in order to ensure issues of concern are identified and addressed.

For the purpose of this study, we reviewed and analyzed several previous research studies pertaining disease outbreaks which adopted the EWS approach. Most studies in EWS disease outbreak concerning weather monitoring and seasonal climate forecasting, epidemiological, social and environmental factors. There are numerous research studies that can be referred for successfully and effectively implement the EWS for disease outbreak. Hence, we came out with seven types of different research studies on EWS which provide early warning and concerning public health as summarized in table 2.

Table 2. Example of KM Processes

Author/ System	Functions	Activities
MiTAP [7]	- collect - analyze - distribute	- data source from news posting, web spider, emails, epidemiological reports - data filtered, translated and categorized - distribute via newsgroup, search engine, web server
ESSENCE II [4]	- collect - detection - alerting - notification - distribution	- data source from hospital emergency rooms, private practice groups, over counter pharmaceuticals, veterinary reports, school absenteeism, sales promotion, weather events and external surveillance system - detection of abnormal health condition - deliver alerts and surveillance via web-site - notify of special events or environmental conditions that trigger warrant changes in detection parameters - internet based distribution
RODS [5]	- collect - model - detection - alerting -display	- data source from Emergency Department - classifiy free text complaint into one of the syndromic categories - detection algorithm - when trigger alert, send email to users - display using geographical screens and temporal information screens
ProMED-mail [9]	- collect - review - verify - disseminate	- receipt information from email subscribers - review and filtering the information - document representation & selection - finalized reports and distribute to subscribers
Brownstein [8]	- acquire - display	- data source is the news acquired automatically every hour - collected data are aggregated and then overlaid in an interactive map
AEGIS [6]	- collect - modeling	- daily visit data from Emergency Department are collected - the historical models are constructed for total daily visit
	- detection - alerting - investigation	counts for each syndrome group - compares current visit levels with predicted models and generate alarms - disseminate visits and alarm information to users - investigate the alarm using authorized client agents
Automatic online new monitoring [28]	- collect - representation & selection - classification & evaluation	- web crawler to spider news articles from internet - news document are transformed - conduct online new classification task and performance are evaluated

2.3 Clinical Diagnostic Process

CD environment deals with interaction between patient and physician involves three main level known as history taking, physical examination, and other investigation [1]. History taking is the initial process in CD where the conversation between patient and physician is captured to obtain general idea of the patient's personality, the kind of disease and the degree of severity. Then, the physical examination is to identify the physical sign of the disease. During this process, skill and experience of the physician is crucial in eliciting sign of disease. Finally, investigation is to further perform a laboratory test or screening test to solve clinical problems and to complement the history taking and physical examination. CD process is knowledge driven and depends heavily on the medical knowledge, intuition, expertise and judgment to diagnose or prognosis and to determine appropriate treatment. Data is collected during a history taking, and information is obtained during the physical examination. Mean while, investigation by laboratory or screening test will also provide data and information for analysis and comes to a conclusion. Therefore, a clinical diagnostic environment can be used as a platform to facilitate EWS because each of the CD process can be used to promote early warning for disease outbreak.

3 Research Methodology

For this study, we are currently performing an initial study to propose the model of the integration between knowledge management system (KMS) and early warning system (EWS) known as KMS@EWS. We conduct a literature reviews (LR) that collect and understand the existing knowledge or information about the title to research. Compile and analyze related topic to learn from others knowledge, showing the paths of previous research and how the study can be related to it. The reviews and analysis are on academic journals, articles, books and information search on tools, technologies and techniques used that relate to the research. The LR is divided into three main steps as illustrated in figure 1. Step 1 is the model formulation which obtained information from the reviews and analysis of general information about KM, EWS, and disease outbreaks. In this step, definition and theories of KM, KMS, early warning, EWS and CD environment are analyzed and seek for common or identical functionalities between KM and EWS. Whereas, analysis on CD environment is to get an overview of CD process workflow in order to use as a platform for model implementation. Besides the definition, we also analyze the technologies and techniques used or required for the application of KM and EWS. We looked and identified the technical perspective requirement in terms of knowledge acquiring, codification and dissemination for the integration between KM and EWS. Step 2 is the development model of KMS@EWS for the CD environment. For the model development, we identified the integration components and provide its functionalities

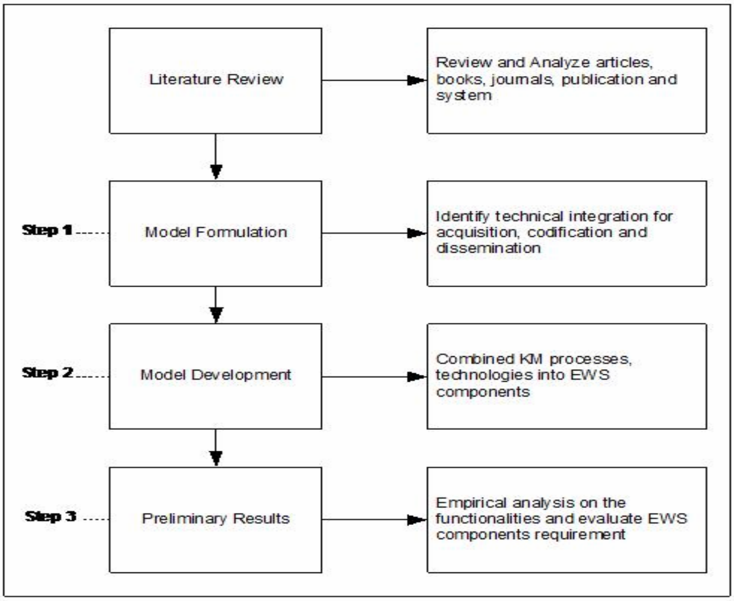

Literature Review
Review and Analyze articles, books, journals, publication and system
Step 1
Model Formulation
Identify technical integration for acquisition, codification and dissemination
Step 2
Model Development
Combined KM processes, technologies into EWS components
Step 3
Preliminary Results
Empirical analysis on the functionalities and evaluate EWS components requirement

Fig. 1. Research Methodology Diagram

based on the EWS framework reviewed earlier. Finally in step 3, we synthesize and map the analysis and findings of KMS and EWS to model the integration. We perform empirical analysis on the EWS functionalities in order to evaluate EWS components requirements. Then, we derived a model of the integration between knowledge management system (KMS) and early warning system (EWS) known as KMS@EWS for clinical diagnostics environment.

4 A Model of KMS and EWS for CD Environment

The concept of KM has been strongly supported as a process for acquiring, organizing and dissemination knowledge in order to promote effectiveness and efficiency [13]. Due to this concept, this study attempt to integrate the KM processes and technologies with EWS four main components. Therefore KM is a solution chosen to address the principles of building integrated model of KMS and EWS in CD process to provide proactive early warning for decision facilitation to react and respond on the outbreak. KMS@EWS will provide early warning when any abnormalities or peculiar pattern of disease and symptoms arise and detected during CD process.

4.1 Understanding KM in Relation to EWS

KM deals with how best to leverage knowledge involves the strategies and processes for identifying, capturing, structuring, sharing and applying knowledge in order to promote effectiveness and efficiency[11] [12] . While, EWS is a process of gathering, sharing and analyzing information to identify a threat or hazard timely and sufficiently in advance for preventive action to be initiated [22]. Hence, KM processes and tools can help strengthen the EWS to facilitate the acquisition, codification and dissemination of knowledge in order to identify and forecast the event, prediction of possible outcomes and timely reporting on the risks for immediate response. Derived from the ISDR guidelines, a complete and effective development of EWS should include 4 main interacting components as mention in Section 2.2. The purpose of these components is:

1. Risk Knowledge (RK): This is first component that plays a key function for data acquisition, analysis, storage and manipulation. A systematic data collection, analysis and information management can provide necessary information a quick and effective reference for response. Furthermore, the proper data handling and information management during the data acquisition is important for risk assessment. In relation to KM, this component is referring to knowledge acquisition with functionalities to collect, create, model, analyze, verify and filtering information.
2. Monitoring and warning service (M&WS): This component deals with a continuous monitoring to generate warnings in a timely fashion. The warning service is to detect abnormalities and able to provide forecasting and predicting when detected. The monitoring and warning service can be related to KM process for knowledge codification with functionalities to categorize, indexing, maintaining and retrieving information. Technologies such as case-based reasoning, rule-based system, data mining and patterning based recognition can be used to perform the forecasting and predicting to generate alert when any abnormalities detected.
3. Disseminate and communication (D&C): The KM process that related to this component is the knowledge dissemination with functionalities to distribute, display, delivering and transfer information to notify the communities. Good infrastructure can enable effective and efficient knowledge dissemination.
4. Response capability (RC): This component involves good coordination, governance for appropriate action plan by related parties and authorities. Additionally, awareness and education to the public on risk mitigation is also needed.

The purpose of these components can be combined into KM processes and technologies to provide a KMS environment as depicted in Figure 2. We propose the KMS environment with the concept for knowledge sharing system organize and distribute knowledge.

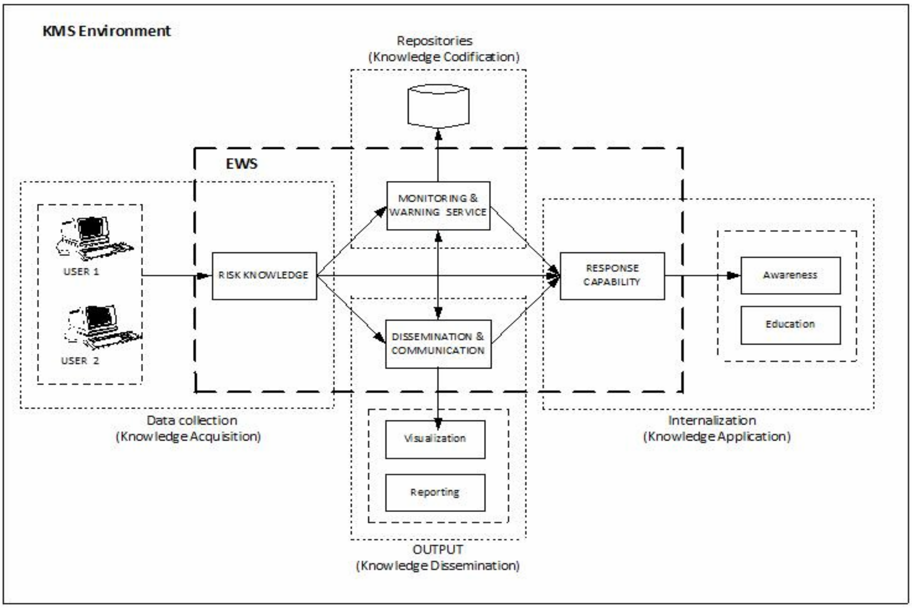


Fig. 2. EWS in KMS Environment Conceptual Diagram

4.2 Model Development

From the analysis on the EWS framework and definition, we identify main functionalities of EWS that can be combined to KM processes and technologies. Drawing from the three-tiered KM architecture from Chua [19] and Kersbergh [20], this study try to combine the three layers of KM architecture, to KM process, to EWS components, and to KM technologies as illustrated in Figure 3. Figure 4 depicted the model implementation of KMS@EWS in CD process. KMS@EWS will generate notification alert at every stage of CD process. Each of the notification alerts will indicate the severity of the risks.

Based on Figure 3, the purpose of three-tiered KM layer are:

1. Knowledge Presentation layer: provide interface to capture, acquire, sharing and disseminate knowledge pertaining disease outbreak to promote early warning. This layer also helps to stimulate communication and collaboration within the CD environment.
2. Knowledge Management layer: concerning the KM activities to acquire knowledge, to create knowledge, to store, codify and organize knowledge, to access and deploy knowledge and lastly to apply and use knowledge.
3. Data Source Layer: consists of various information and knowledge sources in a knowledge bases or repositories.

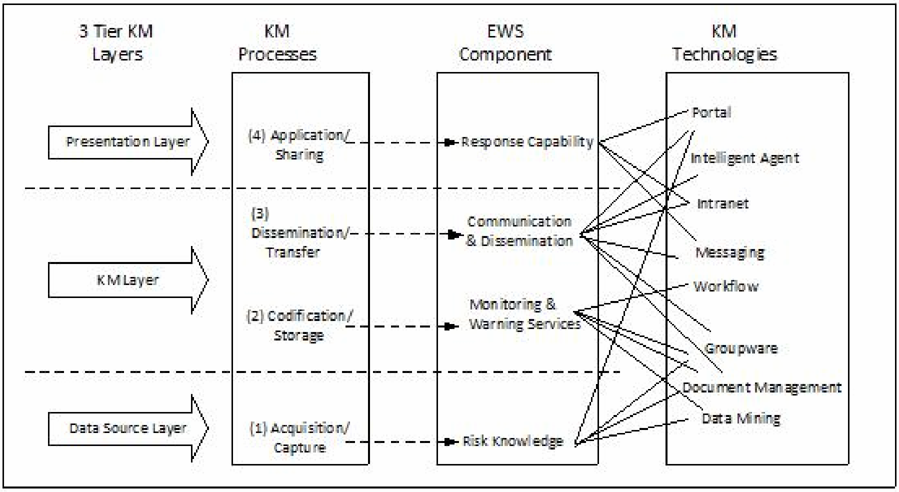


Fig. 3. Combination of KM Architecture and Its Technologies with EWS components

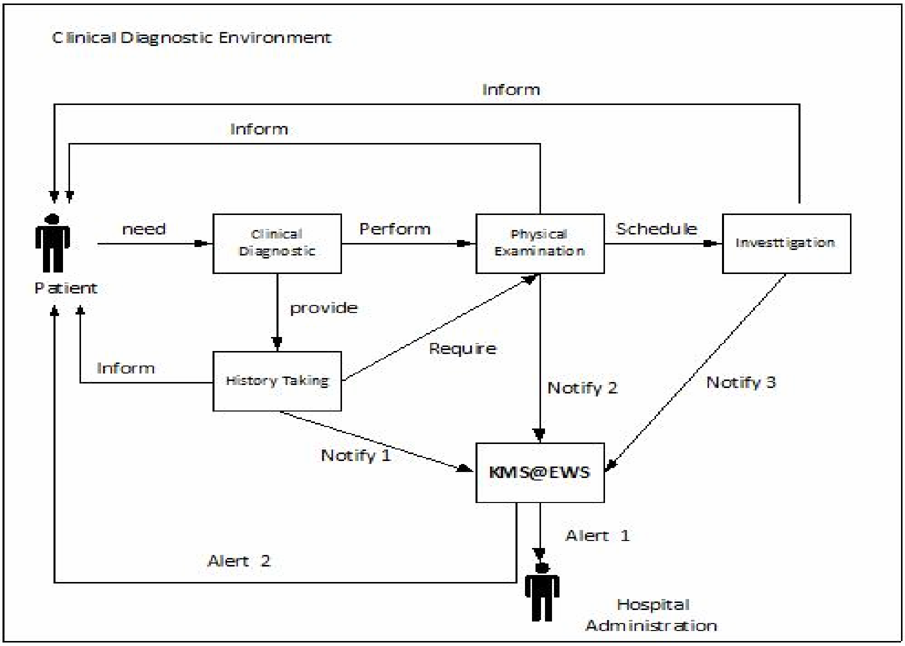


Fig. 4. A Proposed Model for the Implementation of KMS@EWS in CD Process

4.3 Preliminary Results and EWS Model and Its Functionalities

In order to synthesize the KM processes and EWS components, we analyzed the framework of previous research as shown in Section 2.2. Drawing from Table 2, we grouped the EWS functionalities to EWS four main components as in Table 3. The total and percentage from the table exhibit the overall requirements of EWS components.

Table 3. Matrix of EWS Functionalities and Components

Components / Functionalities	RK	M&WS	D&C	RC
Collect	/			
Review	/			
Verify	/			
Analyze	/			
Modelling	/			
Characterization		/		
Classification		/		
Evaluation		/		
Detection		/		
Prediction		/		
Alerting			/	
Notification				/
Investigation				/
Display			/	
Distribution			/	
Total	**5**	**5**	**3**	**2**
Percentage	**33.33**	**33.33**	**20.00**	**13.33**

Based on the matrix, we conduct an empirical analysis on the EWS to look for the most salient component to develop EWS. The result has shown that risk knowledge and monitoring & warning service are equally important (on average about 33 percent) for the EWS development. Both components are critical requirement for acquisition and codification to timely detect and forecast a potential risk. The communication & warning service with average of 20 percent is for delivering and distribute warning messages with the support of good infrastructure. The average 14 percent is the response capability which involves good coordination, governance for appropriate action plan. Figure 5 exhibit the EWS components requirements.

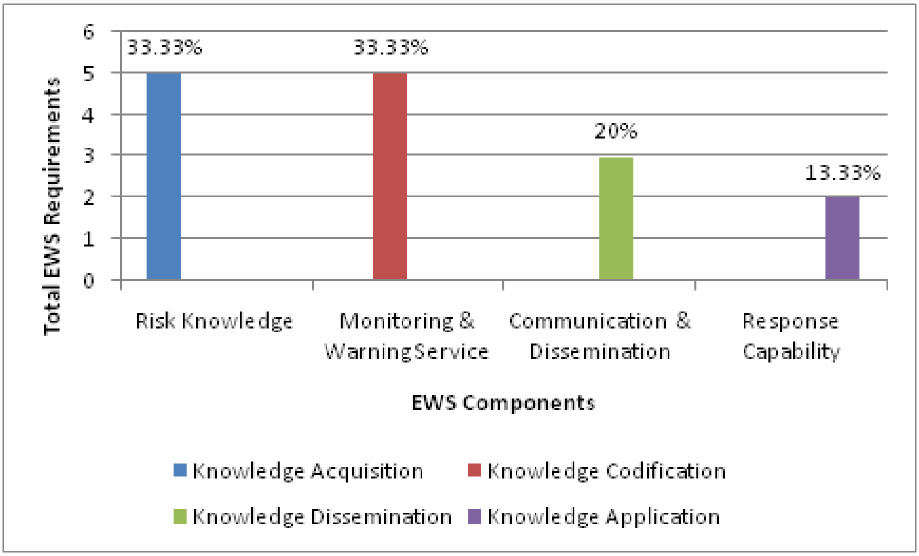


Fig. 5. EWS Components Requirements

5 Conclusion and Future Work

The field of KM has some significant process that can be integrated into and strengthen EWS to enable the knowledge flow in the CD environment. Response to any conflict situation needs knowledge to facilitate a common awareness regarding problems or risks and thus accelerate decision making with appropriate implementation of actions to deal with the problems or risks.

The adoption of technology is improving information access and dissemination, also is increasing the need to better understand how information can be managed and utilized to improve the performance of early warning systems.

To our knowledge which based on the literature reviews, there is currently no specific design of the integration model between KMS and EWS that combine KMS processes and technologies with EWS four main components. Therefore, this paper attempts to introduce the integration model which used CD environment as a platform for model implementation. In order to strengthen the integration and to demonstrate the knowledge flow within this model, we suggest the proposed integration model to be design as a web based EWS that can facilitate the clinical diagnostic. In which this model can demonstrate the main KM processes namely knowledge generation, knowledge codification and knowledge transfer. Furthermore, this model also intends to introduce tool based on Multi-agent System (MAS) to address the integration between KMS with EWS and also to assists the knowledge flow within the integration model in CD environment.

We also propose in future work to validate EWS components against research problems that been highlighted as well as research question in order to achieve the total set of research objective. The criteria of future validation include the reliability, suitability and accuracy in enhancing the capability of KMS@EWS.

It is also hope that the insight study of this integration model could offer a good perspective on the relation of KM processes and technologies with the EWS four main components.

References

1. Realdi, G., Previato, L., Vitturi, N.: Selection of diagnostic tests for clinical decision making and translation to a problem oriented medical record. Clinica Chimica Acta 393(1), 37–43 (2008)
2. Bravata, D., et al.: Systematic Review: Surveillance systems for early detection of bioterrorism-related diseases. Emerg. Infect. Dis. 10, 100–108 (2004)
3. Pavlin, J.A.: Investigation of disease outbreaks detected by syndromic surveillance systems. Journal of Urban Health: Bulletin of the New York Academy of Medicine 80(supplement 1), i107–i114 (2003)
4. Lombardo, J., et al.: A systems overview of the electronic surveillance system for the early notification of community-based epidemics (ESSENCE II). Journal of Urban Health: Bulletin of the New York Academy of Medicine 80(supplement 1), i32–i42 (2003)
5. Tsui, F., et al.: Technical description of RODS: a real-time public health surveillance system. Journal of the American Medical Informatics Association 10(5), 399 (2003)
6. Reis, B., et al.: AEGIS: a robust and scalable real-time public health surveillance system. British Medical Journal 14(5), 581 (2007)
7. Damianos, L., Zarrella, G., Hirschman, L.: The MiTAP System for Monitoring Reports of Disease Outbreak (2006)
8. Brownstein, J., Freifeld, C.: HealthMap: The development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill 12(48), 3322 (2007)
9. Madoff, L.C.: ProMED-Mail: An Early Warning System for Emerging Diseases. Clinical Infectious Diseases 39(2), 227–232 (2004)
10. Chen, H.: Knowledge management systems: a text mining perspective (2001)
11. Abidi, S.S.R.: Knowledge management in healthcare: towards [] knowledge-driven'decision-support services. International Journal of Medical Informatics 63(1-2), 5–18 (2001)
12. Satyadas, A., Harigopal, U., Cassaigne, N.P.: Knowledge management tutorial: an editorial overview. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 31(4), 429–437 (2001)
13. Alavi, M., Leidner, D.E.: Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS Quarterly 25(1), 107–136 (2001)
14. Choo, C.W.: The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions* 1. International Journal of Information Management 16(5), 329–340 (1996)
15. Zack, M.H.: Managing codified knowledge. Sloan Management Review 40(4), 45–58 (1999)

16. Bukowitz, W.R., Williams, R.L.: The Knowledge Management Fieldbook. Prentice Hall, London (2000)
17. McElroy, M.W.: The New Knowledge Management: Complexity, Learning, and Sustainable Innovation. Butterworth Heinemann, Boston (2003)
18. Rudy, L., Ruggles, I.: Knowledge Management Tools. Butterworth Heinemann, Boston (1997)
19. Chua, A.: Knowledge management system architecture: a bridge between KM consultants and technologists. International Journal of Information Management 24(1), 87–98 (2004)
20. Kerschberg, L.: Knowledge management in heterogeneous data warehouse environments. Data Warehousing and Knowledge Discovery, 1–10 (2001)
21. Grasso, V.F., Beck, J.L., Manfredi, G.: Automated decision procedure for earthquake early warning. Engineering Structures 29(12), 3455–3463 (2007)
22. Austin, A.: Early Warning and The Field: A Cargo Cult Science? (2004)
23. ISDR, Hyogo Framework for Action 2005-2015: Building the Resilience of Nations and Communities to Disasters (2005)
24. Grasso, V.F., Singh, A.: Global Environmental Alert Service (GEAS). Advances in Space Research 41(11), 1836–1852 (2008)
25. Li, N., et al.: Developing a knowledge-based early warning system for fish disease/health via water quality management. Expert Systems with Applications 36(3), 6500–6511 (2009)
26. Gierl, L., Schmidt, R.: Geomedical warning system against epidemics. International Journal of Hygiene and Environmental Health 208(4), 287–297 (2005)
27. Ebi, K.L., Schmier, J.K.: A stitch in time: improving public health early warning systems for extreme weather events. Epidemiologic Reviews 27(1), 115 (2005)
28. Zhang, Y., et al.: Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems 47(4), 508–517 (2009)

ICT Evaluation for Knowledge Sharing among Senior Citizens Community

Sharanjit Kaur Dhillon

Information System Department
College of Information Technology
Universiti Tenaga Nasional
Selangor, Malaysia
sharanjit@uniten.edu.my

Abstract. This study identifies problems faced by senior citizens while using computer and identify the most popular computer technologies used among this community. This paper formulates hypothesis based on a survey conducted in early 2010, where main issues were identified as requirements and opportunity of sharing knowledge among this community. Findings shows that specific computer technology requirement and suitable content of a web portal that suits senior citizens is essential so that the older generation who see computers and new technology as what they have the potential to be - a tool for expanding their horizons, learning new skills and finding new interests. Limitation of this paper is identifying the ICT requirements for senior citizens. For future research it is suggested that the study about the content of a web portal and the design guidelines analysis for the portal. This paper examines a proposed knowledge sharing framework to get it adapted for this community.

Keywords: Knowledge Sharing, Knowledge Management, Framework, Senior Citizen.

1 Introduction

ICT continues to be the best hope for developing countries to accelerate the development process there is an emerging need for all sectors of society to find ways to optimize the opportunities which information and communication technology presents. Knowledge and information produced should be shared and delivered fast and information technology must offer the solutions that are able to fulfill the requirements of the organization.

The concept of knowledge sharing arose when people found many challenges in managing knowledge. One of the key challenges in knowledge sharing is how to develop a culture of distributing knowledge within a community. Different individual have different views with the term of knowledge sharing and the perspective of the term. But, knowledge sharing means a commitment to inform, translate, and educate interested colleagues. It is an active listening and learning process, not a technology-driven panacea.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 92–103, 2011.

As an individual, all the senior citizens have their own knowledge, which they usually keep within the records of their mind. Besides, there is no specific platform for them to express what they have for sharing purposes. Due to that case, senior citizens are believed do not have a good interaction between other senior citizens and as a results, they do not gained anything that is stored inside the senior citizens' mind.

The objective of this paper is to identify the problems faced by senior citizens while using computer and to identify the most popular computer technologies used among the senior citizens. At present, there is no accessible platform for the senior citizens to contribute all their knowledge. The problem faced internally within them is how to encourage all the members and where do they have to place their knowledge. Most of the community members and more interested in keeping all their knowledge that they have without considering the importance of sharing the knowledge.

2 Literature Review

Generally, knowledge can be defined as the fact or condition of knowing something with familiarity gained through experience or association. The meaning of knowledge was adopted from Oxford Advance Learner Dictionary. Knowledge is best defined as actionable information-deeper, richer, and more expansive. Actionable implies when and where it is needed to make the right decision, and the right context (Tiwana, 2000). According to (Tiwana, 2000), there are two types of knowledge, which are tacit and explicit knowledge. As for other sources, (Nonaka, 1994) demonstrate the organizational complexities of attempting to manage the dynamic process of knowledge generation. He defines knowledge as possessing one of two main characteristics – tacit or explicit knowledge.

2.1 Knowledge Sharing

According to (Miller, 2002) sharing is a process whereby a resource is given by one party and received by another. For sharing to occur, there must be an exchange; a resource must pass between source and recipient. The term knowledge sharing implies the giving and receiving of information framed within a context by the knowledge of the source. What is received is the information framed by the knowledge of the recipient. Although based on the knowledge of the source, the knowledge received cannot be identical as the process of interpretation is subjective and is framed by our existing knowledge and our identity.

Sharing knowledge is one of the first cultural roadblocks we run into when implementing a KM project or program. The common recipe reads: "A corporate intranet with technology to allow people to create their own home pages encourages sharing. Real sharing implies opportunity for feedback, acceptance of critique, willingness to engage in deep dialog, and the expectation of reciprocity. Sharing requires a level of trust. It is a two way process and forms an integral part of relationship building.

Knowledge sharing means a commitment to inform, translate and educate interested colleagues. It is an active listening and learning process, not a technology-driven panacea. The key to sharing is helping the other party appreciate your context, which is difficult unless the context can be constrained. For example, within a community of practice, there may be agreement on a common language, or there may be sufficient context accumulated in the form of common experience and learning. Information sharing is not only about the technical aspects of work. Tasks, vision, values, goals, contacts, support, feelings, opinions, problems and questions are all part of the sharing experience.

2.2 Senior Citizen

When defined in an official context, "senior citizen" is often used for legal or policy-related reasons in determining who is eligible for certain benefits available to the age group. The term was apparently coined in 1938 during a political campaign. It has come into widespread use in recent decades in legislation, commerce, and common speech. Especially in less formal contexts, it is often abbreviated as "senior(s)", which is also used as an adjective.

The age which qualifies for senior citizen status varies widely. In governmental contexts it is usually associated with an age at which pensions or medical benefits for the elderly become available. In commercial contexts, where it may serve as a marketing device to attract customers, the age is often significantly lower.

In Malaysia, the standard retirement age is currently 58. The senior citizens though are defined as those aged 60 and above. This definition is according to the statement done at “World Assembly on Ageing 1982" at Vienna.

2.3 Common Worries about Technology among Senior Citizens

Some seniors take to the Internet very quickly. The discovery of a way to store numerous photos of family and friends must seem like a godsend. Then there's the convenience of being able to contact loved ones all over the world at a moment's notice. In remote areas Internet access can provide a window to the world.

According to (Twombhy, 2009) seniors with grandchildren or great-grandchildren often adapt better to using modern technology. This is especially true when children are physically available to teach them how to use things like computer. Kids delight in giving their grandparents enough skills to show up their parents. Grandparents appreciate the attention and family interaction.

Teaching technology to senior citizens takes a special kind of patience. Not only are you introducing them to new technology, but you have to bear in mind that these seniors may have problems of their own, which impair their learning skills. As stated by (Stein, 2008), if senior citizens are worried about using a computer, they may also not like mobile phones and other technological aids which could prove very useful on occasion.

The key to getting senior citizens to come to enjoy technology is to remove the fear - which comes about through a lack of understanding. It is advisable to teach them small skills at a time. Some of them , once they pluck up the courage to start, find it hard to stop but for many it is a question of small steps - dialing using a mobile phone and taking one call before moving on to texting and watching someone else send a simple email before having a go.

According to (Stein, 2008) libraries, clubs and small groups offer help and assistance to senior citizens to help them understand technology and these places also mean they meet similar people with the same worries.

Based on a survey that was conducted by (Jokisuu, 2007) with the purpose to explore the reasons that the elderly people have for not adopting various ICTs. The seniors without prior experience of computer usage had several reasons for their lack of use. A relatively large group of senior citizens in that survey stated that they had never used a computer whereas some of them even said that they were not interested in learning to use it either. The results also suggest that age, education and place of living are significant factors in determining whether an older person makes use of ICT.

Most of the non-users were in the older age groups. Another similarity among the non-users was the environment in which they lived: there were less computer-users in sparsely populated area than in towns and other more densely populated areas. In addition, gender did not have a significant effect on computer use.

2.4 ICT Difficulties among Senior Citizens

(Jokisuu, 2007) conducted a survey on a group of respondents to identify the problems with ICT among senior citizens. In this survey the senior citizens were asked to describe the problems they have encountered while using ICT. The responses concerned with technology in general or related specifically to computers, digital television, or mobile phones. Most of the senior citizens mentioned one or two problems that they had observed. Five general categories of problems were identified which is elements of technology, attributes of users, skill requirements, management of technology and technical problems.

2.5 Requirement of Sharing Knowledge among Senior Citizens

A survey was conducted by researcher early this year to identify the requirement and opportunity of sharing knowledge as well as experience using a web based portal as a platform to serve the senior citizens. A total 40 respondents were selected from various locations to analyze a local web based portal designed by Malaysian Government University.

Overall, most of the respondents agreed with the portal arrangement and they all agree that the arrangement of the portal is very pleasing. The icon usage of the portal was also very pleasing according to the respondents. The response for the color usage in the portal really attracts the respondents where most of them were very pleased with the selection of color for the fonts that are available at the portal. The respondents said that this will help those who are color blind.

The selection of the text size available in this portal was pleasing enough for the respondents that are having eyesight problem as they can increase the text size to see the text in the portal. As for the script that is used in the portal, the respondents agree with the usage of Malay language in the portal but they also suggested that the portal should be in English language too so that the web visitors can select from both the languages.

The portal is said to be user friendly and the instructions in the portal is easy to understand. As stated by most of the respondents that it is easy to understand the instructions in the portal because it is in the Malay language and the respondents do understand the language well and the language level used in the portal in not very difficult to understand.

The links that are available in this portal are said to be useful and the content of the portal is said to be appropriate for the targeted community as the links that are available in the website are useful for the community. Very less of the respondents agreed with the quality of the content of the portal. The same reason was given as before that the language should also be available in English.

The information proposed in the website is said to be very little and less helpful for the targeted community because many of them said that the information is less related to them. They were expecting information on financial aids, healthcare and also on facilities provided by government for this community. Same respond was given for the facts and abbreviations used in the portal.

The proposed categories in the portal are said not to be appropriate because too many categories make this community confuse. As the speed of this system is slow, this makes the respondents do not like to surf the portal for long as they have to wait for the page in the portal to load for a long time.

Very less navigation facilities are available in this portal and many of the respondents were not pleased with the usage of the site map available in the portal. They said that the site map was not very useful for them. Lastly, the respondents think that the portal does have some assistive technique for disability navigation that can be useful for the less fortunate users of the site.

3 Methodology

The researcher conducted a survey to gather information and identify the facilities and environment for knowledge sharing in order to develop a web based portal to serve the senior citizens of Johor Bahru, Malaysia. Besides that, the analysis was also done to identify a knowledge sharing framework as a guideline for the portal development.

There are varieties of framework to be treated as guideline in a project. Thus, researcher has looked through many frameworks that are available in the internet and has decided to use framework which was proposed by (Aida, 2009) from the Information System Department of Faculty of Computer Science and Information System in this university. Figure 1 show her framework made for the developing of knowledge portal for the special children needs.

However, as researcher can see the scope and the community of practice is different, so the step taken to make some changes not in the mine body of the framework, but in the framework terms and omitting some parts that is not going to be used in the project or they may be out of the project scope.

The data collection of this study was referred through the extensive reading. Literature research is important to obtain depth understanding about the research's topic. Two ways of conducting the literature research were carried out, which is online research (via internet) and offline research which was conducted at various locations in Johor Bharu and Ipoh. Before the detailed study was done, early observation was used to collect data from the research study. Analysis of the interview as well as the survey that was conducted with the respondents from various places in Johor Bahru city in Johor has been used to determine the content and the design of the website for senior citizens.

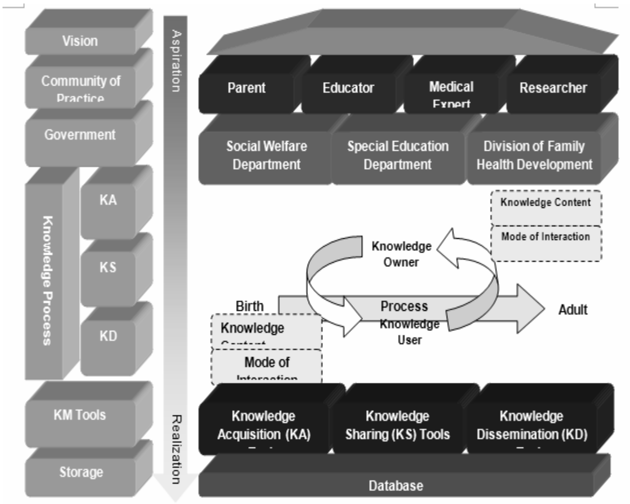


Fig. 1. Proposed Knowledge Sharing Framework (Version 2.0). Framework to develop knowledge portal for the special children needs.

The adaptation of the framework above was analyzed to suit the community of practice of this study. Researcher concentrates more on the knowledge process area of the framework whereby the senior citizens are above the age of 60 until death. They are also the knowledge owners as well as the knowledge users but this does not limit to only senior citizens. The family members of the elderly people, the researchers as well as the medical experts can be categorized in the process of knowledge sharing among the senior citizens.

3.1 Description of Knowledge Sharing Framework v2.0

This framework is divided into three compartments: Compartment 1 at the left side, Compartment 2 at the right side and main arrow between those two compartments. Compartment 1 is comprises of six (6) main components. The components are Vision, Community of Practice (CoPs), Government, Knowledge Process, KM Tools and Storage. Compartment 2 is the content of Compartment 1.

3.2 Vision

“To create effective knowledge sharing practice through K-Sharing”.

3.3 Community of Practice

The author identified 4 community of practice in special children context. The communities are Parents, Educators, Medical Experts and Researchers. The identification of these CoPs is due to their direct involvement in special children life cycle. These communities are communicating among each other for the good of special children.

3.4 Government

This component identifies the government bodies directly linked with the community of practice in aiding the community.

3.5 Knowledge Process

In this component, knowledge process is divided into three phases i.e. Knowledge Acquisition, Knowledge Sharing and Knowledge Dissemination. This study is focusing on knowledge sharing. However, the other two processes were identified as important process which may assist knowledge sharing activity. It is the iterative process. In order to share knowledge, they have to realize what knowledge they wish to share. Thus, knowledge acquisition process needs to be done.

3.6 KM Tools

This component becomes the most important component between the others. It is due to the main objectives of this framework which is to provide the most suitable knowledge sharing tools for certain situations. This component is comprises of three main knowledge management tools namely as: Knowledge Acquisition Tools, Knowledge Sharing Tools and Knowledge Dissemination tools. However, this framework is focusing on knowledge sharing aspect.

3.7 Storage

Database

4 Results

As mentioned earlier that a survey was conducted to gather information and identify the facilities and environment for knowledge sharing in order to develop a web based portal to serve the senior citizens at Johor Bahru. A total of 40 respondents were given survey forms to fill in. The survey was conducted to gather information, feedback and views from the respondents regarding the matters that are being researched. The research also takes into consideration the design of a web portal taking senior citizens in mind.

4.1 Reasons for Not Using Computer

The key reasons stated by most of the senior citizens for not using the computer are shown in Figure 2. The most common reasons for non-access are the categories ‘‘no computer at home’’ and ‘‘not interested’’. In-home computer availability seems to be less important. The two motives may be categorized indicating a lack of means and motivational indifference. Lack of means is a smaller and motivational indifference is a larger issue within the young and old senior population, compared to the middle-aged population.

For the option ‘‘PC is too expensive’’ only 7% of the respondents ticked this option. The reason may be twofold. For instance, financial concerns seem to be less crucial for senior citizens compared to middle-aged people. On the other hand, it may simply be the case that a lack of knowledge about the computer also includes a lack of knowledge about its price.

The researcher also found pronounced for people ticking that they ‘‘do not know what it is”. Thus, a minority would prefer to use the computer but cannot afford it. A majority does not invest in computer technology because they are not interested or lack knowledge about it.

Another important category for not using the computer is ‘‘missing skills’’. Nearly 78% of the non-using senior citizens in this region perhaps might be persuaded to use a computer if they were instructed on how to use it.

With regard to the aging process in Johor this is an alarming percentage. The same holds for about 12% of the respondents who ticked the next category ‘‘I do not know what it is’’. The reason ‘‘too complicated’’ is ticked by almost 49% of the non-users.

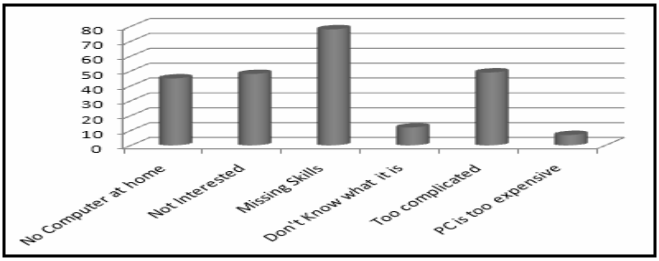


Fig. 2. Reasons for senior citizens not using a computer

4.2 Age-Related Functional Limitations

Getting older can result in several problems such as vision issues. These include decreasing ability to focus on near tasks, changes in color perception and sensitivity: blues/greens become harder to see then reds/yellows and dark blue/black can be indistinguishable, reduction in contrast sensitivity as well as reduction in visual field. From the survey, almost 91% of the respondents are vision impaired is having trouble in reading without glasses most of the time.

Besides vision, hearing loss is also a common problem with the older age people. 43% of the elderly people are having hearing loss and not all of them are using hearing aid. Motor Skill impairment is also identified as another problem where seniors are merely to have Parkinson's – Tremor, rigidity, slow movement, impaired balance and co-ordination and Arthritis (the leading cause of disability in those over 55).

These age people also prone to Cognitive decline where there will be decline in the ability to encode new memories of facts and decline in working memory. Aging may affect memory by changing the way the brain stores information and by making it harder to recall stored information. This will effect on either short-term or long-term memory loss. 27% of the respondents agreed that they are affected with cognitive decline.

4.3 Common Computer Based Technologies among Senior Citizens

Analysis was done to identify what computer based technologies are the respondents mostly familiar with. Besides that, there were questions asked to identify the difficulties they face when using a computer and mostly of the respondents gave the same answer. Figure 3 illustrates the percentages difficulties faced by senior citizens while using a computer. The respondents respond that the monitor screen causes very bad eyesight problem when they sit in front of it. Besides that, the small key on the keyboard are also difficult to see and hit on as responded by them. Fonts of the text on the websites are difficult to see for this community as majority of them has eyesight problems at their age. Sound element is among the least difficulties faced by this community as they are not concern about this issue much.

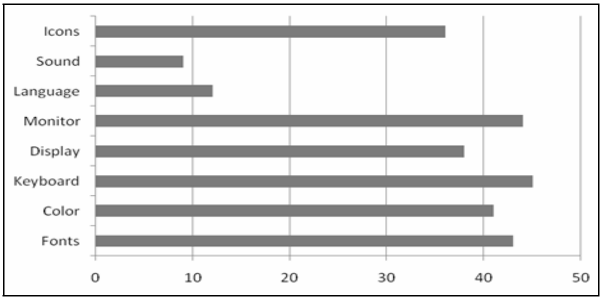


Fig. 3. Major Difficulties Faced By Respondents When Using Computer

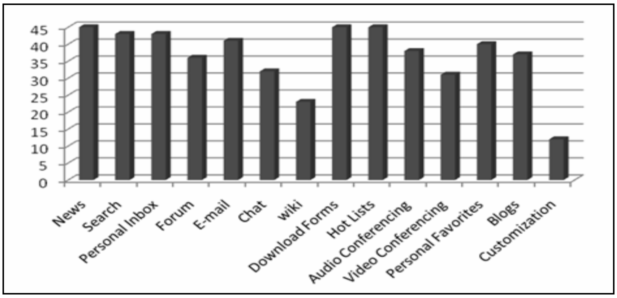


Fig. 4. Knowledge Sharing Portal Features from Respondents Perspectives

Figure 4 shows the percentage of knowledge sharing portal features from the respondent's perspective that was identified among the respondents of the survey. Downloading forms such as forms related to EPF, SOCSO as well as bank loans is the top feature required by the senior citizens. This community is not fussy in getting their webpage customized as they want; therefore customization is the least selected feature chose by the respondents.

4.4 Website Design Requirements for Senior Citizens

There are a few criteria that need to be taken into consideration when designing a website for senior citizens. The most common question that this group of age will ask when visits any website in the Internet are: Are the links clear? Is the text guiding the user to the wrong place? Is the typeface too small?

The researcher identified a checklist of ways Web designers can address the visual and cognitive disabilities that many seniors live with. They include building sites with large, plain typefaces; avoiding the juxtaposition of yellow, blue and green, a color combination that can be difficult to discriminate; and keeping text simple.

It is important to avoid technical jargon at all cost. However, if you employ newer functionality such as tagging for example, don't try to rename it, but provide an easy-to-understand explanation for it. Include instructions in plain English or Bahasa Melayu where necessary and always try to reduce the number of words displayed on the page.

Use simple and short sentences and include bullet points where possible. For links on the homepage or landing pages include a short description to tell site visitors what to expect when following the link.

Buttons must also be made as large and prominent as possible so they become a clear call to action. 3D effects for buttons can help to make them stand out. Also, make links and buttons easy to target and hit by increasing their clickable area.

A dropdown menu can be fiddly and time consuming for site visitors, and can result in people selecting the wrong item by accident. If you have less than 10 items in a dropdown menu use radio buttons if possible. These have the advantage of showing the number of options at a glance without having to click.

A site map gives users a good overall picture of how the site is organized and clearly defines all the resources the website has to offer. The link to the site map can usually be found near the top or the bottom of the page and frequently placed near the link to 'contact us'. Internet savvy senior surfers are aware of site maps and make use of them to gain an overview of the site**.** They will also likely click on a sitemap link when they get lost on the site or if they can't find what they want while browsing.

Web adaptation technology that allows users to personalize their Web interface by altering colors, size, and spacing, as well as turning off animation is very helpful to this community of practice. The technology also can convert text to speech, and eliminate repeated keystrokes caused by hand tremors.

Active phrases should be used in websites -- "view accounts," for example, rather than just "accounts". Besides that features that provide a short pop-up description of where each link will take the user can also be included. Users can opt to have those descriptions read aloud. Seniors also sometimes have trouble finding links. One solution to consider is color change on each link that is already been visited by the user.

Lastly, it is necessary to make the website trustworthy. Senior surfers tend to be more cautious when browsing and can get confused when something unexpected happens such as a new window opening or an application installing.

Firstly, clearly state the purpose of your site on the homepage. Also, offer a brief description with content links, so users know what to expect when following them. Explain in 'large print' how personal information will be handled before asking users to enter it. Make use of the well-known padlock icon to indicate a secure part of the site. Show words such as 'secure', 'safe' and 'confidential' in bold. Offer a content section on security when your site offers financial services.

5 Conclusion

Computers are becoming pervasive throughout society. Since several years, a trend towards an increased distribution of vital information via the Internet can be observed and this trend is unlikely to stop in the near future. With the current work the researcher can understand why older citizens do not use the computer or other ICT technologies. Additionally, some answers are sought as to how individual socioeconomic background determines the likelihood of the technology usage among the older adults.

Overall, the analysis done provides a root for the researcher to get preliminary understanding of the problems that are faced by senior citizens when using a computer as well as in identifying the ICT requirements that suits this community. The proposed knowledge sharing framework can be adapted and changes can be done to suit the community of senior citizens. The researcher hopes to look into the criteria of contents in a website especially for the senior citizens and design the guidelines to be followed to be used for the next study. Applying web accessibility for this community of practice should be taken into consideration and further research should be included on this criterion for the next study as well.

References

1. Suzana, A.: The Selection of Knowledge Sharing Tools for Special Children Community. Universiti Teknologi Malaysia Skudai Johor (2009)
2. Amrit, T.: The Knowledge Management Toolkit. Prentice Hall, Upper Saddle River (2000)
3. Emily, S.: Commentary: Is it all about aging? Impact of technology on successful aging, pp. 28–41. Springer, New York (2005)
4. Hesh, J.: ICT Problems among Senior Citizens. Prentice Hall, Englewood Cliffs (2007)
5. Martin, M.: Culture as an Issue in Knowledge Sharing: A Means of Competitive Advantage. University of Luton, UK (2007)
6. National Miller, W.C.: Fostering Intellectual Capital. HR Focus 755(1), 9–10 (2002)
7. Nonaka, I., Takeuchi, H.: The Knowledge Creating Company. Oxford University Press, Oxford (1994)
8. Pew, A.: Problems of collaboration and knowledge transfer in global cooperative ventures. Organization Studies 18(6), 973–996 (2007)

9. Phyllis, T.: The Social Life of Information. Harvard Business School Press, Boston (2009)
10. Stein, S.: Leveraging tacit organizational knowledge among senior citizens. Journal of Management Information Systems, 9–24 (2008)
11. Tullis, A.: Collection and connection: The anatomy of knowledge sharing in professional service. Social Development Journal, 61–72 (2007)

Subthreshold SRAM Designs for Cryptography Security Computations

Adnan Abdul-Aziz Gutub[1,2]

[1] Center of Research Excellence in Hajj and Omrah, Umm Al-Qura
University, Makkah 21955, P. O. Box 6287, Saudi Arabia
[2] Associate Researcher at Center Of Excellence in Information Assurance (CoEIA),
King Saud University, Riyadh, Saudi Arabia
aagutub@uqu.edu.sa

Abstract. Cryptography and Security hardware designing is in continues need for efficient power utilization which is previously achieved by giving a range of trade-off between speed and power consumption. This paper presents the idea of considering subthreshold SRAM memory modules to gain ultra-low-power capable systems. The paper proposes modifying available crypto security hardware architectures to reconfigurable domain-specific SRAM memory designs. Although reliability is still a problem, we focus on the idea to design flexible crypto hardware to gain the speed as well as the reduced power consumption.

Keywords: Cryptography hardware, Subthreshold SRAM, Low-power architecture, Efficient crypto computation. Security arithmetic signal processing.

1 Introduction

Saving Power is becoming a target for most modern cryptographic computations hardware designs especially with the rapid increase in performance and transistor count [1,2]. The prediction relating power consumption with technology advancement and Crypto-key size increase is that "power consumption would increase quadratically every technology generation" [3]. Although this prediction is changing but still the power consumption is becoming a real problem. In 1980's, the power consumption increase was reported approximately 30% every year. However, this pace reduced in the 1990's to around 13% per year. Lately, it was found that this rate kept holding until nowadays and the power consumption per processors exceed the 100 watt [4]. Reliability [5], is becoming a parameter reported to affect the balance of performance and energy utilization. This presentation will consider reliability briefly later in this work.

It is known that efficiency of hardware power consumption cannot anymore depend on device technology and circuit optimization alone. Computer architecture and electronics engineering are also involved in providing new solutions to the increasing power utilization problems [6]. Furthermore, the development cost of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 104–110, 2011.

system design is increasing due to crypto-system complexity, where hardware modeling and verifications is becoming increasingly difficult and time consuming [3]. In fact, the analysis of power and performance at early stages of hardware designing is necessary to avoid starting again every time [7].

Proper cryptography hardware designing goes all the way from the top-level where structured or behavioral hardware description is given passing by circuit optimizations at logical level or gate level, down to semiconductor devices and their technology. All these design levels need to explore low power design methods independently, so that the complete crypto hardware system could benefit from the total power efficiency gained. Many technology tools have been developed for industrial general designing purposes, however, not many of them are acknowledged for authentic academic research. Accordingly, power estimation studies at architecture level are becoming a more important research subject [3,7].

2 CMOS SRAM for Crypto Designing

Cryptographic hardware normally faces the problem of power consumption [8], which is beveled to be efficiently considered when involved in the designing phase. The power consumption of crypto memory, i.e. CMOS circuits' as an example, is known to be affected by two components, namely the subthreshold leakage and the dynamic (charging/discharging) factors [4]. As the technology is improving, the supply voltage, VDD, is decreasing, affecting the threshold voltage, VTH, to decrease too. To keep the crypto-computation circuit performance (speed) to a certain practical level with lowering VDD and VTH, the subthreshold leakage component is involved heavily [4]. REBEL [9] is an example. It is a network based cryptography (block encryption) function which uses reconfigurable gates instead of substitution boxes.

REBEL hardware approach had the advantage of the key size that can be much greater than the block size, with its security to be reduced to Boolean square root problem. REBEL design also showed resistant to known cryptanalytic attacks. The hardware of REBEL model compared between ASIC and FPGA implementations to evaluate its area, power and throughput. Relating REBEL to the SRAM focus, REBEL used two methods to store the crypto key, i.e. registers as well as SRAM. Interestingly the SRAM key storage should high efficiency, where the hardware area decreased and the computation throughput increased [9].

Generally, the subthreshold CMOS transistors in crypto hardware design and operation is getting progressively more. important due to their essentiality in portable small devices (i.e. notebooks, mobiles, smartcards…etc) [10], and all low power applications (i.e. Encryption chips on smart-credit cards, wireless sensor nodes, bio-informatics, security surveillance, medical and health examining, industrial monitoring …etc) [11,7,4].

Operating the transistor in subthreshold is generally based on its leakage current, which is applicable whenever the hardware is compact and does not involve in intensive computation because of the subthreshold natural performance degradation [12]. One of the best hardware modules to operate in this subthreshold transistor region is the static random access memory (SRAM), which is known with its low standby leakage current [10]. In fact, low-power SRAM designs is becoming

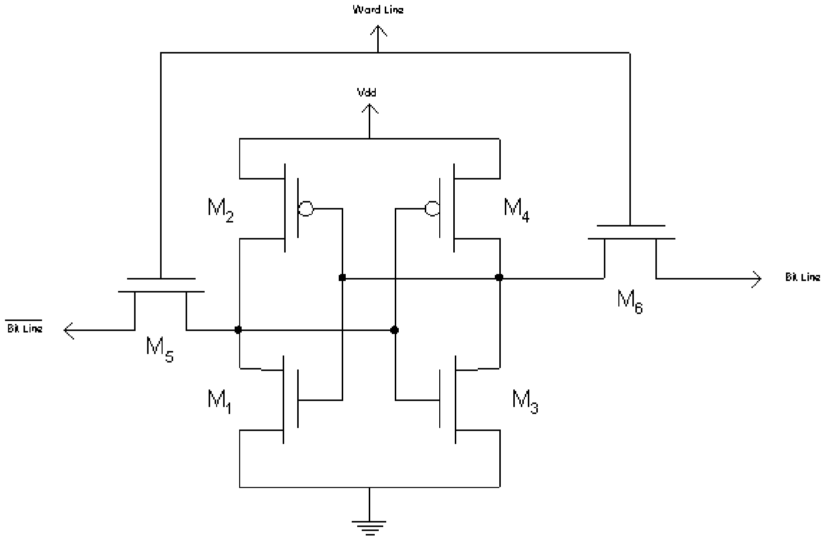


Fig. 1. Standard 6T SRAM Cell

essential, since 95% of the area in the system on chip designs is expected to be consumed by memory units [4].

The conventional SRAM cell is made of six-transistor (6T SRAM) as shown in Figure 1. This 6T SRAM cell is fast compared to DRAM but suffers high power consumption making it unpractical for future low energy application needs. This 6T SRAMs as is, is having difficulty in adjusting to the rising requirement for larger and larger memory capacity applications [10].

In response to this low energy memory requirement, researchers are trying to develop an SRAM cell operating with subthreshold transistors to reduce the cell power consumption.

3 SRAM Potential

Power reduction of SRAM memory cells can be performed by maintaining the standard 6T SRAM cell as is and changing the voltages or transistors sizes, or by modifying the SRAM cell transistors design it self [11]. Changing the voltages method is mainly performed in two different ways; one with increasing V_{DD} and V_{TH}, and the other with controlling any of the cell voltages, i.e. V_{DD}, V_{SS}, V_{GG} or V_{BB}, of the SRAM cell. Increasing the voltages V_{DD} and V_{TH} (shifting the voltage swing) benefits in increasing the speed of the cell, which will naturally reduce the leakage power consumption. "This approach, however, is not scalable in a long run, since we cannot use miniaturized devices with high V_{DD}" [4].

On the other hand, reducing the power consumption through controlling one of the SRAM voltages is preferred to benefit from cutting off the supply voltage of a cell when it is un-selected [4]. For example, D. Ho. in [11] presented a comparison study to decrease the power consumption through reducing the leakage current in the standard 6T SRAM but on 90nm technology scale. Several power reduction techniques have been investigated, such as scaling the supply voltage, sizing transistor gate length, and implementing sleep transistors. Scaling supply voltages gave good efficiency in power consumption but degraded the stability of the SRAM cell tremendously. However, transistor sizing and adding a sleep transistor before connecting the cell to ground gave interesting promising results, as shown in Figure 2.

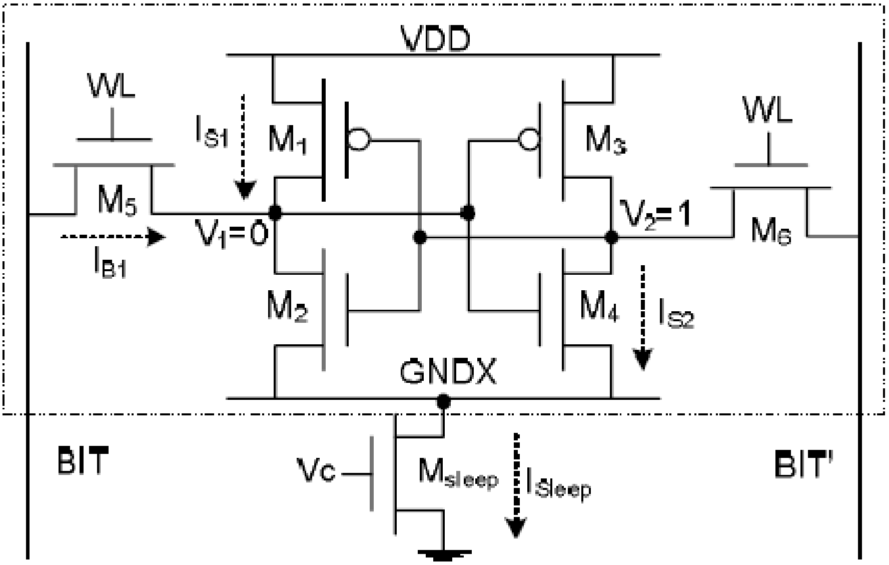


Fig. 2. Standard 6T SRAM Cell with the addition of Sleep Transistor [11]

Note that the study in [11] did not consider the SRAM cell speed which is expected to be affected accordingly. Several attempts have been proposed to modify the 6T SRAM cell transistor structure to gain different benefits. Some SRAM designs tune the transistor sizes [11].

Several others change the standard design number of transistors and invent new structure [10] or add power efficiency transistors [11]. For example, Arash Mazreah in [10] proposed an SRAM cell with 4 transistors (4T SRAM, as shown in Figure 3) with same design rules of the standard 6T SRAM. The main aim of their 4T SRAM is to reduce the cell size claiming to reduce the power consumption. The memory reading and writing operation of data in this 4T design is not performed normally; i.e. the reading is performed from one side while the writing is from the other. The power consumption reduction is gained from lowering the swing voltages on the word lines, making the operation need of different voltage levels, which is unpractical to most reliable VLSI designs [4]. This may also affect the reliability [5], which can be a serious concern in crypto applications as described next.

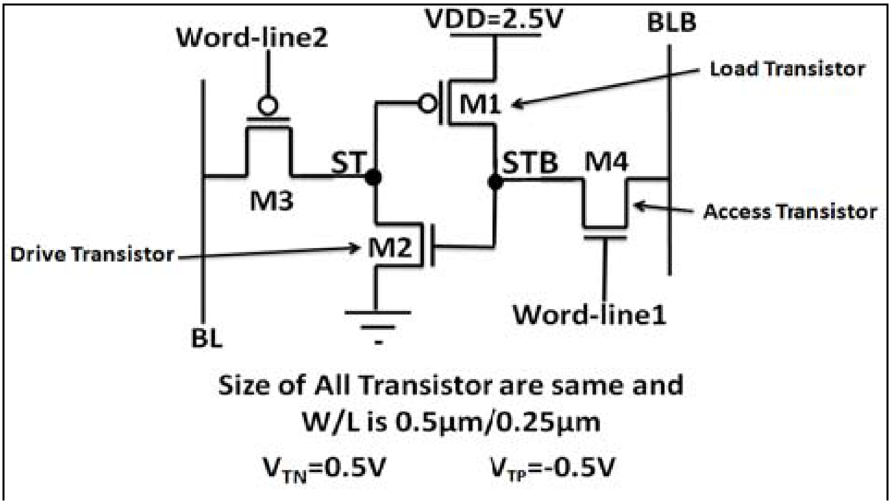


Fig. 3. Modified SRAM cell with 4T proposed by Arash et al. [10]

4 Reliability of Low-Power SRAM

As the transistor feature size scales down, reliability (immunity to soft error), is becoming a critical problem. "Soft errors or transient errors are circuit errors caused due to excess charge carriers induced primarily by external radiations" [5]. Soft errors can change the values of the bits stored leading to functionality failures [5], which are very serious in crypto applications.

As low power SRAM designs are saving energy and reducing the supply voltage and the node capacitance, the transistors are becoming more sensitive to soft errors. The reader is referred to the study in [5] for details on the current low-power design techniques and their effect on reliability. It is noted that, as the reliability and soft errors are becoming to be related and noticeable, low-power designing should put more importance to it as a design dimension of reliability-aware low-power SRAM hardware.

5 Remarks

The demand for security in portable power-constrained situations is increasing. Designing of low power architecture can be achieved through several means, such as pipelining, redundancy, data encoding, and clocking. Pipelining allows voltage scaling which may increases throughput because frequency could be increased resulting lower supply voltage instead [7]. Redundancy minimize shared resources to lower signal activity and buses affecting power consumption to be optimized. Data encoding is helpful in energy efficient state encoding which can reduce the effect on the bits to the minimum, such as using Gray code or One hot encoding. Clocking can be useful when not connected to all, i.e. gated clocks or self-timed circuits [7]; where all low power architecture means suffer new problems and challenges.

A problem, for example, in all hardware architecture power reduction is that it is lacking consistency, i.e. in cryptography and security hardware designing low-power consideration resulted in the need to develop specific energy-efficient algorithm-flexible hardware. Reconfigurable Domain-specific SRAM memory designs are what is needed to provide the required flexibility. However, it may not payback without gaining the high overhead costs related to the generic reprogrammable designs resulting in implementations capable of performing the entire suite of cryptographic primitives over all crypto arithmetic operations.

The technology is moving toward ultra-low-power mode where the hardware processors power consumption should be reduced much. Measured performance and energy efficiency indicate a comparable level of performance to most reported dedicated hardware implementations, while providing all of the flexibility of a software-based implementation [1, 8, 9].

6 Conclusion

This paper is addressing the current need to consider saving energy in the design phase of cryptography computations hardware architectures. Building a specified VLSI design for limited power application is opening the door for low power SRAM memory designs where memory is playing a big role in energy consumption and can be well thought-out as a promising solution. The paper discussed several techniques to save power in SRAM memory designs such as pipelining, redundancy, data encoding, and clocking where all options are having advantages and drawbacks based on the specific cryptography situations and application. In fact, cryptography arithmetic in general is becoming complex and power hungry. It is in real need for efficient power utilization which is achieved in the past through the trade-off between speed, area and power consumption. We focused on the idea of considering SRAM memory modules in subthreshold operation to benefit from ultra-low-power capable systems.

The work presents the idea of modifying available cryptography hardware security architectures to reconfigurable domain-specific SRAM memory designs. We focus on the initiative to design flexible security hardware to gain performance as well as the reduced energy consumption. We propose to consider the reliability issue, which is still a problem, as future research.

Acknowledgments. Special thanks to Professor Bashir M Al-Hashimi, the director of the Pervasive Systems Centre (PSC) at the School of Electronics and Computer Science in University of Southampton, for hosting me during my visit to the UK. Thanks to the collaboration between the British council in Saudi Arabia and KFUPM for sponsoring my travel and living expenses during this research period. Appreciation goes to Dr Biswajit Mishra for introducing me to the subthreshold design area and all fruitful discussions promising for interesting contributions. Thanks to Center of Research Excellence in Hajj and Omrah (HajjCoRE), Umm Al-Qura University (UQU), Makkah, for moral support toward the achievements in this work.

References

1. Goodman, J., Chandrakasan, A.P.: An energy-efficient reconfigurable public-key cryptography processor. IEEE Journal of Solid-State Circuits 36(11), 1808–1820 (2001)
2. Nakagome, Y., Horiguchi, M., Kawahara, T., Itoh, K.: Review and future prospects of low-voltage RAM circuits. IBM J. Res. & Dev. 47(5/6), 6 (2003)
3. Iwama, C.: A Framework for Architecture Level Power Estimation. Thesis, Tanaka & Sakai Lab, University of Tokyo (2002)
4. Sakurai, T.: Perspectives of Low-Power VLSI's. IEICE Trans. on Electronics E87-C(4), 429–436 (2004)
5. Yang, S., Wang, W., Lu, T., Wolf, W., Xie, Y.: Case study of reliability-aware and low-power design. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 16(7), 861–873 (2008)
6. De, V., Borkar, S.: Technology and Design Challenges for Low Power and High Performance. In: Proceedings of the International Symposium on Low Power Electronics and Design, pp. 163–168 (1999)
7. Wolkerstorfer, J.: Low Power Future's hardware challenge. Lecture presentation 09 in the VLSI-Design course, Institute for Applied Information Processing and Communications (IAIK) – VLSI & Security, Graz University of Technology, Austria (2008)
8. Gutub, A., Ibrahim, M.K.: Power-time flexible architecture for GF(2k) elliptic curve cryptosystem computation. In: Proceedings of the 13th ACM Great Lakes Symposium on VLSI, Washington, D.C., USA, April 28-29, pp. 237–240 (2003)
9. Gomathisankaran, M., Keung, K., Tyagi, A.: REBEL - Reconfigurable Block Encryption Logic. In: International Conference on Security and Cryptography (SECRYPT), Porto, Portugal, July 26-29 (2008)
10. Mazreah, A.A., Shalmani, M.T.M., Barati, H., Barati, A.: A Novel Four-Transistor SRAM Cell with Low Dynamic Power Consumption. International Journal of Electronics, Circuits and Systems (IJECS) 2(3), 144–148 (2008)
11. Ho, D., Iniewski, K., Kasnavi, S., Ivanov, A., Natarajan, S.: Ultra-low power 90nm 6T SRAM cell for wireless sensor network applications. In: IEEE International Symposium on Circuits and Systems (ISCAS), May 21-24 (2006)
12. Mohan, N.: Modeling Subthreshold and Gate Leakages in MOS Transistors. Course Project Report, ECE-730, Submitted to Prof. John S. Hamel, Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada (2007)

A Development of Integrated Learning System for Visual Impaired

Wan Fatimah Wan Ahmad, Rustam Asnawi, and Sufia Ruhayani Binti Zulkefli

Department of Computer & Information Sciences
Universiti Teknologi PETRONAS,
31750 Tronoh, Perak
fatimhd@petronas.com.my, rustam@uny.ac.id,
sufia.ruhayani@gmail.com

Abstract. Integrated Learning System (ILS) is a system that integrates several functions of the multimedia elements such as audio, text and slide show presentation to support teaching/learning process. This paper describes a development of ILS model for visually impaired students. In this context, one of the most unique functions of the system is the Text to Voice feature which is able to "read" texts from e-slides written in Bahasa Malaysia for learning a topic in History. Waterfall Model has been chosen as the methodology for developing the system and a prototype of the ILS was introduced. Delphi programming, one of the rapid application programming tools that support object-oriented design is used to develop the system. Microsoft Access is used to manage the Database of audio files containing all the voices in Bahasa Malaysia. The prototype will benefit the visually impaired to enjoy the benefits of computer technology in using ILS.

Keywords: Integrated Learning System, visual impaired, multimedia, learning.

1 Introduction

With the advancement of technology, the visual impaired students are encouraged to use electronic educational material in their learning by the use of multimedia and computer technology as the main tool [1]. Students with visual impairment face difficulties in accessing educational material. Visual impairment (or vision impairment) is vision loss (of a person) to such a degree that additional supports are needed due to a significant limitation of visual capability resulting from either disease, trauma, or congenital or degenerative conditions that cannot be corrected by conventional means, such as refractive correction, medication, or surgery [2].

Microsoft PowerPoint slides are not useful for visually impaired people because the content is visually displayed, and it is not equipped with auditory text content. A system that has Text to Voice feature will come in handy, in which the system can "read" the text to the visual impaired students. Recognizing the significance of such system, this study proposes an ILS that incorporates a system that can "read" text stream. Focusing on people with visual impairment, information such as how they

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 111–121, 2011.

learn and how they interact with technologies were obtained from previous research projects which have conducted observations for 3 to 5 years on the blinds who interact with technologies.

The choice of Microsoft PowerPoint slide is mostly due to its wide utilization as the tool for delivering educational material. In fact, it can be categorized as an interactive learning medium will beneficial for both normal and visually impaired students. Normal students can use this learning system as an alternative way to study since besides reading the slides they are also able to listen to the text audio as well.

Decades ago, the medium of learning between educators and students in universities were blackboard, white board and OHP Hardware. Nowadays, almost all universities in Malaysia are using Microsoft PowerPoint as the medium to deliver lessons. It is possible that in the future, schools will also adopt the same method, whereby teachers will use Microsoft PowerPoint to deliver lessons instead of using writing boards. While such change is possible in schools for normal students, it would not be possible in schools for visually impaired students since the output of the existing Microsoft PowerPoint slides are visual.

Currently, schools that cater for visual impaired students are using JAWS software that is provided by the government. JAWS or Job Access With Speech is produced by the Blind and Low Vision Group at Freedom Scientific of St. Petersburg, Florida, USA. It provides the user with access to the information displayed on the screen via text-to-speech or by means of a Braille display and allows for comprehensive keyboard interaction with the computer. The JAWS software is almost like Microsoft Narrator but it has more advanced functions. The government has considered that JAWS is suitable to support the blinds and visual impaired for navigation and reading using a computer. However, the software is in English language, while in Malaysia, the language used is in Bahasa Malaysia. Therefore, the pronunciations of words are very different which may confuse the visual impaired students because they are different from the teacher's pronounces.

In order to solve this problem, an ILS has been developed. ILS is a system that integrates several functions of the multimedia elements such as audio, text and slide show presentation to support teaching/learning process. In this study, the ILS is specially designed in such a way as to encourage the visually impaired students to use learning materials in e-slide format. Principally, the ILS is developed to overcome several problems:

1. The visual impaired are not able to use the technology or computer that is available for other students.
2. The visual impaired will be left behind in using any technology such as Microsoft PowerPoint. In other words, they will not be using any technology to assist them in learning.

Therefore, the objective of this study is to report on the development of an ILS for visually impaired students. This study focuses on secondary history subject which is part of the Malaysian syllabus. ILS integrates the written materials developed in Microsoft PowerPoint with Text to Voice processing.

2 Literature Review

The number of people with visual impairment has reached up to 135 million compared to the world's population which is approximately 6 billion [3].Technology-Related Assistance for Individuals with Disabilities Act of 1988 (also call as Tech Act) has been defined as the first Assistive Technology Device for people with visual impairments [4].

Computers and technologies are usually developed for normal people; however they should also include Human Computer Interaction (HCI) systems for the visually impaired or blinds. For example, W3C's Web Accessibility Guideline (WCAG) has provided a general guideline for providing a universal way in to computing technology that consisted of the specification on shapes and colors. According to WCAG [5], a system that was specially developed for visual impaired people will be totally different from normal people.

While [6] mentioned that in order to learn about visual impaired people, it is important to include some critical extreme values of the relevant characteristics. This shows that in order to gather the requirements of such ILS, the visually impaired or the blinds should be involved. There is a couple of ways to determine the interaction for visually impaired people such as tactile or audio [1]. This research has shown that principally tactile is more effective than audio.

Meanwhile, according to [4] there are seven categories of Assistive Technology (AT) devices; positioning, mobility, augmentative and alternative communication, computer access, adaptive toys and games, adaptive environments, and instructional aides. Related to visually impaired people, the accessibility of computer devices is the most important category to be considered in developing the ILS. In this context, the accessibility to computer devices can be interpreted as the visually impaired can "read" the text written in an e-slide file such as the Microsoft PowerPoint.

Researchers have studied different matters on the use of technologies and they have identified several points such as blind acceptance of technologies, learning method of the blinds, ICT and its effect, development of Text to Voice, and relationship between the blinds and the visually impaired. Each of these is addressed in the subsequent sections.

2.1 Blind Acceptance towards Technologies

[7] has revealed that although it is not easy for the visual impaired to interact with technologies, but the effort and passion make them capable to learn it. This paper has shown that the visual impaired people can learn about technologies and computers. Another researcher [8] has developed a system called AudioStoryTeller. Basically, the system will help young blind students to read story book with smaller device than Laptop or Personal Computer (PC). AudioStoryTeller also considers how the blinds interact or response to any provided technology that is specially designed for them. A complete system was tested on the blinds to assess the usefulness of that particular system to them. The product is quite established in the market, which proves that it is successful in assisting the blind to "read" a story. Therefore, the results of [7] was considered in developing the ILS for the blinds.

2.2 Learning Method

The learning process for the visual impaired should include instructional design, communication bridges, skill development simulations, distance learning practices and discovery learning [8]. Department of Allied Health and Science [9], UNC School of Medicine conducted a research called The Deaf-Blind Model Classroom Project. In that research, two persons volunteered to test the system, and observations were made on both of them. Both of them, from the age group 10 to 15 years old, who never knew the alphabets before, yet succeeded to make sentences within 1 to 3 years of using the system.

The research done by UNC School of Medicine was for the beginning of the learning process. It may take years to observe the whole learning process step by step. In this paper, a summary of the observations made by other researchers is provided, such as from UNC School of Medicine.

The next section describes the learning processes of the visual impaired.

2.2.1 Instructional Design

Generally, instructional design (ID) is a set of actions that treat knowledge familiarity, value, and request directions at the maximum potential. In this context, ID is purposely focused on the set of activities that takes places on how the visual impaired students learn. In non-technological way, the blinds read using Braille. The Braille is a symbol of dots that represents the "ABC" alphabet. The visual impaired/blinds read by feeling the dots with their fingers. By using Braille they are able to use the keyboard, remember the alphabets and type as normal people do. They can use the keyboard by remembering each position of the alphabets. On the computer's keyboard, there are two dots on "F" and "J" alphabets, which will provide the clue to the positions of other alphabets.

2.2.2 Communication Bridges

The visual impaired has less problem to communicate compared to other physical disabilities. They can easily communicate and express what they want or dislike with others. However, the main communicate issue with the visual impaired is describing pictures or something that needs visual assessment to interpret them. For example, it is very challenging to talk about colors to visual impaired as they have never seen them before. It is almost pointless to make them understand about colors as they cannot visualize anything. Thus, when designing an ILS system it is important to consider those areas or subjects that are beneficial to them, such as the alphabet, weather, knowledge and etc.

2.2.3 Skills Development Simulations

Skills Development Simulations [8] conducted a study on job interviews. The interviews were conducted in the same manner as oral examinations, where student were tested on a case study and is requested to propose solutions to the case study. The purpose was to determine whether the students had understood the information that had been given. Since it is not possible to carry out test for the visual impaired using paper (perhaps Braille paper will be provided), interviews is most appropriate to test these students' knowledge. This method is similar to human computer interaction between visual impaired and the ILS, whereby the blinds can use computer directly (by clicking the mouse and typing words).

2.2.4 Distance Learning

Distance Learning can be applied when a teacher is not in the same room with the students. Video conferencing can also be another method of teaching. Although video conferencing provides both voice and video streaming, unfortunately for the blinds, only voice is beneficial for them. Through video conferencing, the teacher (who may not blind) can observe the students' learning process. Then again, to reduce the cost of development and enhancing efficiency of the ILS, the Distance Learning can be applied only through audio call between visual impaired students and the teacher.

2.2.5 Discovery Learning

In discovery learning, [8] introduces interactive learning experiences such as games, non-fiction stories, and video segments. To make the learning more interesting and fun, games can also be incorporated in the teaching and learning processes for the visual impaired. For example, to make them remember, it is very interesting to make the learning experiences through non-fiction story of "snow white", which can be played in the video. As ILS is developed with several features of audio and text, so it is possible to implement discovery learning for the visual impaired. This may make the learning experience interesting and fun.

Besides, the system can also be made available to normal students whereby instead of reading, they can listen to the audio while doing something else. Nowadays, it has become a trend to study while listening to music. Using the ILS, normal students can read and learn while listening to the voice reading the written text.. This may be a very effective way to study and at the same time it may also increase students' performances.

2.3 Student, ICT and Its Impact

According to [10], many researches on ICT and its impacts have been conducted, but the real effects have not yet been exposed. These may be caused by the inconsistency of methods used leading to uncertainty of results in some research studies. Hence, [10] have suggested using more consistent and normal standards in conducting research on the effect of students being exposed to ICT and technologies. Through all that, the most important thing while doing this research is to understand the impact on students in terms of their thinking, knowledge, understanding and acting processes. Their attitude towards ICT and technologies also has some impacts on the students' learning process [11, 12].

3 Methodology

Waterfall Development Model has been adopted into the development of ILS. Figure 1 shows the system architecture of ILS. The user will interact with the system. Inside the machine (laptop), there will be a database of words in Bahasa Malaysia. Text to

Voice routine will be resided in the machine. The functionality testing of the system has been performed at Sekolah Kebangsaan Sultan Yussof and Sekolah Sri Mutiara. 5 participants are involved in this study.

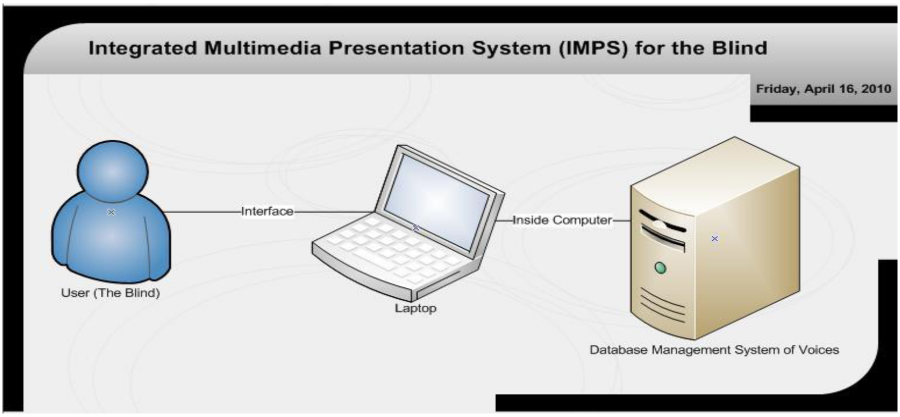

Integrated Multimedia Presentation System (IMPS) for the Blind
Friday, April 16, 2010
User (The Blind)
Interface
Laptop
Inside Computer
Database Management System of Voices

Fig. 1. System architecture

4 Integrated Learning System

In this paper, ILS is a system that integrates two components, audio and text. The learning system enable the visual impaired students to "read" the text on a Microsoft PowerPoint slide by incorporating Text to Voice feature that reads the text to them. Waterfall Model has been chosen as the methodology for developing the system and a prototype of the ILS. Delphi programming, one of the rapid application programming tools that support object-oriented design is used to develop the system. Microsoft Access is used to manage the Database of audio files containing all the voices in Bahasa Malaysia. The difference between JAWS and ILS is that JAWS is using the JAWS Scripting Language.

The interaction between the ILS and visually impaired student has applied the combination of both audio and tactile as recommended by [1]. In the concept of ILS, audio is exploited in order to benefit the visually impairment or the blinds to enable them to "read" the text. Some short or hot keys are provided to the blinds to support accessibility to the system. The short key will be pressed by the blind students through the keyboard device as an alternative way to execute their choices. After conducting a few feasibility analyses such as economic, technical and time perspectives, it is more feasible and less risk to produce ILS with combined modes of interaction.

Four modules have been developed namely: Slide to Text Module, Parsing module, Match Module and Database Module. Details for each module are given in the following sections.

4.1 Slide to Text Module

In this module, the Microsoft PowerPoint slides have to extract the texts from sentences. In normal Power Point slide, few text boxes will be placed on a slide. This module will gather all sentences in the text boxes on all slides and convert into string of sentences.

4.2 Parsing Module

In this module, it will accept any string of words and chunk them into one single word in an array. The purpose of the system is to store all words or sentences into an array to make them easier to be used in Match Module later. For example a slide of Microsoft PowerPoint has 6 sentences and each sentence has 10 words, so the total words in that particular slide in an array would be 60 words.

4.3 Match Module

In this module, it will take the array stored by Parsing Module and matches with the database. Some researchers used normal looping to make the system, which always produces errors. Therefore Match Module has adopted timer technique. Timer actually acts as looping, it produces start and end time of the voicing slide. The advantage is that when using timer, the system will be free from the previous errors.

4.4 Database Module

Database module provides all the words in Bahasa Malaysia for the specific learning material of this project. The learning material is Form one History covering chapter

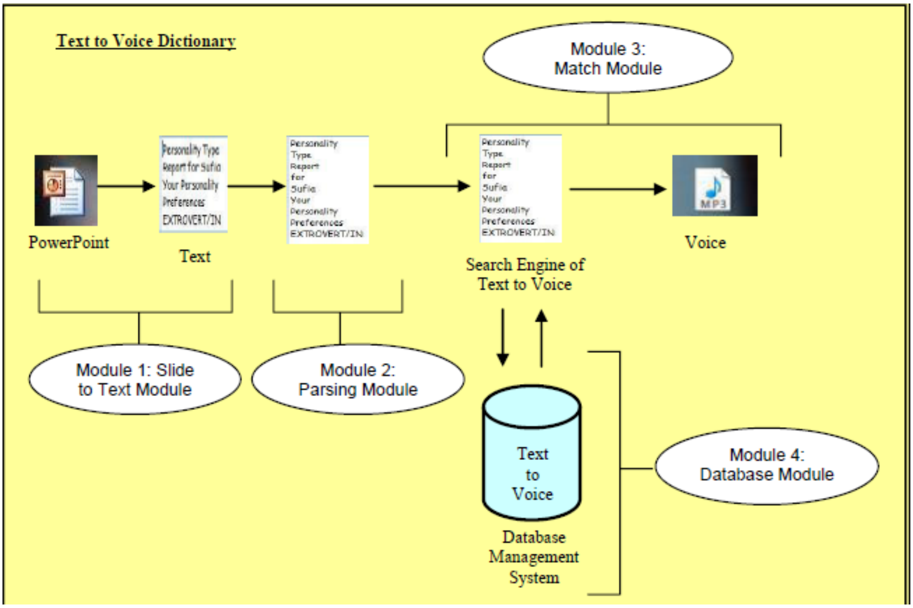

Text to Voice Dictionary
Module 3: Match Module
Personality Type Report for Sufia Your Personality Preferences EXTROVERT/IN
Personality Type Report for Sufia Your Personality Preferences EXTROVERT/IN
Personality Type Report for Sufia Your Personality Preferences EXTROVERT/IN
MP3
PowerPoint
Text
Search Engine of Text to Voice
Voice
Module 1: Slide to Text Module
Module 2: Parsing Module
Text to Voice
Database Management System
Module 4: Database Module

Fig. 2. Text to Speech of ILS for the Blinds

one to chapter four. Several audio files are saved in a specific file. The files can be retrieved using Microsoft Access. So basically in Microsoft Access, there are maps of words and related specific audio MP3 file.

Figure 2 shows the relation between modules for developing Text to Voice in Bahasa Malaysia.

5 Results and Discussions

5.1 The Interfaces of the ILS

Figure 3 shows the main page of ILS. Basically it can execute Text to Voice through Microsoft PowerPoint instead of normal text. Regardless of the shape and number of text box on the Microsoft PowerPoint slide, the ILS is able to "read" all text written on the slide. Figure 4(a) and 4(b) show the interfaces of the snapshots in running ILS.

5.2 Feedback on the Functional Testing

The testing has been conducted at Sekolah Kebangsaan Sultan Yussof and Sekolah Sri Mutiara. 5 teachers were asked to go through the ILS and feedbacks and interviews on the ILS were taken after they have finished with the system.

The feedbacks were: 1) the system should include a voicing notification of slide transition in order to inform the students that a slide transition is taking place. This way, the system will also be able to control the voice speed; 2) a proper introduction can be provided before starting the slide; 3) to provide background music rather than merely "reading" a text only; 4) ILS does have an impact on the visual impaired students and it is capable to boost the students' self esteem.

A positive comment from the teachers include the students would be able to study independently using this system. This presents a new method to the visual impaired students, in particular, whereby the computer is no longer only limited to normal students. Visual impaired students are also able to benefit from the computer to enhance their learning process.

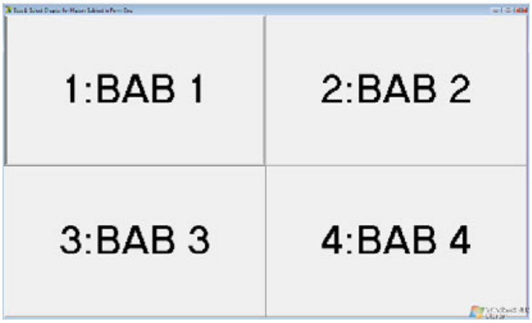
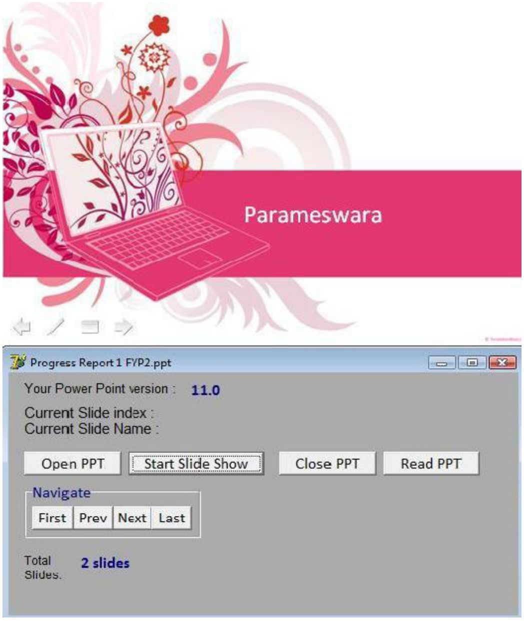


Fig. 3. Main page of ILS



Parameswara

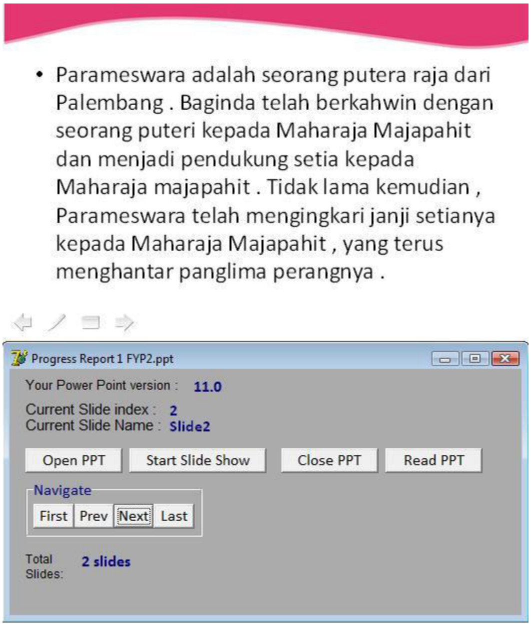
Progress Report 1 FYP2.ppt

Your Power Point version : 11.0

Current Slide index :
Current Slide Name :

Total Slides: 2 slides

Fig. 4(a). Snapshot of executed ILS



- Parameswara adalah seorang putera raja dari Palembang . Baginda telah berkahwin dengan seorang puteri kepada Maharaja Majapahit dan menjadi pendukung setia kepada Maharaja majapahit . Tidak lama kemudian , Parameswara telah mengingkari janji setianya kepada Maharaja Majapahit , yang terus menghantar panglima perangnya .

Progress Report 1 FYP2.ppt

Your Power Point version : 11.0

Current Slide index : 2
Current Slide Name : Slide2

Total Slides: 2 slides

Fig. 4(b). Snapshot of executed ILS

6 Conclusion

The paper has discussed the development of ILS for visual impaired students. The functional testing indicated the potential of the system, as well as the positive feedbacks from the teachers. The ILS is not only limited to visual impaired students but it can also be used by non-blind students as an alternative interactive learning system. The system does not only provide an easy way for visual impaired students to study e-slide materials, but it would also help to boost their self esteems by being able to study independently using the system. The proposed ILS system complements the current JAWS software, because JAWS read Malay text using English language. Since the ILS provides Text to Voice feature in Bahasa Malaysia, this software is more appropriate for learning purposes in visual impairment environment in Malaysia. Future work includes testing with the real users and enhancement will be made according to the feedback. ILS will benefit the visual impaired to enjoy the benefits of computer and multimedia technology.

Acknowledgment

The authors would like to thank Universiti Teknologi PETRONAS and all those involved in this study either directly or indirectly.

References

1. Baldonado, M., Chang, C., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. International Journal of Digital Libraries 1(2) (1997)
2. Arditi, A., Rosenthal, B.: Developing an objective definition of visual impairment. In: Vision 1996: Proceedings of the International Low Vision Conference, pp. 331–334. ONCE, Madrid (1998)
3. van Leeuwen, J. (ed.): Computer Science Today: Recent Trends and Developments. LNCS, vol. 1000. Springer, Heidelberg (1995)
4. Bruce, K.B., Cardelli, L., Pierce, B.C.: Comparing Object Encodings. In: Ito, T., Abadi, M. (eds.) TACS 1997. LNCS, vol. 1281, pp. 415–438. Springer, Heidelberg (1997)
5. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs, 3rd edn. Springer, Heidelberg (1996)
6. Shinohara, K., Tenenberg, J.: Observing Sara: A Case Study of a Blind Person's Interactions with Technology. In: Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility, Tempe, Arizona, USA, pp. 171–178 (2007)
7. Sánchez, J., Galáz, I.: AudioStoryTeller: Enforcing Blind Children Reading Skills. In: Stephanidis, C. (ed.) HCI 2007. LNCS, vol. 4556, pp. 786–795. Springer, Heidelberg (2007)
8. Patron, B.S.: Snapshot of Interactive Multimedia at Work Across the Curriculum in Deaf Education: Implications for Public Address Training. Journal of Educational Multimedia and Hypermedia 15(2), 159–173 (2006)
9. School of Medic UNC Winter, vol. 13(2) (2006), http://www.med.unc.edu/ahs/clds/projects/north-carolina-deaf-blind-project/db-case-studies

10. Cox, M.J., Marshall, G.: Effects of ICT: Do we know what we should know? Journal Education and Information Technologies 12(12), 59–70 (2007)
11. Pearson, M., Naylor, S.: Changing contexts: Teacher professional development and ICT pedagogy. Journal Education and Information Technologies 11(3-4), 283–291 (2006)
12. Watson, D.M.: Pedagogy before Technology: Re-thinking the Relationship between ICT and Teaching. Journal Education and Information Technologies 6(4), 251–266 (2001)

Fingerprint Singularity Detection: A Comparative Study

Ali Ismail Awad[1] and Kensuke Baba[2]

[1] Graduate School of Information Science and Electrical Engineering,
Kyushu University, Japan
[2] Research and Development Division, Kyushu University Library, Japan
{awad,baba}@soc.ait.kyushu-u.ac.jp

Abstract. A singular point or singularity on fingerprint is considered as a fingerprint landmark due its scale, shift, and rotation immutability. It is used for both fingerprint classification and alignment in automatic fingerprint identification systems. This paper presents a comparative study between two singular point detection methods available in the literature. The Poincaré index method is the most popular approach, and the complex filter is another proposed method applied on the complex directional images. The maximum complex filter response is highly related to the regions with abrupt changes in the ridge orientations. These regions have a high probability to contain a singular point. The optimum detection method in both processing time and detection accuracy will be updated to suite our efficient classification method. The experimental evaluation for both methods proves that the accuracy achieved by complex filter is up to 95% with considerable processing time compared to 90% with Poincaré index method.

Keywords: Fingerprints, Singular Point, Poincaré Index, Complex Filters.

1 Introduction

Fingerprint is the dominant trait between different biometrics like iris, retina, and face. It has been widely used for personal recognition in forensic and civilian applications because of its uniqueness, immutability, and low cost. Moreover, it also has a valuable source of information that easy to be extracted without violating user privacy. Embedded fingerprint systems that supporting instant identification in large databases are increasingly used, and reducing the matching time is a key issue of any identification system. Fingerprint classification becomes indispensable choice for reducing the matching time between the input image and the large database.

Fingerprint structure is defined by the interleaved ridges and furrows constructed on the finger tip. It falls under two categories, local and global structure. Fingerprint global features are considered as a coarse level structure, and it highly depends on the ridge flow orientation inside overall fingerprint image. Singular Points (SP) or singularities are the most important global characteristics of a fingerprint. A core point is defined as the topmost point of the innermost curving ridge, and a delta point is the center of triangular regions where three different direction flows meet [1]. A global

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 122–132, 2011.

structure of different fingerprint images with core and delta points is shown in Fig. 1. A singular point area is generally defined as a region where the ridge curvature is higher than normal and where the direction of the ridge changes rapidly.

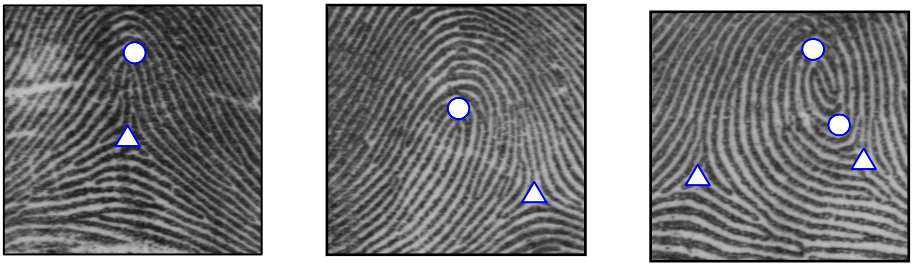


Fig. 1. Global structure of fingerprint images with different singular points locations

Singular points detection is an essential concept of fingerprint recognition and classification, however, the singular detection is sensitive to different conditions such as noise, sensor type, and fingerprint status. In this paper, we are interested only in fingerprint classification applications. Principal axes technique [2] has been used as a secondary stage of singularity-based classification algorithm. The algorithm takes into count not only the number of SP, but also the relation between them and their orientations on (x, y) coordinates. Pseudo ridge tracing [3] has proposed the usage of ridge tracing beside the orientation field analysis as a supplement to compensate the missing singularities and improve the classification process. Singular point characteristics can be used to construct a feature vector for training different learning based classification approaches such as Multilayer Preceptron (MPL) Neural Networks [4], and Bidirectional Associative Memory (BAM) Neural Networks [5].

Many proposed approaches for detecting SP are available in the literature, and most of them are working on the orientation field images. From the SP definition, the interested regions with high probability to include SP must have abrupt changes in the directions orientation inside. Poincaré Index (PI) [6] is the most famous approach of SP detection. The Poincaré index is calculated for each block by summing the directions changes around the selected block. Driven from the former definition, Complex Filters (CF) [7] is another proposed method to extract regions with high orientation changes using a first order complex filter. These two methods have been especially selected due to the lack in processing time evaluation in both. The contributions of this paper fall under catching the optimum singularity detection method in terms of the detection and localization accuracies, and the computational complexity. Moreover, detecting the processing time bottlenecks of both algorithms will open the doors to improve their performance dramatically. In addition to its applicability by many researchers in the area of fingerprints, the optimum algorithm will be updated and applied to our proposed classification method [8], [9] that uses the SP location as a base for fingerprint image partitioning to achieve higher robustness for fingerprint image shift and rotation.

The reminder part of this paper is organized as follows: Section 2 explains the Poincaré index method as a most dominant one in the literature. Section 3 presents the background of singular point localization using a complex directional image and complex filtering technique. The experimental results of the two methods, overall performance evaluation, and the comparative study are reported in Section 4. Finally, conclusions and future work are reported in Section 5.

2 Poincaré Index Method

Poincaré index-based method is the most popular approach for detecting fingerprint singularities (both core and delta points). Many researchers such as [10-12] have applied and also modified this method from different prospective. The Poincaré index has the ability to detect both types of singularity, however, it is very sensitive to noise and the variations of the grayscale level inside the fingerprint image.

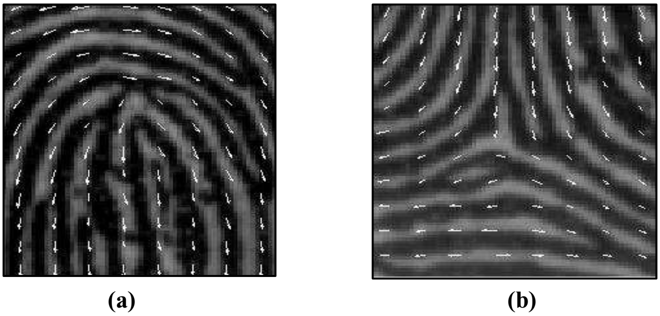

(a) **(b)**

Fig. 2. Orientation field around different fingerprint regions: (a) ridge orientations around core point, (b) ridge orientations around delta point

Since Poincaré index is working on the direction changes, the first step prior to implementing PI calculation is to extract the directional image (orientation filed) corresponding to the input fingerprint. Pixel gradient method [10] is the common way to extract pixel orientation in fingerprint images. Gradient calculation is varying from using simple method like "Sobel filter" into more complex filters such as "Priwett" [13]. In our implementations, and for simplicity, directions filter mask has been used to estimate the orientation fingerprint image. Figs. 2(a) and 2(b) show an example of ridges orientations around core and delta points, respectively. We assume that $\theta(i,j)$ is pixel orientation of any element in the directional image at pixel (i,j), where $0 \le \theta(i,j) < \pi$. Let (i_k, j_k) be the elements selected for calculating the Poincaré index of a point (i,j) for $0 \le k \le N-1$. Then, the Poincaré index can be calculated as follows:

$$Poincaré(i,j) = \frac{1}{2\pi}\sum_{k=0}^{N-1}\Delta(k), \tag{1}$$

where Δ (k) is the accumulative changes in the pixel orientations. It is mathematically represented as:

$$\Delta(k) = \begin{cases} \delta(k) & \mathit{if}\ |\delta(k)| < \pi / 2 \\ \pi + \delta(k) & \mathit{if}\ \delta(k) \leq -\pi / 2 \\ \pi - \delta(k) & \mathit{otherwise} \end{cases} \tag{2}$$

and

$$\delta(k) = \theta(i_{(k+1)\bmod(N)}, j_{(k+1)\bmod(N)}) - \theta(i_k, j_k) \ . \tag{3}$$

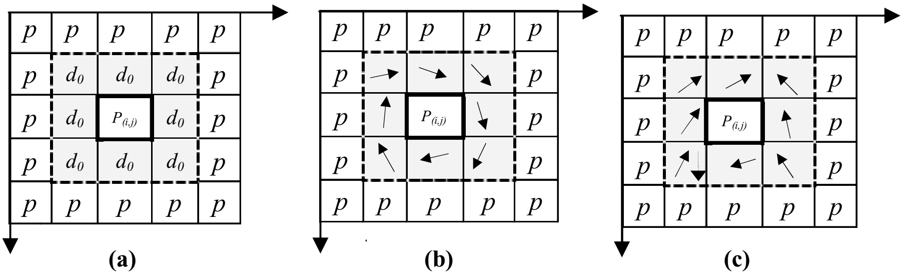


Fig. 3. Representations of different Poincaré indices: (a) General calculation scheme, (b) Poincaré index representation of (+1), (c) Poincaré index representation of (+1/2)

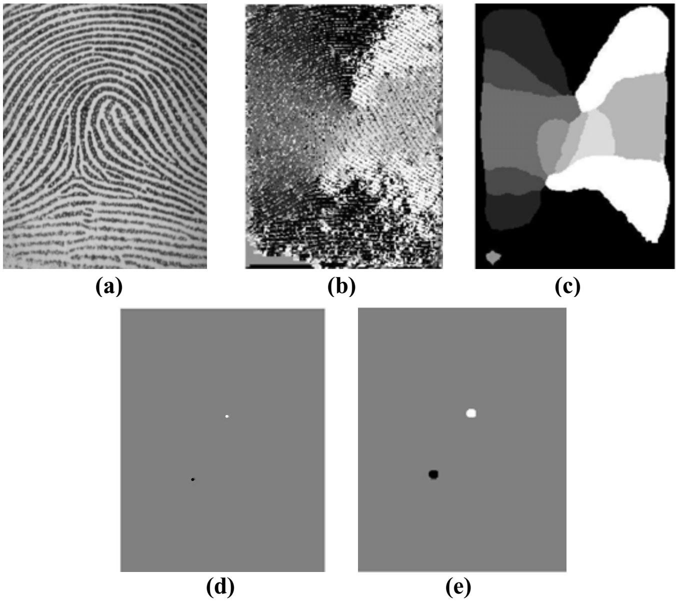


Fig. 4. Sample output of Poincaré index method: (a) input fingerprint image, (b) estimated orientation image with (8 directions), (c) normalized orientation image (θ = 0 black, θ = 7×π/8 white), (d) Detected SPs using PI calculated on small scale (block size = 8×8 pixels), (e) Detected SPs using PI calculated on large scale (block size = 32×32 pixels)

Poincaré index may have four different values: 0, -1/2, +1/2, and +1 corresponding to no singular point, a core point, a delta point, and probability of two singular points respectively. Fig. 3(a) shows the rule of direction estimation for Poincaré index, where Fig. 3(b) represents the PI equivalent to +1 (two singularities may become available), and Fig. 3(c) represents the pixels orientations of PI equivalent to +1/2 (delta point). Fig. 4 demonstrates the output images for different PI execution steps: Figs. 4(a) and 4(b) show the input image and its corresponding orientation field with orientation rage ($\theta = 0, \pi/8, \ldots, 7\times\pi/8$), respectively. Fig. 4(c) shows the normalized directional image, as all closed directions are grouped into the closer one. Although this process is time consumer, it is needed to detect the abrupt changes between the different areas. Fig. 4(d) and 4(e) show the PI calculated on two different scales.

3 Complex Filter Method

Complex Filter (CF) with Gaussian window [7] has been implemented on the complex directional field to extract the symmetry singular points in fingerprint image. The main steps of the complex filter implementations are: orientation field estimation, complex filter construction, and the convolution between both complex filter and directional image. To speed up the total consumed time, convolution process has been performed in frequency domain.

3.1 Complex Orientation Field Estimation

Many proposed approach for complex orientation field estimation can be found in the literature. To construct the complex directional image, assume that $G_x(i, j)$ and $G_y(i, j)$ are the pixel derivations at pint (i, j) in both *x* and *y* directions, respectively. Then, the complex directional image can be calculated as:

$$z(x, y) = (G_x + iG_y)^2, \tag{4}$$

where *i* is the imaginary unit. Figs. 5(a) and 5(b) show the original input image and the final extracted directional image corresponding to the input image in Fig. 4(a).

3.2 Complex Filter Construction

One advantage of using complex filter instead of the traditional Poincaré index is the possibility to detect both location and direction of the singular point [7]. The other advantage is the complex filter can be designed to extract core or delta point individually, while Poincaré index should perform all calculations first, and then decide the singular type. In this research we are interested only in core point detection, since most of fingerprints have this type of singularity. Therefore, we put focus on implementing the complex filter for core point. The polynomial explanation of the complex filter with order *m* in (*x*, *y*) yield as:

$$(x + iy)^m g(x, y), \tag{5}$$

where $g(x, y)$ is a Gaussian window that can be expressed as:

$$g(x, y) = \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \tag{6}$$

where σ is the variance of Gaussian window. Since Gaussian function is the only function which is orientation isotropic and separable, it does not introduce any orientation dependency. Fig 6(a) shows the 3D plot of absolute complex filter response in spatial domain with selected filter size as (32×32) pixels. In contrary, Fig. 6(b) shows the output image after convoluting the complex filter with the directional image that shown in Fig. 5(b).

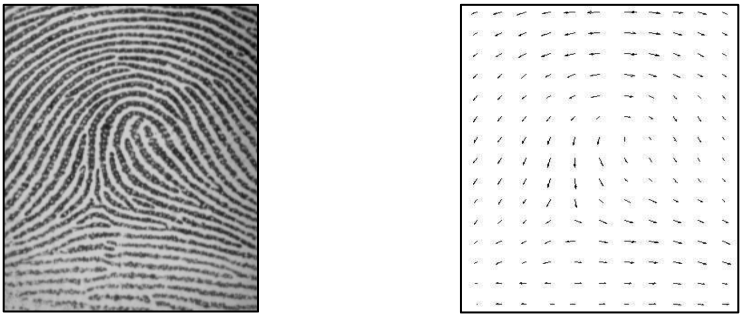
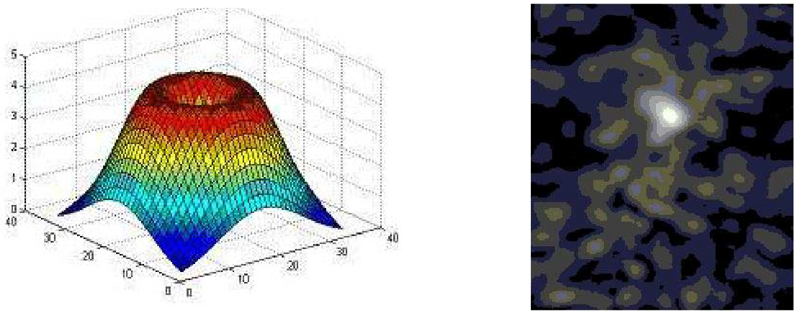

(a)
(b)

Fig. 5. Orientation field estimation using gradient method: (a) original input image, (b) corresponding directional image of (a)



(a)
(b)

Fig. 6. Complex filter and its output response: (a) spatial domain representation of first order complex filter, (b) magnitude of filter response after its convolution with the orientation image in Fig. 5(b)

In Fig. 6(b), the maximum filter response at the real core point is represented as a brighter area, and it can be consistently localized by extracting (x, y) coordinates of the maximum points inside the convolution output array.

4 Experimental Results and Comparisons

A good reference point localization approach should be robust for noise, shift, and rotation. In addition, it should consistently and accurately detect a unique reference point for all types of fingerprints including plain arch fingerprint in which no common singular points exist. The key issue of any fingerprint identification system is the response time. Singular point can be used in both classification and matching stages to reduce the response time, however, it may become a worst option if the processing time of extracting these singularities is not taken into account. Through our experimental work, we will shade lights on the processing time as an important evaluation factor of singularity detection algorithms beside the total detection accuracy.

In the experimental work, both algorithms have been evaluated using two Fingerprint Verification Competition 2002 (FVC2002) [14] subsets. DB1_B and DB2_B are two FVC2002 subsets captured by optical sensors "TouchView II" by Identix and "FX2000" by Biometrika respectively. Each data set contains 80 images, and all images have been used to evaluate our both algorithms. Poincaré index has been evaluated using the optimal block size N=8. A first order m=1 complex filter has been constructed with an optimal size as (32x32) pixels, and the Gaussian variance is empirically sets to *7* to achieve the maximum performance.

4.1 Processing Time

Due to the separable property of Gaussian window, a 1D complex convolution has been used instead of 2D one to speed up the calculation process. All output results in the following tables have been generated using Intel® Core™ 2 Due Processor (T9300, 2.5 GHz, 6 MB L2 cash), 3 GB of RAM, Windows XP® Pro 32 bit, and Matlab® R2009b version.

Table 1. Processing time measurements of singular point detection algorithms in DB1_B

Method	**80 Fingerprint images (DB1_B)**			
	Min time / Fingerprint	Max time / Fingerprint	Mean time / Fingerprint	Total time / Database
Poincaré Index	2.21 sec	4.52 sec	3.14 sec	**270** sec
Complex Filter	0.11 sec	1.20 sec	0.13 sec	**10.92** sec

Table 2. Processing time measurements of singular point detection algorithms in DB2_B

Method	**80 Fingerprint images (DB2_B)**			
	Min time / Fingerprint	Max time / Fingerprint	Mean time / Fingerprint	Total time / Database
Poincaré Index	2.40 sec	07.1 sec	4.52 sec	**320** sec
Complex Filter	0.12 sec	1.35 sec	0.12 sec	**9.96** sec

Tables 1 and 2 point out the huge difference in processing time between Poincaré index method and the complex filter one for two different databases in terms of minimum, maximum and mean processing times. From both tables, we have believed that Poincaré index method imposes some processing time impact on the total response time of the system that uses singular point in classification or alignment. The over all performance of the automatic fingerprint identification system will get degraded in terms of the processing time after embedding Poincaré index method for detecting fingerprint singularity.

4.2 Bottlenecks Detection

Detecting processing time bottleneck is an important contribution of this work. To achieve that, we have measured the time consumed by each processing step in both algorithms. Fig. 7 shows the output of the measurement process. By Fig. 7(a), we realize that the orientation field normalization is the core point of Poincaré index. By Fig. 7(b), we digest that image segmentation considered as the bottleneck of complex filter implementations. In order to improve the computational time for both algorithms, we should put some efforts to enhance these two bottlenecks before going further to use any of them in fingerprint classification or alignment.

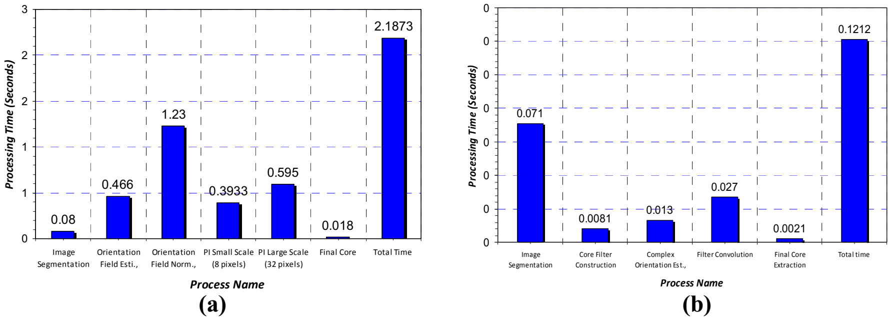


Fig. 7. Processing time measurements for each processing step in both methods: (a) Poincaré index method, (b) Complex filter method

4.3 Singularity Detection Accuracy

As for accuracy, Tables 3 and 4 show the accuracy measurements of both algorithms in the two selected databases. The algorithm accuracy is measured as the number of images in which the singularities have been correctly detected to the total images in the database. By both Tables, the complex filter based algorithm has an advanced step over the Poincaré index method. There is another term of accuracy that can be considered which is the accurate localization of the singularity itself. All fingerprints in the two selected databases have been visually inspected to judge this type of accuracy. By evaluating both algorithms in this context, we found that both of the evaluated algorithms produce closed locations of the detected singularities. The tolerance between the two methods is up to ±16 pixels.

Table 3. Accuracy measurement of singular point detection algorithms in DB1_B

Database	80 Fingerprint images (DB1_B)			
	Correctly detected	Incorrectly detected	Ambiguous	Total Accuracy %
Poincaré Index	68	12	0	**85**
Complex Filter	73	7	0	**91.25**

Table 4. Accuracy measurement of singular point detection algorithms in DB2_B

Database	80 Fingerprint images (DB2_B)			
	Correctly detected	Incorrectly detected	Ambiguous	Total Accuracy %
Poincaré Index	72	8	0	**90**
Complex Filter	76	4	0	**95**

4.4 Other Considerations

In this subsection, we report the robustness of both algorithms in terms of noise, shift, and rotation. Fig. 8 shows singular point detection by both explained methods under the above conditions. By Figs. 8(a), and 8(b), we realize that CF performs well under little noise, but both methods fail under the heavy noise. By Fig. 8(c), it is clear that PI fails under shifting condition even for a good image. Finally, from Figs. 8(d) and 8(e) we observe that image rotation imposes a little effect on PI accuracy. In contrary, the accuracy of CF is immutable under rotation condition.

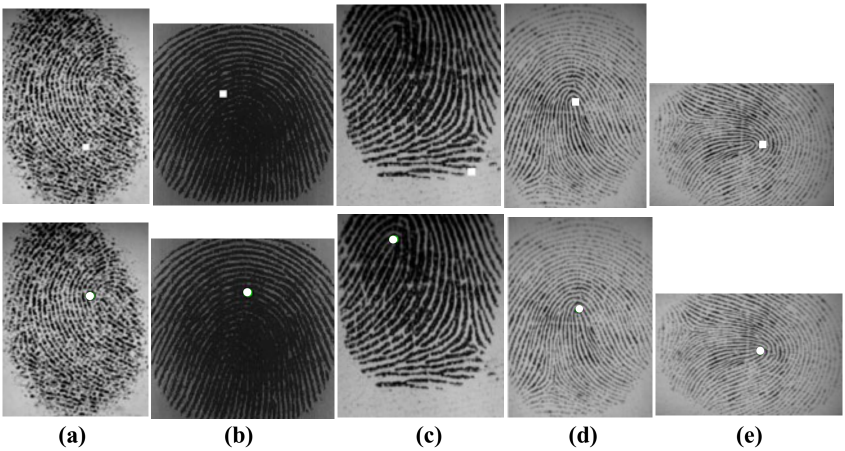


Fig. 8. Performance evaluation of PI (Top) and CF (Down) under different conditions: (a) little noise, (b) heavy noise, (c) shifted up image, (d) normal image, (e) rotated image by 90º

5 Conclusions and Future Work

Fingerprint singular point is considered as a landmark of fingerprint topology, it is scale, shift, and rotation invariant. These singularities can be used for both classification and alignment processes in the automatic fingerprint identification systems. This paper was seeking for catching the best singular point detection algorithm in terms of processing time and detection accuracy. The selected algorithm will be developed and extended for a novel fingerprint classification technique. The comparison has been carried out on two singularity detection algorithms by Poincaré index and complex filters. Our experimental wok concludes that the complex filter method is working very well and it achieves high detection accuracy in a very considerable processing time. Driven from this conclusion, the complex filter method will be orientated to suite our classification technique, and moreover, parallel programming implementation will become an available idea for improving its efficiency, and its accuracies.

Acknowledgement

This work was partially supported by CREST program of Japan Science and Technology Agency (JST) from 2008 to 2012 and the Grant-in-Aid for Young Scientists (B) No. 22700149 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2010 to 2012.

References

1. Wang, L., Dai, M.: Application of a new type of singular points in fingerprint classification. Pattern Recognition Letters 28, 1640–1650 (2007)
2. Klimanee, C., Nguyen, D.T.: Classification of Fingerprints Using Singular Points and Their Principal Axes. In: Proceedings of 2004 IEEE Consumer Communications and Networking Conference (CCNC 2004): Consumer Networking, IEEE, Las Vegas (2004)
3. Zhanga, Q., Yan, H.: Fingerprint Classification based on Extraction and Analysis of Singularities and Pseudo Ridges. Pattern Recognition 37, 2233–2243 (2004)
4. Sarbadhikari, S.N., Basak, J., Pal, S.K., Kundu, M.K.: Noisy Fingerprints Classification with Directional Based Features Using MLP. Neural Computing & Applications 7, 180–191 (1998)
5. Kristensen, T., Borthen, J., Fyllingsnes, K.: Comparison of neural network based fingerprint classification techniques. In: International Joint Conference on Neural Networks (IJCNN), pp. 1043–1048. IEEE, Orlando (2007)
6. Kawagoe, M., Tojo, A.: Fingerprint pattern classification. Pattern Recognition 17, 295–303 (1984)
7. Nilsson, K., Josef, B.: Localization of corresponding points in fingerprints by complex filtering. Pattern Recognition Letters 24, 2135–2144 (2003)
8. Awad, A.I., Baba, K.: Toward An Efficient Fingerprint Classification. In: Albert, M. (ed.) Biometrics - Unique and Diverse Applications in Nature, Science, and Technology. InTech (2011)

9. Awad, A.I., Mustafaa, M., Moness, M.: A New Fingerprint Classification Approach Based on Fast Fourier Transformer. In: Proceedings of the 5th International Conference on Informatics and Systems. Faculty of Computers & Information, pp. 78–83. Cairo University, Cairo (2008)
10. Liu, M., Jiang, X., Kot, A.C.: Fingerprint Reference-Point Detection. EURASIP Journal on Applied Signal Processing, 498–509 (2005)
11. Parka, C.-H., Leeb, J.-J., Smitha, M.J.T., Parkc, K.-H.: Singular Point Detection by Shape Analysis of Directional Fields in Fingerprints. Pattern Recognition 39, 839–855 (2006)
12. Liu, M.: Fingerprint classification based on Adaboost learning from singularity features. Pattern Recognition 43, 1062–1070 (2010)
13. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using Matlab. Prentice Hall, Englewood Cliffs (2003)
14. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer, Heidelberg (2009)

The Development of Software Evaluation and Selection Framework for Supporting COTS-Based Systems: The Theoretical Framework

Fauziah Baharom, Jamaiah Hj. Yahaya, and Feras Tarawneh

Sciences and Information Technology
College of Arts and Sciences
Universiti Utara Malaysia (UUM)
Sintok, Malaysia
{fauziah,jamaiah}@uum.edu.my,
feras79tara@hotmail.com

Abstract. As a result of increasing demands on COTS technology, there is an increasingly huge market of COTS software. Therefore, one of the most critical activities in COTS-based system development is the COTS evaluation and selection. Unfortunately, most existing methods that have been proposed in previous studies for evaluating and selecting COTS software are still have many limitations to be applicable and used in the industry. So without an effective method to select and evaluate COTS software, the time spent for selecting the correct COTS software may offset the advantages of using it. This paper outlines and discusses the common problems in existing methods and the main processes and criteria (non-functional requirements) that are required for evaluating and selecting COTS software through theoretical and empirical studies which goal is to develop new framework to evaluate and select COTS software.

Keywords: COTS; COTS evaluation and selection; theoretical framework.

1 Introduction

In the last decade, the software functionalities have become more complex because the rapidly changing in the customers' demands and the software technology in the market is evolved very fast. Therefore, a new approach has been produced as an alternative software development approach which is based on integrating pre-packaged solutions, usually known as Commercial-Off-The-Shelf (COTS) software.

COTS software is a term referred to as pieces of reused software that are developed and supported by outside suppliers (so-called vendors) to provide additional functionalities within a final system. COTS software can be a compiler, software tool and operating system. COTS-based system is developed based on selecting, adapting, and integrating one or more COTS software and this process is also called as COTS-Based System Development (CBSD) [1]. CBSD changes the way of building software

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 133–143, 2011.

development from in house-development to pre-existing COTS software which are tested many times by many other users. Thus, this approach grants the opportunity to lower the costs, time and effort for developing systems, also CBSD enhances the reliability, flexibility, and reusability of the systems [2].

However, many challenges are faced by the organizations during the development of their systems via COTS software. One of main challenges is lack of abilities to select the most suitable COTS software that meets their requirements. This challenge occurs due to many similar COTS software in the market with different capabilities and qualities characteristics [2]. In addition, any wrong decision for selecting COTS software will reflect negatively on the project as entire by increasing of the cost, time, effort, and also effect negatively on the performance and quality of the final system [3]. Therefore, most of the organizations are interested in the evaluation and selection process and considered it as one of the critical success factors in CBSD [4], [5].

However, the evaluation and selection COTS software process have many problems such as rapid evolvement of the COTS software market [6], the "Black box" nature of COTS software, evolving requirements during COTS evaluation [7], ineffective evaluation criteria, and lack of well-defined and systematic COTS software evaluation and selection process [8].

This paper presents the common problems in existing methods for evaluating and selecting COTS software, the objectives and the methodology for proposing new framework to evaluate and select COTS software based on theoretical and empirical studies. It presents the problem statements, objectives and literature review, follows by the research methodology, the theoretical framework and conclusion.

2 Problem Statements

Based on previous studies, there are many models have been proposed to handle the COTS software evaluation and selection problems, but none of these models have been accepted and considered as formal method for evaluating and selecting COTS software to be applicable in industry. The evaluation and selection of COTS software is still performed using ad-hoc manners [9], [3], [10], such as depending on the experiences of developer team or their intuition. Therefore, lack in systematic, repeatable, and well-defined process for evaluating and selecting COTS software in the industry keep the organizations under the pressure [10], and the development team has lack of experiences to plan for the selection process in detail [4].

Despite of many methods have been proposed previously to evaluate and select COTS software, there are some issues and problems that are still not considered by these methods such as identifying the mismatches between user requirements and COTS features, lack of handling non-functional requirements to distinguish between COTS software alternatives, and lack of managing and learning from previous selection cases knowledge.

2.1 Mismatches Problem

Identifying the mismatches between COTS features and customer requirements does an important role for supporting the decision making in COTS software selection process [1], [11]. Mismatches are defined as a shortages or excesses of COTS features against customer requirements [4]. Thus, addressing and better understanding of these mismatches earlier will support and provide valuable insight on the decision of COTS software selection and thereby, reduces the risk of project failure. Also most of these mismatches are solved after selecting the COTS software which makes the latter activities like adaptation and integration in CBSD easier [11], [4]. In reality, existing methods for evaluating and selecting COTS software neglect the mismatches between COTS features and customer requirements [12].

2.2 Non-functional Requirements Problem

The common limitation of current methods is more concerned on functionality and cost criteria over the non-functional requirements [13]. Non-functional requirements define the overall characteristics or attributes of the system such as quality attributes, vendor attributes, organizations attributes [14]. However, non-functional requirements play important role to distinguish similarities of COTS software alternative and facilitate the COTS software evaluation and selection [13]. Non-functional requirements are related to the software as complete characteristics rather than the individual characteristics, and often might be a deciding factor on the survival of software. Conversely, most existing methods for evaluating and selecting COTS software not provide sufficient support for these requirements [13].

2.3 Lack of Learning from Previous COTS Software Selection Cases Knowledge

Learning from past software selection cases helps both of the evaluators and decision makers in current software selection. It is necessary to know how the past software components were chosen and what are the successful criteria and techniques that were used before beginning the current evaluation and selection process. Moreover, previous selection cases provide information about set of vendor attributes that will be very important in current selection case such as vendor reputation, vendor sustainability and vendor credibility [15]. However, most of the existing methods for evaluating and selecting COTS software recommended documentation of COTS evaluation and selection process without showing what is the storing mechanism and how to manage the information. Therefore, these methods lack of benefits and lack of management of information over the previous selection cases. Thus, they are difficult to learn from previous selection cases that can support good and effective decision making [3].

3 Objectives

The main objective of this work is to propose a new framework to support and improve the COTS software evaluation and selection processes in industry. To achieve this objective several specific objectives have been addressed: (1) identifying the processes that are support the COTS software evaluation and selection; (2) to determine the criteria or requirements are required for successful evaluation and selection process; (3) to propose suitable method and technique to address the mismatches between COTS features and customer requirements; (4) to develop a simple repository to manage information from previous selection cases that will support the decision making process.

4 Literature Review

A new framework for COTS software evaluating and selecting will be proposed based on determining the main processes and factors that supporting COTS software evaluation and selection. Also the common methods for evaluating and selecting COTS software will be considered when building the framework.

4.1 Main Processes for Evaluating and Selecting COTS Software

Based on previous studies, several processes for evaluating and selecting COTS software are shared by existing methods for COTS software selecting. These processes can be ordered as iteratively, sequentially, or overlapping. However, the common processes for evaluating and selecting COTS software can be classified in terms of four general processes:

a) Preparation process: this process identifies and determines each of evaluation criteria and potential COTS software candidates to further detail evaluation. All of these are achieved by set of activities such as defining the evaluation criteria, searching COTS software alternatives, and screening the COTS software alternatives to select the most suitable alternatives that can be estimated in further detail with available resources [16].

b) Evaluation process: this process plays a vital role to determine how well each of the COTS software alternatives achieves the evaluation criteria [12]. Therefore, the main objective of this process is to estimate each COTS software alternative against the evaluation criteria in more detail and sort these alternatives based on their importance [17].

c) Selection process: the outputs of the evaluation process are several kinds of data such as facts, checklists, weights, opinions. Those kinds of data should be consolidated and interpreted into information [12]. However, the decision maker requires knowing about previous selection cases and identifying mismatches between COTS features and customer requirements in order to select the fitness COTS software. The recommendations in the last to the manager should include either use COTS software solutions or building the software [16], [17].

d) Supporting process: this process includes set of activities that support the valuation and selection processes, such of these activities: documentations and evaluation planning that includes forming the evaluation team, making the evaluation chart, and determining the stakeholders [12], [16].

4.2 COTS Software Evaluation Criteria

Establish the evaluation criteria are very important task for understanding, evaluating, and selecting the suitable COTS software. Evaluation criteria are decomposed through the evaluation criteria definition in a hierarchical decomposition, which starting from high level requirements until producing pieces of well-defined measurement information. Evaluation criteria are defined based on careful analysis of many influencing factors such as application requirements, application architecture, project objectives and constraints (budget and schedule)[17]. However, there is lack of providing a general list of evaluation criteria that can be used to evaluate and select COTS software. Therefore, the evaluation criteria can be categorized into several main groups [13]:

a) Functional requirements: that are defined as what the software expected to do, or the services that are provided by the software in order to support the goals of users and their activities and tasks [18]. Conversely, functional requirements are not considered as a distinct characteristic between COTS alternatives.

b) Non-functional requirements (NFRs): existing methods for evaluating and selecting COTS software have lack of dealing with NFRs [13]. Many empirical reports stated that there is lack or incorrect dealing with NFRs which cases failure, delays, or increases final cost and effort of projects [19]. According to Beus-Dukic [13], NFRs can be classified into four groups: (1) quality attributes (reliability, usability, maintainability), (2) architecture requirements (scalability, evolvability, portability), (3) domain requirements (majority, popularity of COTS in particular domain), (4) organizational requirements: customer organization (characteristic of existing hardware platform, legacy application kind), vendor organization (vendor stability, vendor reputation, upgrade policy of the product).

4.3 Existing Methods for Evaluating and Selecting COTS Software

Many methods have been carried out dealing with evaluation and selection COTS software in the previous studies. These methods can be clustered into requirements-driven approaches represented by Procurement-Oriented Requirements Engineering (PORE) [5] and COTS-based Requirements Engineering (CRE) [20], while Off-The-Shelf-Option framework (OTSO) [17] and Social-Technical Approach to COTS software Evaluation (STACE) [21] represented the architecture-driven approaches [3].

a) OTSO

OTSO [17] is the first widespread method for evaluating and selecting COTS software. It supports many techniques which are used for determining the evaluation criteria, cost and benefits estimation of candidates, and supports decision making like Analytical Hierarchy Process technique (AHP) [22]. OTSO method is considered as an important milestone and basis model for the other methods. However, OTSO method has several limitations such as (1) lack of considering non-functional requirements like vendor aspects which consider the functional and cost aspects; (2) it doesn't provide specific technique about how to handle the extra or unrequired features (mismatches problem); (3) it also depended on AHP technique to provide the decision making although this technique has several limitations like it not efficiently in large number of comparisons.

b) PORE

Procurement-Oriented Requirements Engineering method [5] is a template-based method to select COTS software. It is based on an iterative process of requirements elicitation and product selection. PORE method integrates set of techniques, methods and tools, such as: multi-criteria decision making techniques, knowledge engineering techniques, and requirements acquisition techniques. Also PORE method offers guidelines for designing product evaluation test cases. Conversely, PORE is not clear in specifying requirements and eliminating the products (i.e. do not capture the decision rationale). PORE based on templates to acquire and evaluate COTS alternatives, but these templates provide only initial view of steps to do a systematic evaluation.

c) STACE

Social-Technical Approach to COTS software Evaluation (STACE) [21] was developed to address the lack of attention in non-technical issues for COTS software like organization issues and social issues. On the other hand, the main limitation of this method is the lack of a process of requirements acquisition and specification. Moreover, it is not clear how to deal with mismatches problem, also this approach does not provide or use systematic analysis of COTS alternatives during the assessment when using a decision-making technique.

d) CRE

COTS-based Requirements Engineering (CRE) [20] is an iterative COTS software selection approach that chooses COTS software by rejection. CRE consider time restriction, domain coverage, vendor guaranties, and cost rating through the evaluation process. However, CRE approach does the balance between the evaluated cost and benefits without any guidance that explain how to satisfy it. Also this approach has a lack of supporting experiences and information sharing between stakeholders. Furthermore, the decision will be more complex and a large number of final situations as a resulting for dealing with large number of COTS alternatives. CRE has less ability to handle COTS software selection and it is most suitable for requirements elicitation.

In general, despite the similarities between these methods by sharing several processes, factors, and techniques, the missing issues are not considered and still not addressed by these approaches such as identifying mismatches between COTS features and customer requirements. In additional, those methods are concern on functionality and cost criteria over the non-functional criteria. Also, these methods have limitations to provide suitable and systematic mechanism to manage and learn from previous selection cases in order to support the decision making.

5 Methodology

The primary purpose of this work is to propose a new framework to support improvement in COTS software evaluation and selection processes in software industry. In order to do so, it is necessary to identify the processes and evaluation criteria that support COTS software evaluation and selection.

In this research, the deductive approach [23] will be used, because it is suitable to be applied in developing a model, where theories or concepts will be derived from theoretical and empirical findings. Then the proposed model will be applied and evaluated in real environment. This methodology consists of four stages: (1) theoretical study, (2) empirical study, (3) framework development, and (4) framework evaluating.

5.1 Theoretical Study

In this phase, previous studies will be reviewed in depth and focused on several related topics like such as COTS-based systems, evaluation and selection process, existing methods for evaluating and selecting COTS software. The main aim of this phase is to identify and analyze the common processes and evaluation criteria that have been used in evaluating and selecting COTS software. Additionally, the deep analysis on the existing methods and models for evaluating and selecting COTS software will be carried out. Also the related issues to evaluation and selection COTS software such as evaluation strategies, mechanisms, guidance, and templates that facilitate implementing evaluation and selection process will be investigated in this study.

The deliverables of this study will be a set of theoretical processes and factors for evaluation and selection COTS software, and common limitation of existing methods and models. Moreover, the questionnaire that is required in the next phase will be designed and tested using pilot study.

5.2 Empirical Study

The overall purpose of this study is to investigate the current practices of COTS software evaluation and selection which are related to the use of process, factors, and relevance issues (mechanisms, templates). It aims to determine the importance of the current theoretical processes and factors related to the evaluation and selection COTS software in practice. It is important to understand the current evaluation and selection situation in practice and the problems that are faced by the organizations.

In this phase, self-administered questionnaires will be used because it has several advantages such as cost effectiveness, ease to analysis, coverage a wide area, and it supports a high degree of secrecy [24]. The survey will be conducted in Jordan where the respondents are from IT organizations (that have experience with COTS-based systems) including the IT manager, developers, and other software practitioners. The data collected will be coded and entered to Statistical Package for Social Science (SPSS) software for analyzing it.

5.3 Framework Development

In this stage, the findings from theoretical and empirical studies such as the successful processes, factors, and related issues (templates, guidance, and mechanisms) to COTS software evaluation and selection will be used for developing a new framework. The development of new framework for evaluating and selecting COTS software aims to bridge the gap between state-of-the-art and state-of-the-practice. However, the proposed framework will be constructed by integrating set of processes and their activities, also the relationship between them will be established. The successful factors (NFRs) for evaluating and selecting COTS software will be determined and established. The suitable technique will be used for determine mismatches between COTS features and customer requirements. Moreover, a simple repository tool will developed in order to control and learn from previous selection cases to support the decision making. The proposed framework will be supported by set of guidance, mechanism, evaluation strategies and templates to facilitate evaluation and selection COTS software in real life.

5.4 Framework Evaluating

The aim of this phase is to evaluate the effectiveness and acceptability of the proposed framework in real environment. The evaluation will facilitate improvement and refinement to the proposed framework. A case study will be adopted as a qualitative method because this method is preferred when the researcher cannot control or manipulate the relevant behavioural events [25]. The evaluation process will start by determining the criteria that will be used to evaluate the framework. However, interviews will be adopted as a data collection method because its flexibility and adaptability, open-end questions will be used among the interviews in order to void the interview bias. The data will be entered into a software tool for analysis (e.g. ATALAS/it) and the modification and refinement will be conducted if required.

6 Theoretical Framework

Based on literature review, we propose a theoretical framework for evaluating and selecting COTS software which include two studies: theoretical study that focuses on three main issues: processes (activities and techniques), evaluation criteria (non-functional requirements) and previous frameworks. The second study is the empirical study that used survey and case study to investigate these elements in real life. Figure 1 shows how these studies are used to achieve the aims of this research.

The theoretical framework shows the main issues that should be considered when developing a new framework for evaluating and selecting COTS software. One of these issues is a set of processes that should be followed to select more fitness COTS software. This diagram shows all the main processes that have not been included by the most previous methods like preparing process, and supporting process. In this research, the selection process will be more focused because the final decision about selecting COTS software is prepared at this process [16]. The decision makers face many challenges to decide the suitable COTS product. The main challenge is how to identify the mismatches between the COTS software features and customer requirements to select the most fitness COTS software with minimum cost and effort to adapt and integrate with other components. Moreover, Learning from the previous COTS software evaluation and selection cases helps the evaluators and decision makers to understand how the past software components were chosen and which

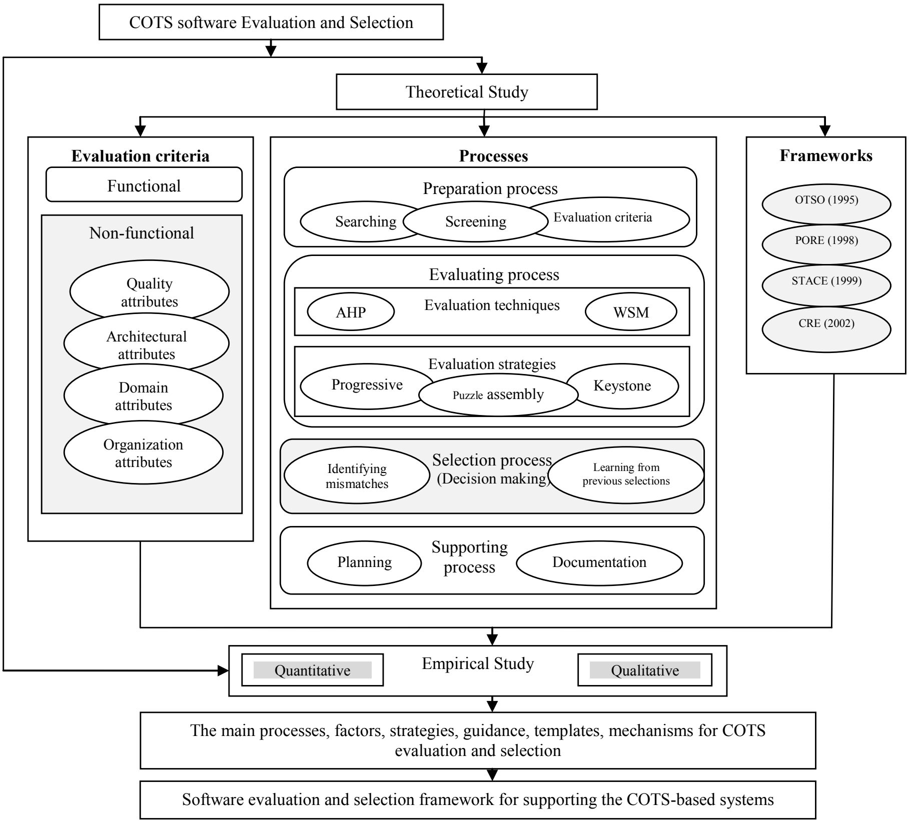


Fig. 1. The proposed theoretical framework

successful criteria and techniques were used. Known about past selection cases contributes to identify the best vendors and supports greatly the experiences of evolution and selection team [4], [15].

The evaluation criteria play a vital role during the evaluation and selection COTS software. As the theoretical framework shown, the evaluation criteria are classified into functional and non-functional criteria [14]. The non-functional criteria are considered vital because they play important role to distinguish between the COTS software such as quality attribute (reliability and efficiency), domain attributes (maturity and security), architectural attributes (portability and integrity), and organization attributes (vendor attributes) [13].

On the other hand, for eliciting and synthesise current practices of COTS evaluation and selection the empirical study will be conducted based on theoretical study by using quantitative study (questionnaire) and qualitative study (case study) to get the successful processes, criteria, techniques, strategies, and mechanisms for building a new framework for evaluating and selecting COTS software.

7 Conclusions and Future Work

This paper has investigated the common problems in the existing methods for selecting COTS software, and classified the main processes and criteria that are required for selecting COTS software. Also we propose the method for developing a new framework for evaluating and selecting COTS software. The theoretical framework of this research has also been presented. Our next step is to conduct set of questionnaires in order to examine theoretical processes and criteria, and elicit current practice of the evaluation and selection of COTS software in real life. Findings from this survey as well as the findings from literature study will be used to develop new framework for evaluating and selecting COTS software to support COTS-based system development.

References

1. Kvale, A., Li, J., Conradi, R.: A Case Study on Building COTS-Based System Using Aspect-Oriented Programming. In: Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, New, Mexico, pp. 1491–1498 (2005)
2. Wanyama, T., Far, B.: An Empirical Study to Compare Three Methods for Selecting COTS Software Components. International Journal of Computing and ICT Research 2, 34 (2008)
3. Neubauer, T., Stummer, C.: Interactive Decision Support for Multi Objective COTS Selection. In: HICSS 2007 40th Annual Hawaii International Conference on System Sciences, p. 283b (2007)
4. Mohamed, A., Ruhe, G., Eberlein, A.: Optimized Mismatch Resolution for COTS Selection. Software Process Improvement and Practice 13, 157 (2008)
5. Maiden, N., Ncube, C.: Acquiring COTS Software Selection Requirements. IEEE Software 15, 46–56 (1998)
6. Ulkuniemi, P., Seppanen, V.: COTS Component Acquisition in an Emerging Market. IEEE Software 21, 76–82 (2004)
7. Meinke, K.: A Stochastic Theory of Black-Box Software Testing. Algebra, Meaning and Computation, 578–595 (2006)
8. Vijayalakshmi, K., Ramaraj, N., Amuthakkannan, R.: Improvement of Component Selection Process Using Genetic Algorithm for Component-Based Software Development. International Journal of Information Systems and Change Management 3, 63–80 (2008)
9. Couts, C., Gerdes, P.: Integrating COTS Software: Lessons from a Large Healthcare Organization. IT Professional 12, 50–58 (2010)
10. Kunda, D.: STACE: Social Technical Approach to COTS Software Evaluation. LNCS, pp. 64–84 (2003)
11. Alves, C., Finkelstein, A.: Investigating Conflicts in COTS Decision-Making. International Journal of Software Engineering and Knowledge Engineering 13, 473–493 (2003)
12. Mohamed, A., Ruhe, G., Eberlein, A.: COTS Selection: Past, Present, and Future. In: 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS 2007), Tucson, Arizona, pp. 103–114 (2007)
13. Beus-Dukic, L.: Non-Functional Requirements for COTS Software Components. In: Proceedings of ICSE Workshop on COTS Software, pp. 4–5 (2000)

14. Pavlovski, C., Zou, J.: Non-Functional Requirements in Business Process Modelling. In: 5th Asia-Pacific Conference on Conceptual Modelling (APCCM 2008), Wollongong, NSW, Australia, pp. 103–112 (2008)
15. Alghamdi, A.: An Empirical Process for Evaluating and Selecting AWEM Environments. Evaluation Stage 19, 17–37 (2007)
16. Land, R., Blankers, L., Chaudron, M., Crnkovic, I.: COTS Selection Best Practices in Literature and in Industry. In: Mei, H. (ed.) ICSR 2008. LNCS, vol. 5030, pp. 100–111. Springer, Heidelberg (2008)
17. Kontio, J.: OTSO: A Systematic Process for Reusable Software Component Selection. University of Maryland, College Park (1995)
18. Franch, X., Botella, P.: Putting Non-Functional Requirements into Software Architecture, p. 60. IEEE Computer Society, Los Alamitos (1998)
19. Kassab, M., Daneva, M., Ormandjieva, O.: Early Quantitative Assessment of Non-Functional Requirements, Centre for Telemetric and Information Technology, University of Twente, Enschede, Technical Report TR-CTIT-07-35, Citeseer (2007)
20. Alves, C., Castro, J.: CRE: A Systematic Method for COTS Components Selection. In: de Janeiro, R. (ed.) XV Brazilian Symposium on Software Engineering (SBES), Brazil (2001)
21. Kunda, D., Brooks, L.: Applying Social-Technical Approach for COTS Selection. In: Proceedings of the 4th UKAIS Conference (1999)
22. Saaty, T.: The Analytic Hierarchy Process: Planning, Priority Setting, p. 287. MacGraw-Hill, New York (1980)
23. Trochim, W.M.K.: Deduction & Induction Thinking, `http://www.socialresearchmethods.net/kb/dedind.php` (retrieved April 10, 2010)
24. Kirakowski, J.: Questionnaires in Usability Engineering, `http://www.ucc.ie/hfrg/resources/qfaq1.html` (retrieved April 5, 2010)
25. Yin, R.: Case Study Research. In: Design and Methods, 3rd edn. Sage, London (2003)

Recipe Generation from Small Samples: Incorporating an Improved Weighted Kernel Regression with Correlation Factor

Mohd Ibrahim Shapiai[1], Zuwairie Ibrahim[1], Marzuki Khalid[1], Lee Wen Jau[2], Soon-Chuan Ong[2], and Vladimir Pavlovich[3]

[1] Centre of Artificial Intelligent and Robotics (CAIRO), Universiti Teknologi Malaysia,
Jalan Semarak, 54100, Kuala Lumpur, Malaysia
ibrahimfke@gmail.com, zuwairiee@fke.utm.my,
marzuki.khalid@utm.my

[2] ATTD Automation (APAC) Pathfinding,
Intel Technology Sdn. Bhd. Kulim, Penang, Malaysia
{wen.jau.lee,soon.chuan.ong}@intel.com

[3] Department of Computer Science, Rutgers University, NJ 08854, New Jersey, United States
vladimir@cs.rutgers

Abstract. The cost of the experimental setup during the assembly process development of a chipset, particularly the under-fill process, can often result in insufficient data samples. In INTEL Malaysia, for example, the historical chipset data from an under-fill process consist of only a few samples. As a result, existing machine learning algorithms cannot be applied in this setting. To solve this problem, predictive modeling algorithm called Weighted Kernel Regression with correlation factor (WKRCF), which is based on Nadaraya-Watson kernel regression (NWKR), is proposed. The correlation factor reflected the important features by changing the bandwidth of the kernel as a function of the output. Even though only four samples are used during the training stage, the WKRCF provides an accurate prediction as compared with other techniques including the NWKR and the artificial neural networks with back-propagation algorithm (ANNBP). Thus, the proposed approach is beneficial for recipe generation in an assembly process development.

Keywords: Recipe Generation, Predictive Modeling, Weighted Kernel Regression, Small Samples, Correlation Factor.

1 Introduction

Recipe generation provides the key references needed by engineers to set up a new experiment for a new product and plays an important role in determining the success of product development. Currently, the ingredients chosen for the recipe mainly depend on the engineer's knowledge. Optimizing the input parameters will facilitate the engineering decision needed to fulfill certain requirements. As the assembly process for the chipset is rapidly progressing towards smaller scales and greater

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 144–154, 2011.

complexity, the accuracy and efficiency requirement are more vital. For example, a semiconductor process flow requires about hundreds of fabrication operations steps with a lead-time of a few months. In addition, device fabrication and manufacturing costs continue to escalate. In addition to the usual strategy of increasing the wafer size and shrinking devices to reduce the cost per transistor, automation and modeling are becoming more important. Fowler [1] has revealed that the productivity improvement strategy of a semiconductor manufacturing is based on operational improvement at the front-end of wafer fabrication; this strategy accounts for almost half of the total annual productivity improvement target.

The use of artificial intelligence techniques for process modeling during the downstream assembly and all the involved tests is expected to reduce the overall manufacturing cost. Introducing intelligent modeling to the assembly process promises to accelerate the engineering decisions even at early stages when very few collected samples are available. Inherently, intelligent modeling can improve equipment and resource utilization. The uses of the existing algorithms such as Gradient Boosting Trees (GBT) and Random Forest (RF) have enjoyed some successes for Intel class test yield prediction and analysis [2-3]. In general, the development of recipe generation for assembly processes requires only limited samples. However, most of the existing algorithms [4] and ANNBP [5-6] are hindered by the limited number of available samples. Hence, we also highlighted the limitation of ANNBP especially when dealing with small samples due to the non-deterministic nature [7] of the ANNBP. In other words, the performance of existing algorithms degrades because the sample size is insufficient [8-9]. Also, it is difficult to obtain an accurate model [6] and create the conflict between the number of samples and the complexity of the algorithms when dealing with fewer data [10].

In INTEL Malaysia, the under-fill process shown in Figure 1, which consists of six input parameters with a small and sparse data set, is considered. Those input parameters are die size (dimension of die), gap height, the number of bumps, dispense distance, dispense weight, and the output is the dispense tongue length. In practice, it is difficult to define the input-output relationship, and improperly determined input setting parameters frequently cause the yield to be ‘excess epoxy’, ‘epoxy on die’, or ‘insufficient epoxy’. Notably, the experiment usually involves large samples, and it is rather expensive to determine the recipe that prevents the tongue generated during the under-fill process from touching the keep out zone (KOZ), as illustrated in Figure 2. Hence, it is important to develop a cost-effective method to arrive at the optimal setting.

The objective of this study is to construct a predictive model of tongue length that allows the determination of various input parameter settings. The established model can be employed later by engineers during recipe generation to determine the feasible input setting parameters.

This problem being solved can be categorized as learning from small samples and this problem has gained increasing attention in many fields, such as in assembly process sparse prediction modeling [4], biological activities prediction [11], engine control simulation [9] and pulp and paper industry [12]. A predictive algorithm proposed in this study is based on the NWKR [8] to solve small samples problem. The implementation of WKRCF is different in such a way that the observed samples are defined in the kernel matrix. Also, the correlation factor between input and output which is obtained from insufficient samples is embedded to the kernel matrix.

The remainder of this paper is organized as follows. A brief review of the NWKR is given in Section 2. The proposed WKRCF is presented in Section 3. Section 4 includes the implementation of the proposed approach and the experimental results. Finally, the conclusions are provided in Section 5.

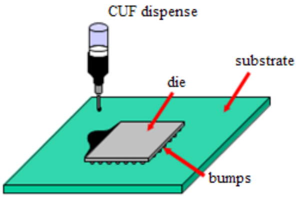


Fig. 1. Illustration of an under-fill process in an assembly process

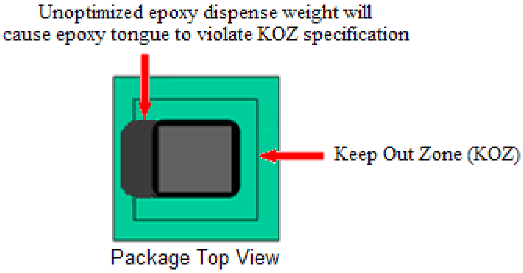

Unoptimized epoxy dispense weight will cause epoxy tongue to violate KOZ specification
Keep Out Zone (KOZ)
Package Top View

Fig. 2. Illustration of an epoxy tongue that touch the keep out zone

2 Nadaraya Watson Kernel Regression

Kernel regression, particularly the NWKR [13-14], is a non-parametric technique in statistics used to estimate the conditional expectation of a random variable. It is an estimation technique used to fit the given samples. It finds a function, *f(.)* such that the approximated function is best-fit to match the given samples. Thus, it allows interpolation and approximation somewhat beyond the samples. A Gaussian kernel, for example, assigns weights to any arbitrary samples based on the distance from the given samples based on the Eq. (1):

$$K(x, x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x - x_i)^2}{h}\right) \tag{1}$$

where h denotes the smoothing parameter, and K is a Gaussian kernel that is used to assign a weight, which is based on Euclidean distance, to any arbitrary samples. The closer the arbitrary sample is to any given samples, the heavier its assigned weight is.

x_i is a list of observed independent variables and x is an arbitrary point to be estimated.

The dependent variable, $\hat{y}$, corresponds to any arbitrary x values and can be estimated by using the following equations: Eq. (2), Eq. (3) and Eq. (4)

$$\hat{y}(x) = \sum_{i=1}^{n} y_i \hat{w}(x, x_i) \tag{2}$$

$$\hat{w}(x, x_i) = \frac{K(x, x_i)}{\sum_{i=1}^{n} K(x, x_i)} \tag{3}$$

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} y_i K(x, x_i)}{\sum_{i=1}^{n} K(x, x_i)} \tag{4}$$

where n is the number of observed samples. For the estimation more than two dimensions, consider d-dimensional, the i$^{\text{th}}$ observation for each of d independent variable is given in the vector X_i as given in Eq. (5).

$$X_i = \begin{bmatrix} X_i^1 \\ \vdots \\ X_i^p \\ \vdots \\ X_i^d \end{bmatrix} \quad i = 1,2,\ldots,n \tag{5}$$

The estimated value of $\hat{y}$ can be calculated using Eq. (6).

$$\hat{y}(X, X_i) = \frac{\sum_{i=1}^{n} y_i \left(\prod_{p=1}^{d} K(X^p, X_i^p) \right)}{\sum_{i=1}^{n} \left(\prod_{p=1}^{d} K(X^p, X_i^p) \right)} \tag{6}$$

3 The Proposed Weighted Kernel Regression with Correlation Factor

An overview of the proposed technique is given in Figure 3. The proposed technique requires a series of steps to develop the prediction model.

3.1 Training Phase

With an insufficient number of samples, popular model selection methods such as cross validation cannot be used [15-16]. As for NWKR, it is important to compromise between smoothness and fitness in selecting the smoothing parameter h [17]. The WKRCF provides an easy method of tuning the hyper-parameters of the proposed model when dealing with small samples. The smoothing parameter for the proposed technique can be estimated using Eq.(7)

$$h = \sum_{i=1}^{n} \left(\|X_i\|^2 - \overline{\|X\|^2} \right)^2 \tag{7}$$

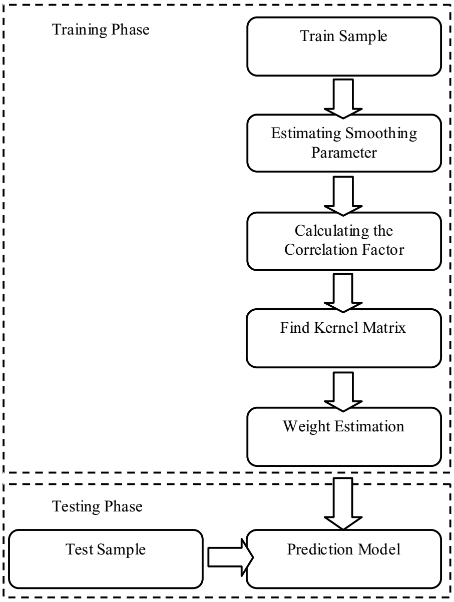


Fig. 3. Overview of the proposed technique, WKRCF

The correlation factor is introduced to adaptively set the smoothing parameter, h, of the Gaussian Kernel Function for WKRCF. Initially, the correlation coefficient for each input parameter must be calculated as follows

$$r_{x_p y} = \frac{\operatorname{cov}(x_i, y)}{\sigma_{x_p} \sigma_y} = \frac{E\{(x_p - \bar{x}_p)(y - \bar{y})\}}{\sigma_{x_p} \sigma_y} \tag{8}$$

where $r_{x_p y}$ is the correlation coefficient, x_p and $\overline{x}_p$ are the input value of one particular dimension and the corresponding mean value of the set of x_p respectively, y and $\overline{y}$ are the output value and the corresponding mean value of the set of y , σ_{x_p} is the standard deviation of x_p and σ_y is the standard deviation of y .

Thus, the correlation factor, c_p , for each input parameter is then can be defined as:

$$c_p = \frac{r_{x_p y}}{\sum_{k=1}^{d} r_{x_k y}} \tag{9}$$

The adaptive Gaussian kernel and the corresponding kernel matrix, *A = [a*a_{ij}*],* where *i = j = 1,..., n,,* are given by Eq. (10), and Eq. (11), respectively.

$$K(x, x_i, c_p) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x - x_i)^2}{h * c_p} \right) \tag{10}$$

$$a_{ij} = \begin{cases} \dfrac{\prod_{p=1}^{d} K(X_i^p, X_j^p, c_p)}{\sum_{l=1}^{n} \left[\prod_{p=1}^{d} K(X_{i \vee j}^p, X_l^p, c_p) \right]} & if\ i \neq j \\ \dfrac{1}{\sum_{l=1}^{n} \left[\prod_{p=1}^{d} K(X_{i \vee j}^p, X_l^p, c_p) \right]} & if\ i = j \end{cases} \tag{11}$$

The kernel matrix A transforms the linearity of the observed samples to nonlinear problems by mapping the data into a higher dimensional feature space based on Eq. (3) with subject to the correlation factor.

Once the kernel matrix is found, it is necessary to introduce the estimated weight. The estimated weight is determined from the kernel matrix. The weight is updated iteratively by comparing the estimated values $\hat{y}_i$ to the actual value y_i. When the difference converges to a minimum value or after reaching the predefined iteration value, the training to estimate the weight will be stopped. Initially, arbitrary values are assigned to the weights. The weight is defined in a column vector, as shown in Eq. (12).

$$W = [w_1 \quad w_2 \quad \cdots \quad w_n]^T \tag{12}$$

The estimated $\hat{y}_i$, the error equation and the estimated weight equation are given by Eq. (13), Eq. (14), and Eq. (15), respectively.

$$\hat{y}_i = \sum_{i=1}^{n} w_j a_{ij} \tag{13}$$

$$E(W) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i) \tag{14}$$

$$\hat{W}(X) = \arg\min_{W} E(W) \tag{15}$$

3.2 Testing Phase

Once the optimum weight is obtained, the model is ready to predict any unseen samples (test samples). The test samples can be predicted using Eq. (16).

$$\hat{y}(X, \hat{W}) = \frac{\sum_{i=1}^{n} \hat{w}_i \left(\prod_{p=1}^{d} K\left(X^p, X_i^p, c_p\right) \right)}{\sum_{i=1}^{n} \left(\prod_{p=1}^{d} K\left(X^p, X_i^p, c_p\right) \right)} \tag{16}$$

4 Experiment and Results

4.1 Experiment Setup

Initially, all the parameter settings for each predictive modeling algorithm are predefined. The parameter settings are summarised in Table 1.

4.2 Performance Measure

A simple but useful concept from [4] is used to evaluate the performance of the prediction based on the error of the acceptance rate, E, within the accuracy of the guard band, B, as given in Eq. (17).

$$E = \left| \frac{predict - actual}{predict} \right| \times 100\% \le B \tag{17}$$

Within a specified acceptance rate, the coverage accuracy, as given in Equation (18), is calculated to determine how many samples fulfill the setting of guard band value.

$$C = \frac{total\ number\ of\ accepts}{total\ number\ of\ predictions}\times 100\% \qquad (18)$$

Table 1. Parameter settings for each of the function approximation algorithms

Technique	Parameter Settings
WKRCF	$h=\sum_{i=1}^{n}\left(\lVert X_i\rVert^2-\overline{\lVert X\rVert^2}\right)^2$, iteration = 1000 (whichever is reached first)
NWKR	$h=\sum_{i=1}^{n}\left(\lVert X_i\rVert^2-\overline{\lVert X\rVert^2}\right)^2$
ANNBP	Input Layer (6 nodes), One Hidden Layer (15 nodes with sigmoid function), Output Layer (1 node with linear function), momentum rate = 0.9, learning rate = 0.7 and stopping criteria either training error MSE < 10e-6 or iteration = 1000 (whichever is reached first)

4.3 Results

The historical DOE data set obtained from INTEL Malaysia, shown in Table 2, was employed in the experiment. The total number of available samples was ten and four training samples from the first four rows were chosen because those training samples covered the minimum and the maximum range of the input and output values. This was a relevant assumption because the problem became an interpolation problem

Table 2. The historical DOE data set; x and y are the dimension sizes, gh is the gap height, nb, the number of bumps, dd, the distance dispense, sw, the amount of epoxy; the output, the length of the tongue

x	*y*	*gh*	*nb*	*dd*	*sw*	*output*
14795.66	13475.28	3035.64	6782064	67870	61700	256305.3
17238.98	17238.98	3134.36	6782064	80210	49360	166709.3
6170	6170	3072.66	662658	49360	17276	114980.7
16671.34	16362.84	3356.48	6415566	74040	61700	250800.1
14795.66	13475.28	3035.64	6782064	67870	49360	237581.9
17238.98	17238.98	3134.36	6782064	80210	55530	243672.4
16671.34	16362.84	3356.48	6415566	74040	49360	215971.4
16671.34	16362.84	3356.48	6415566	67870	61700	246692
14795.66	13475.28	3035.64	6782064	67870	40722	199574.8
14795.66	13475.28	3035.64	6782064	67870	57998	251815.5

based on the observed samples. The calculated correlation factors are given in Table 3. The remaining samples were then used to measure the performance of the proposed model. The tabulated results are shown in Table 4.

The presented results show the coverage accuracy of the proposed model for three different guard band values. The WKRCF achieves the highest coverage accuracy even when dealing with very small and sparse dataset. It is also found that ANNBP tended to produce inconsistent predictions when dealing with small samples. Hence, we only reported the best coverage accuracy for ANNBP in this study. Meanwhile, NWKR has the over-smoothing and over-fitting effects due to the limited samples. As a result, NWKR has the worst prediction quality.

Table 3. Calculated correlation factor of the Intel Dataset

	c_1	c_2	c_3	c_4	c_5	c_6
Correlation Factor	0.1762	0.1570	0.0955	0.1983	0.1364	0.2365

Table 4. The coverage accuracy of the presented techniques

Technique	*Sample Size*		*Coverage Accuracy, C (%)*		
	Train	Test	B=8%	B=12%	B=15%
WKRCF	4	6	50	100	100
NWKR	4	6	16.67	33.33	33.33
ANNBP	4	6	50	83.33	83.33

Introducing the correlation factor reflected the particular input relationship against the output by contributing more weight in predicting the output. Theoretically, the introduction of correlation factor agreed with the nature of the dataset, in which the length of the tongue from the under-fill process is highly correlated with the amount of the dispensed epoxy. In other words, the ‘sw’ input feature contributed substantially to the length of the tongue as shown in Table 3. Explicitly, the calculated coverage accuracy also agrees with the assumption of the correlation factor.

The chosen guard band values provide an indicator for the engineer and facilitate the establishment of a new experiment for a new product at a certain confidence level. As a result, the experiment conducted here to model the under-fill process will allow the full use of resources and indirectly reduce costs by creating a recipe from the proposed model.

5 Conclusion

Because of limited information, learning from small samples is extremely difficult, especially the under-fill process of an assembly process. This study shows that the modified version of kernel regression, namely WKRCF is superior to existing technique. The WKRCF requires a training process to find the optimum weight before the model is ready to use for the recipe generation process development. Incorporating the correlation factor to the kernel matrix reflected the dependencies of the highly correlated input parameter to the output. This assumption provides necessary information when there is no training sample available. In the future, a technique to systematically generate artificial samples will be investigated to increase the number of relevant samples and thereby improve the prediction of the model.

Acknowledgments. This work was financially supported by the UTM-INTEL Research Fund (Vote 73332) and the Ministry of Higher Education (MOHE) by the Fundamental Research Grant Scheme (FRGS) (Vote 78564).

References

1. Fowler, J.W.: Modeling and Analysis of Semiconductor Manufacturing. Dagstuhl Seminar (2002)
2. Kuan, Y.W., Chew, L.C., Jau, L.W.: Method for proposing sort screen thresholds based on modeling etest/sort-class in semiconductor manufacturing. In: Automation Science and Engineering, CASE 2008, pp. 236–241. IEEE, Los Alamitos (2008)
3. Yip, W., Law, K., Lee, W.: Forecasting Final/Class Yield Based on Fabrication Process E-Test and Sort Data. In: Automation Science and Engineering, CASE 2007, pp. 478–483. IEEE, Los Alamitos (2007)
4. Lee, W., Ong, S.: Learning from small data sets to improve assembly semiconductor manufacturing processes. In: 2nd ICCAE 2010, pp. 50–54 (2010)
5. Huang, C., Moraga, C.: A diffusion-neural-network for learning from small samples. International Journal of Approximate Reasoning 35, 137–161 (2004)
6. Zhou, J., Huang, J.: Incorporating priori knowledge into linear programming support vector regression. In: Computing and Integrated Systems (ICISS), pp. 591–595. IEEE, Los Alamitos (2010)
7. Graczyk, M., Lasota, T., Telec, Z., Trawiński, B.: Nonparametric statistical analysis of machine learning algorithms for regression problems. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6276, pp. 111–120. Springer, Heidelberg (2010)
8. Shapiai, M., Ibrahim, Z., Khalid, M., Jau, L., Pavlovich, V.: A Non-linear Function Approximation from Small Samples Based on Nadaraya-Watson Kernel Regression. In: 2nd International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN 2010), pp. 28–32. IEEE, Los Alamitos (2010)
9. Bloch, G., Lauer, F., Colin, G., Chamaillard, Y.: Support vector regression from simulation data and few experimental samples. Information Sciences 178, 3813–3827 (2008)
10. Sun, Z., Zhang, Z., Wang, H.: Incorporating prior knowledge into kernel based regression. Acta Automatica Sinica 34, 1515–1521 (2008)
11. Andonie, R., Fabry-Asztalos, L., Abdul-Wahid, C., Abdul-Wahid, S., Barker, G., Magill, L.: Fuzzy ARTMAP prediction of biological activities for potential HIV-1 protease inhibitors using a small molecular dataset. IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics (2009)
12. Lanouette, R., Thibault, J., Valade, J.: Process modeling with neural networks using small experimental datasets. Computers & Chemical Engineering 23, 1167–1176 (1999)

13. Watson, G.: Smooth regression analysis. Sankhy: The Indian Journal of Statistics, Series A 26, 359–372 (1964)
14. Nadaraya, È.: On estimating regression. Teoriya Veroyatnostei i ee Primeneniya 9, 157–159 (1964)
15. Jang, M., Cho, S.: Observational Learning Algorithm for an Ensemble of Neural Networks. Pattern Analysis & Applications 5, 154–167 (2002)
16. Isaksson, A., Wallman, M., Göransson, H., Gustafsson, M.: Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognition Letters 29, 1960–1965 (2008)
17. Zhang, J., Huang, X., Zhou, C.: An improved kernel regression method based on Taylor expansion. Applied Mathematics and Computation 193, 419–429 (2007)

Applying Feature Selection Methods to Improve the Predictive Model of a Direct Marketing Problem

Ding-Wen Tan[1], Yee-Wai Sim[2], and William Yeoh[3]

[1] Department of Physical and Mathematical Science, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
tandw@utar.edu.my
[2] Department of Information Systems, Universiti Tunku Abdul Rahman, Jalan Universiti, Bandar Barat, 31900 Kampar, Perak, Malaysia
simyw@utar.edu.my
[3] School of Information Systems, Deakin University, 70 Elgar Road, Burwood, Victoria 3125, Australia
william.yeoh@deakin.edu.au

Abstract. The ability to forecast job advertisement demands is vital to enhance the customer retention rate for recruitment companies. On top of that, it is uneconomical to cold call every individual on a regular basis for companies with a large pool of customers. This paper presents a novel approach in predicting the re-ordering demand of a potential group of SMEs customers in a large online recruitment company. Two feature selection techniques, namely Correlation-based Feature Selection (CFS) and Subset Consistency (SC) Feature Selection, were applied to predictive models in this study. The predictive models were compared with other similar models in the absence of feature selections. Results of various experiments show that those models using feature selections generally outperform those without feature selections. The results support the authors' hypothesis that the predictive model can perform better and further ahead than similar methods that exclude feature selection.

Keywords: data mining, direct marketing, feature selection, artificial neural networks, decision trees.

1 Introduction

Recently data mining approaches have been widely used by contemporary enterprises in an effort to enhance the customer retention rate. For example, data mining techniques were applied in direct mailing campaigns (as reported in Kim & Street [1] and Bella et al. [2]), and in telecommunication service and financial industries for churn prediction and customer retention (as presented in Cox[3], Hung et al. [4], Au & Chan [5], and Kumar & Ravi [6]).

However, there is limited research in the area of highly imbalanced class distribution and non-binary classification related to direct marketing, particularly for online job recruitment companies. One of the typical services provided by an online recruitment company is online job advertisement posting on its websites. In this

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 155–167, 2011.

research, the authors were expected to assist an online job recruitment company to increase the retention rate of a particular group of SMEs customers (known as group *X* in this paper). In fact, the company is one of the largest online job recruitment companies in Southeast Asia. The definition of group *X* follows certain criteria determined by the company internally. Normally, those relatively small companies or those companies which contribute relatively small amount of purchases are categorized into group *X*. The reason why the company suffers from low retention rate of group *X* is because the customers are generally having less demand for posting job advertisements and not buying the job advertisement package regularly as compared to large companies. As a result, it is difficult to predict when the customers want to buy a job advertisement package. Therefore, this research aims to build a predictive model to overcome this problem.

Several challenges exist in this research, as discussed in Ling & Li [7]. First issue is the highly imbalanced class distribution. The ratios of "Yes" customers (positive instances) to all customers in both datasets for training and testing are only 4.73% (538 "Yes" customers out of all 11377 customers) and 9.92% (1507 "Yes" customers out of all 15199 customers), respectively. Consequently, some predictive models would give meaningless result that all customers are always classified as "No" customers (negative instances). The second problem is that if binary classification is used then only two classes, "Yes" customers and "No" customers, will be found in the result. If the telesales team of the company wants to make calls to about 10% of customers for selling job advertisement packages, but only 5% of customers are classified as "Yes" customers, then binary classification is not a good choice. The solution proposed by Ling & Li [7] is to use data mining algorithms which can produce probability estimate, certainty factor, or any other confidence measure to rank the customers so that a ranked list of customers from most likely buyers to least likely buyers can be generated according to the likelihood. Thus, the company is able to pick any number of likely buyers from the ranked list for telesales.

In a data mining process, there is a technique named feature selection used for selecting an appropriate subset of features from the full set of all available features. Such technique is used to build a more robust learning model. Basically, the main objective of applying a feature selection technique is to enhance the performance of a predictive model by selecting and attaining the optimal or near-optimal feature subset. Feature selection has been widely investigated and discussed by researchers, such as Blum & Langley [8] and Kohavi & John [9]. In this study, one of the main objectives is to apply the two feature selection methods and determine if they can improve the performance of the predictive model.

The rest of this paper comprises four main sections. Section 2 provides literature review on feature selection. Section 3 explains the process of data collection, followed by description of how the experiments were conducted. Section 4 discusses the details of selected evaluation approaches and presents the results. In section 5, the performance of the models with and without feature selection methods are compared, followed by conclusions of this study.

2 Literature Review

Pattern recognition or classification has been widely applied in various fields including computer vision, bioinformatics, data mining, and marketing [10]. According to Jain et al. [10], feature selection problem is defined as selection of a feature subset which can result in the minimum classification error from the full set of features. In other words, feature selection can also be simply defined as the effort to get a set of features (from all available features) that enables the predictive model in producing robust results, as applied in this study.

According to Guyon & Elisseeff [11], in most cases, a search strategy and an evaluator are the two main components of a feature selection technique. A search strategy performs a feature subset search within the full set of all available features. Whilst an evaluator assesses the performance of the feature subset produced by the search strategy.

Elashoff et al. [12] and Toussaint [13] demonstrated that the best feature subset may not necessarily contain the best feature. In addition, Guyon & Elisseeff [11] provided a finding which states that combination of a desired or undesired feature with another undesired feature can produce a feature subset of good performance. Therefore, to evaluate whether a feature subset is good or not, the merit which belongs to the whole feature subset should be considered, instead of simply selecting the best features (from the ranked list generated by any single feature evaluator) to produce a so-called “best feature subset”.

In the work of Tirelli & Pessani [14], four single feature evaluators were used to assess each feature of the nineteen features. The final feature subset consisted of nine features. However, the reason why the authors chose the nine features was that all the nine features were the top nine features on the ranked lists generated by the four evaluators. It is not so suitable to apply single feature evaluator in that way. For instance, a feature subset of less features, say, the combination of the top five features on the ranking generated by information gain method, may result in better performance. Any single feature evaluator cannot tell which feature subset with at least two features is the best. Hence, to make use of the ranked list generated by a single feature evaluator, the authors suggested two approaches:

1. To choose a feature subset evaluator and an appropriate search method, such as forward selection, backward elimination, and “RankSearch” reported in Witten & Frank [15].
2. To use the top feature on the ranked list as input feature and adopt a learning algorithm particularly for a single feature, like 1R (see Holte [16]).

In addition, feature selection can be categorized into embedded, wrapper, and filter models as discussed in Guyon & Elisseeff [11] and Liu & Motoda [17]. If a feature selection process was included in the predictive model building process, such approach will be classified as embedded model. As for the wrapper model, it evaluates each selected feature subset by running a predictive model on them. On the other hand, a filter model is similar to the wrapper model, but instead of evaluating against a model, a simpler filter is used to test the effectiveness of the selected feature subset.

In this study, filter models were adopted due to the consideration of expense in computation as recommended by Chen & Liu [18]. This is because wrapper models can be computationally expensive when compared to the other two models, especially if those complex predictive models such as Artificial Neural Networks and Support Vector Machine are used and the dataset involved is large. Two techniques that fall under the filter model category, namely Correlation-based Feature Selection (CFS) and Subset Consistency (SC), were applied to the predictive models and their performances were compared. On the other hand, Decision Tree (DT) learning algorithm was used to build predictive model in this study. In fact, DT is an embedded model of feature selection. It selects desired features based on information gain and builds a DT model at the same time. Obviously, DT learning algorithm has its own way to select features. However, Liu & Setiono [19] and Tirelli & Pessani [14] gave examples showing that DT model with an additional feature selection method generally outperformed the one without any additional feature selection method. Thus, the authors wanted to see if the two chosen feature selection methods can improve the performance of their DT models.

According to Hall [20], CFS measures how good a set of features is by assessing the predictive ability of every feature separately while taking account the redundancy level among them at the same time. CFS is always searching for an ideal feature subset which comprises features that are highly correlated with the target feature, and they are not intercorrelated or slightly intercorrelated. Besides, experiments conducted by Hall [20] showed that CFS could significantly reduce the number of input features and preserve or improve the performance of a predictive model at the same time.

As for SC, WEKA provides an evaluator to measure the degree of inconsistency of a feature subset with respect to the target values. More precisely, according to Liu & Setiono [19], if two instances hold exactly the same values for all features but have different target values, they are said to be inconsistent. Suppose k is the number of the group of inconsistent instances. For each group of inconsistent instances, the inconsistency count is calculated by subtracting the largest number of instances which hold the same target values from the number of the instances. The following is a formula for computing the inconsistency rate of a feature subset,

$$Inconsistency\ Rate = (IC_1 + IC_2 + \ldots + IC_k) / N ,$$

where IC denotes the inconsistency count and N is the total number of instances. Since the full feature set always has the highest consistency, SC can be used to look for the smallest feature subset which has the same consistency as that of the full feature set. Moreover, Liu & Setiono [19] did some experiments to show that SC was able to reduce the number of features significantly, and the performances of both ID3 and C4.5 learning algorithms were improved after using SC.

3 Data Collection and Experiments

As for the data set in this study, the authors had randomly selected training instances from 11,377 available instances in building the predictive model and employed another set of 15,199 available instances for the experiment of predicting re-ordering demands. Each instance had two kinds of features: characteristic and dynamic features. Characteristic feature describes the nature of a customer and it is mostly static, such as customer's industry type. Dynamic feature describes customer's "behaviour" in a certain period.

Basically, the data mining process of this research followed the "CRoss Industry Standard Process for Data Mining" (CRISP-DM) as described in Chapman et al. [21]. In doing so, first, business problem and objectives of this research were identified. Then, data for this research were studied and cleaned to ensure high quality data are ready for modeling and evaluation process.

The telesales team that participated in this research failed to identify the ideal time to call some existing customers who were having reorder demand. In other words, customer retention rate was low. Thus, the company needs a predictive model to figure out the propriety and time to offer existing customers with new products.

To illustrate, assume that it is now Dec 31, 2009. The authors would like to predict and target likely buyers of the company's products in the following month, January 2010. Thus, in the dataset for training, target feature "Cat_Ord" indicates if a customer bought at least one product (job advertisement package) in January 2009 whereas target feature "Cat_Ord" in the dataset for testing represents the same customer's purchasing record, but it is for January 2010. If a customer fulfils the above-mentioned criterion, it is labeled as "Yes" customer; otherwise, "No" customer. Incidentally, all the dynamic features were retrieved based on the target feature. For example, the feature "Ord1mb4" in the training set holds the number of product purchased by a customer in Dec 2008.

Ling & Li [7] provided examples and showed that the ratio of positive instances to negative instances needs to be 1 in order to obtain the best result (the best lift index). In these experiments, training set comprised all 538 "Yes" instances and 538 "No" instances which had been selected randomly. Five different random seeds were used to generate five different training sets.

On the other hand, the authors built predictive models with and without feature selection, and the models were evaluated based on three measures, namely area under the ROC curve, lift index, and precision for the top decile.

Artificial Neural Networks (ANNs) and Decision Trees (DTs) were used to build predictive model. According to the comprehensive literature review of Ngai et al. [22], ANNs and DTs were the top two frequently used learning algorithms in the context of customer relationship management (CRM). Since the main purpose of this study is to investigate the usefulness of feature selection in the authors' data mining research, the parameters of both learning algorithms were fixed. Figure 1 illustrates the process of model building. Five training sets were generated randomly and duplicated. Each batch of the training sets went through the process of feature selection with different feature selection methods. After that, the training sets with selected features would be duplicated again for model building by using the learning algorithms. Since there were 5 different training sets, 3 feature selection methods, and 2 learning algorithms, 30 models were built in these experiments.

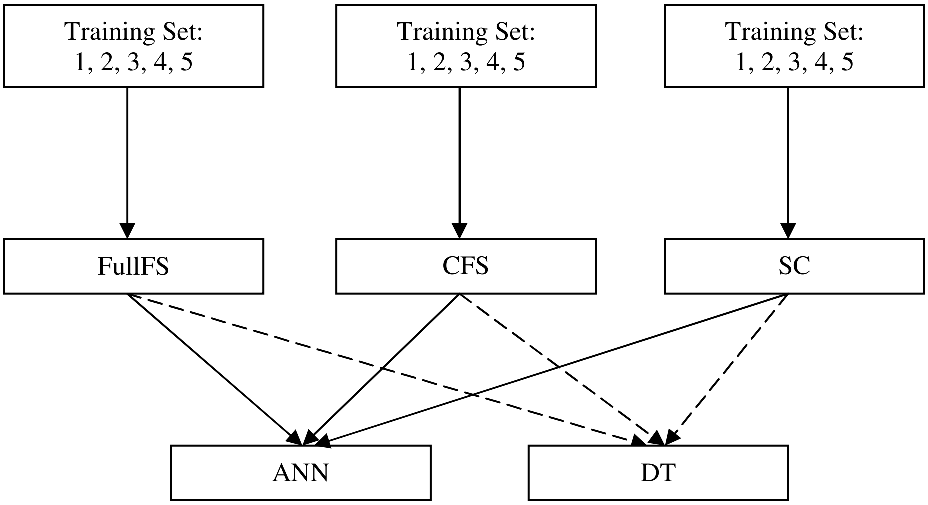


Fig. 1. The model building process (FullFS = full feature set (without feature selection); CFS = correlation-based feature selection; SC = subset consistency; ANN = artificial neural network; DT = decision tree)

4 Evaluations and Outcomes

This section introduces the ways to evaluate the predictive models and presents all of the experiment results.

4.1 Prediction Accuracy

Generally, prediction accuracy or overall success rate can be computed by dividing the number of correct classifications by the total number of instances as described in Witten & Frank [15]. Here, the prediction accuracy can be expressed as the following equation.

$$Prediction\ Accuracy = \frac{Number\ of\ true\ "Yes"\ customers + Number\ of\ true\ "No"\ customers}{Total\ number\ of\ customers} \tag{1}$$

It is not so suitable for evaluating the models' performance of this study because of the highly imbalanced data. Only 9.92% of all customers are "Yes" customers and the rest are "No" customers. If a model simply classify all customers as "No" customers then the classification accuracy or prediction accuracy is very high, that is, 90.08%. It is meaningless since the main concern here is "Yes" customer.

4.2 Lift Index

Ling & Li [7] suggested lift index instead of prediction accuracy if someone is dealing with highly imbalanced data provided that the corresponding predictive model is able to rank customers by certain likelihood or confidence measure. The lift index was defined as

$$S_{lift} = (1 \times S_1 + 0.9 \times S_2 + \ldots + 0.1 \times S_{10}) / \sum_{i}^{10} S_i \text{ ,} \tag{2}$$

where $i = 1, 2, 3, \ldots, 10$; S_i denotes the number of "Yes" customers in $i\text{-}th$ decile of the ranked list of likely buyers.

4.3 Area under the ROC Curve (AUC)

As defined by Witten & Frank [15], a Receiver Operating Characteristic (ROC) curve is plotted by true positive (TP) rate versus false positive (FP) rate. In this research,

$$TP\ rate = \frac{Number\ of\ true\ \text{"Yes"}\ customers(TP)}{Number\ of\ true\ \text{"Yes"}\ customers\,(TP) + Number\ of\ false\ \text{"No"}\ customers(FN)} \times 100\% \tag{3}$$

$$FP\ rate = \frac{Number\ of\ false\ \text{"Yes"}\ customers\,(FP)}{Number\ of\ false\ \text{"Yes"}\ customers\,(FP) + Number\ of\ true\ \text{"No"}\ customers\,(TN)} \times 100\% \tag{4}$$

Provost & Fawcett [23] and Huang & Ling [24] suggested using ROC curve and the area under ROC curve (AUC) instead of prediction accuracy to evaluate learning algorithms. In short, the simple rule is that with a bigger AUC, the performance is better.

4.4 Precision for the Top Decile

According to Witten & Frank [15], precision can be defined as

$$Precision = \frac{Number\ of\ true\ \text{"Yes"}\ customers\,(TP)}{Number\ of\ true\ \text{"Yes"}\ customers\,(TP) + Number\ of\ false\ \text{"Yes"}\ customers\,(FP)} \times 100\% \tag{5}$$

In this research, the authors paid their attention to the precision for the top decile of likely buyers on the ranked list. Since, the telesales team of the job ad company usually would call only 10% of all of the customers in group *X* because of the constraints of calling expenses and manpower. This measure was used to assess each predictive model built for this research and compared with the distribution rate of true "Yes" customers in testing set, 9.92%.

Two open source tools, KNIME (2.2.2) and Waikato Environment for Knowledge Analysis (WEKA) (3.6) were used to do and complete all of the experiments. In particular, the authors used KNIME to complete data preparation and calculate the area under ROC curve. On the other hand, all of the feature selection methods and learning algorithms (for model building) were provided by WEKA. Both tools were available on their official websites.

4.5 Feature Subsets Selected by the FS Methods

The original feature set consisted of 19 features in total. Detailed description of each available feature is in Table 1. The authors used exhaustive search and the two evaluators, CFS and SC, to select feature subsets for the five training sets. The results are shown in Table 2.

Table 1. Description of feature

Feature	Description	Value
Cat_Ord	It shows if the customer bought a product in the target month T.	Nominal feature. Yes or No.
Industry	It is the industry that the customer belongs to.	Nominal feature. 60 categories.
State	It is the state that the customer locates to.	Nominal feature. 15 categories include 13 states, Kuala Lumpur, and Federal Territory of Labuan.
CompanyType	It denotes the company type of the customer.	Nominal feature. Corp = Corporate; RF = Recruitment Firm.
SalesPersonID	It shows the sales person who dealt with the customer.	Nominal feature.
Ordin3mb4	It indicates the total number of products a customer purchased in previous 3 months.	Numeric feature (integer).
Ordin6mb4	It indicates the total number of products a customer purchased in previous 6 months.	Numeric feature (integer).
Ordin12mb4	It indicates the total number of products a customer purchased in previous 12 months.	Numeric feature (integer).
Ord1mb4	It indicates how many products a customer purchased in the month T-1.	Numeric feature (integer).
Ord2mb4	It indicates how many products a customer purchased in the month T-2.	Numeric feature (integer).
Ord3mb4	It indicates how many products a customer purchased in the month T-3.	Numeric feature (integer).
Ord4mb4	It indicates how many products a customer purchased in the month T-4.	Numeric feature (integer).
Ord5mb4	It indicates how many products a customer purchased in the month T-5.	Numeric feature (integer).
Ord6mb4	It indicates how many products a customer purchased in the month T-6.	Numeric feature (integer).
Ord7mb4	It indicates how many products a customer purchased in the month T-7.	Numeric feature (integer).
Ord8mb4	It indicates how many products a customer purchased in the month T-8.	Numeric feature (integer).

Table 1.*(continued)*

Ord9mb4	It indicates how many products a customer purchased in the month T-9.	Numeric feature (integer).
Ord10mb4	It indicates how many products a customer purchased in the month T-10.	Numeric feature (integer).
Ord11mb4	It indicates how many products a customer purchased in the month T-11.	Numeric feature (integer).
Ord12mb4	It indicates how many products a customer purchased in the month T-12.	Numeric feature (integer).

Table 2. Feature subsets selected for the training sets

Training Set	Evaluator	Selected Features
1	CFS	(9 features) Industry, SalesPersonID, Ordin3mb4, Ordin12mb4, Ord12mb4, Ord11mb4, Ord6mb4, Ord5mb4, Ord2mb4
	SC	(11 features) State, Industry, SalesPersonID, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord12mb4, Ord11mb4, Ord6mb4, Ord5mb4, Ord2mb4
2	CFS	(6 features) Industry, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord5mb4, Ord2mb4
	SC	(10 features) State, Industry, SalesPersonID, Ordin6mb4, Ordin12mb4, Ord6mb4, Ord5mb4, Ord3mb4, Ord2mb4, Ord1mb4
3	CFS	(10 features) Industry, SalesPersonID, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord12mb4, Ord6mb4, Ord5mb4, Ord3mb4, Ord2mb4
	SC	(10 features) State, Industry, SalesPersonID, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord12mb4, Ord6mb4, Ord5mb4, Ord3mb4
4	CFS	(6 features) Industry, SalesPersonID, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord2mb4
	SC	(8 features) State, Industry, SalesPersonID, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord5mb4, Ord3mb4
5	CFS	(6 features) Industry, Ordin3mb4, Ordin6mb4, Ordin12mb4, Ord12mb4, Ord2mb4
	SC	(11 features) State, Industry, SalesPersonID, Ordin6mb4, Ordin12mb4, Ord12mb4, Ord9mb4, Ord6mb4, Ord5mb4, Ord3mb4, Ord2mb4

4.6 Performance Outcomes

The performance results of all of the models built with different training sets are shown in Tables 3.

Table 3. Performances of the ANN models and DT models based on different training sets, the number in the brackets denotes the training set number (FullFS = full feature set (without feature selection); CFS = correlation-based feature selection; SC = subset consistency; AUC = area under the ROC curve; LI = lift index; PTD = precision for the top decile)

Model	Method	AUC	LI	PTD
ANN(1)	FullFS	49.73%	0.5482	9.01%
	CFS	50.28%	0.5536	9.87%
	SC	51.49%	0.5636	9.61%
ANN(2)	FullFS	51.13%	0.5590	10.07%
	CFS	50.18%	0.5515	10.86%
	SC	50.97%	0.5595	10.20%
ANN(3)	FullFS	50.13%	0.5515	10.13%
	CFS	50.24%	0.5517	10.33%
	SC	49.56%	0.5459	10.72%
ANN(4)	FullFS	49.66%	0.5468	11.18%
	CFS	51.07%	0.5599	9.87%
	SC	50.86%	0.5573	10.00%
ANN(5)	FullFS	51.82%	0.5657	12.24%
	CFS	50.46%	0.5543	11.38%
	SC	52.43%	0.5715	11.84%
DT(1)	FullFS	51.91%	0.5734	14.93%
	CFS	53.24%	0.5825	11.25%
	SC	52.31%	0.5734	14.61%
DT(2)	FullFS	54.27%	0.5917	13.95%
	CFS	56.53%	0.6166	14.41%
	SC	52.77%	0.5792	10.53%
DT(3)	FullFS	52.23%	0.5745	15.39%
	CFS	53.39%	0.5822	11.38%
	SC	54.07%	0.5904	11.38%
DT(4)	FullFS	53.49%	0.5829	13.62%
	CFS	52.60%	0.5774	11.97%
	SC	54.35%	0.5932	10.79%
DT(5)	FullFS	54.92%	0.5979	13.88%
	CFS	56.07%	0.6111	19.08%
	SC	54.73%	0.5944	13.42%

Taking into account AUC, seven models with CFS produced better result than the corresponding models without feature selection (i.e. ANN(1), ANN(3), ANN(4), DT(1), DT(2), DT(3), and DT(5)). Three models with CFS produced worse result than the corresponding models without feature selection (i.e. ANN(2), ANN(5), and DT(4)). On the other hand, six models with SC produced better result than the corresponding models without feature selection (i.e. ANN(1), ANN(4), ANN(5), DT(1), DT(3), and DT(4)). Four models with SC produced worse result than the corresponding models without feature selection (i.e. ANN(2), ANN(3), DT(2), and DT(5)).

With lift index, seven models with CFS produced better result than the corresponding models without feature selection (i.e. ANN(1), ANN(3), ANN(4), DT(1), DT(2), DT(3), and DT(5)). Three models with CFS produced worse result than the corresponding models without feature selection (i.e. ANN(2), ANN(5), and DT(4)). On the other hand, six models with SC produced better result than the corresponding models without feature selection and one model with SC gave the same result (up to four decimal places) as the one by model without feature selection (i.e. ANN(1), ANN(2), ANN(4), ANN(5), DT(1), DT(3), and DT(4)). Three models with SC produced worse result than the corresponding models without feature selection (i.e. ANN(3), DT(2), and DT(5)).

As mentioned previously, for practical application, the authors need to pay more attention to the precision for the top decile. To consider the precision for the top decile, five models with CFS produced better result than the corresponding models without feature selection (i.e. ANN(1), ANN(2), ANN(3), DT(2), and DT(5)). Other five models with CFS produced worse result than the corresponding models without feature selection (i.e. ANN(4), ANN(5), DT(1), DT(3), and DT(4)). On the other hand, only three models with SC produced better result than the corresponding models without feature selection (i.e. ANN(1), ANN(2), and ANN(3)). Seven models with SC produced worse result than the corresponding models without feature selection (i.e. ANN(4), ANN(5), DT(1), DT(2), DT(3), DT(4), and DT(5)). It showed that SC might not be a good choice to serve as an additional feature selection method for DT model.

Nevertheless, all of the models generate better performance than the probability of a random choice (9.92%) except for the ANN(4) model with CFS and the three ANN(1) models. The best precision for the top decile, 19.08% was achieved by the DT(5) model with CFS.

5 Discussion and Conclusions

In this study, both feature selection methods have successfully reduced the number of features for the predictive models from eight to thirteen features. In particular, feature subset of the above-mentioned model which produced the best precision for the top decile is one of the smallest feature subset which comprises only six features. It showed that feature selection did not only reduce model complexity but also can improve model performance in the experiments of this study.

According to the results of the experiments, it was also clear that CFS outperformed SC most of the time. Moreover, the time for running CFS was not long and it was much less than the time for running SC. Thus, CFS is worthwhile to be investigated and applied in data mining research. Incidentally, DT model outperformed ANN model in these experiments. It might be because the DT model could handle these data better or both feature selection methods failed to select good feature subset for the ANN model.

Like most studies, this study has some limitations too. First of all, the authors used only experiments to show that using ANN model or DT model can solve the problem of this research and the performance of the predictive model can be improved by using feature selection method. Secondly, in this study, only filter models and an embedded model of feature selection were considered. In the future, the authors would like to investigate wrapper models of feature selection that usually involve lengthy processing time. As for the next phase of this research work, the authors will proceed to the stage of deployment of the predictive model in the job advertisement company. That is, to actually apply the predictive model in the participating organization and evaluate the outcome of the actual sales versus the predicted sales. The authors aim to prove that the predictive model is not only theoretically viable, but practically useful to the research participating company and the industry alike.

References

1. Kim, Y.S., Street, W.N.: An Intelligent System for Customer Targeting: A Data Mining Approach. Decision Support Systems 37, 215–228 (2004)
2. Bella, A., Ferri, C., Hernández-Orallo, J., Ramírez-Quintana, M.J.: Joint Cutoff Probabilistic Estimation Using Simulation: A Mailing Campaign Application. In: Yin, H., Tino, P., Corchado, E., Byrne, W., Yao, X. (eds.) IDEAL 2007. LNCS, vol. 4881, pp. 609–619. Springer, Heidelberg (2007)
3. Cox, L.A.: Data Mining and Causal Modeling of Customer Behaviors. Telecommunication Systems 21, 349–381 (2002)
4. Hung, S.Y., Yen, C., Wang, H.Y.: Applying Data Mining to Telecom Churn Management. Expert Systems with Applications 31, 515–524 (2006)
5. Au, W.H., Chan, K.C.C.: Mining Fuzzy Association Rules in A Bank-Account Database. IEEE Transactions on Fuzzy Systems 11, 238–248 (2003)
6. Kumar, D.A., Ravi, V.: Predicting Credit Card Customer Churn in Banks Using Data Mining. Int. J. Data Anal. Tech. Strateg. 1, 4–28 (2008)
7. Ling, C.X., Li, C.: Data Mining for Direct Marketing: Problems and Solutions. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp. 73–79 (1998)
8. Blum, A.L., Langley, P.: Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence 97, 245–271 (1997)
9. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. Artificial Intelligence 97, 273–324 (1997)
10. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 4–37 (2000)
11. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. J. Mach. Learn. Res. 3, 1157–1182 (2003)
12. Elashoff, J.D., Elashoff, R.M., Goldman, G.E.: On the Choice of Variables in Classification Problems with Dichotomous Variables. Biometrika 54, 668–670 (1967)
13. Toussaint, G.: Note on Optimal Selection of Independent Binary-Valued Features for Pattern Recognition (Corresp.). IEEE Transactions on Information Theory 17, 618 (1971)
14. Tirelli, T., Pessani, D.: Importance of Feature Selection in Decision-Tree and Artificial-Neural-Network Ecological Applications. Alburnus Alburnus Alborella: A Practical Example. Ecological Informatics (in press, corrected proof, Available online November 30, 2010)

15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
16. Holte, R.C.: Very Simple Classification Rules Perform Well on Most Commonly Used Datasets. Machine Learning 11, 63–91 (1993)
17. Liu, H., Motoda, H.: Computational Methods of Feature Selection. Chapman & Hall/CRC, Boca Raton (2007)
18. Chen, K., Liu, H.: Towards An Evolutionary Algorithm: Comparison of Two Feature Selection Algorithms. In: Proceedings of the Congress on Evolutionary Computation, vol. 2, pp. 1309–1313 (1999)
19. Liu, H., Setiono, R.: A Probabilistic Approach to Feature Selection - A Filter Solution. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 319–327 (1996)
20. Hall, M.: Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. In: Langley, P. (ed.) Proceedings of the Seventeenth International Conference on Machine Learning, pp. 359–366 (2000)
21. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0: Step-by-Step Data Mining Guide. CRISP-DM consortium (2000)
22. Ngai, E.W.T., Xiu, L., Chau, D.C.K.: Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification. Expert Systems with Applications 36, 2592–2602 (2009)
23. Provost, F., Fawcett, T.: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In: Heckerman, D., Mannila, H., Pregibon, D., Uthurusamy, R. (eds.) Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, pp. 43–48. AAAI Press, California (1997)
24. Huang, J., Ling, C.X.: Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering 17, 299–310 (2005)

Visualizing MDS Coordinates of Multiple Information in Color for Attribute-Based Two-Dimensional Space

M. Bakri, C. Haron, Siti Z.Z. Abidin, and Zamalia Mahmud

Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Malaysia
bakri891015@gmail.com, {zaleha,zamalia}@tmsk.uitm.edu.my

Abstract. Multidimensional scaling (MDS) is often used by researchers to provide a visual representation of the pattern of proximities (i.e., similarities or distances) among a set of objects normally via 2-D space map. Such representation often relates to survey questions involving subjects' multiple responses toward certain attributes in a study. However, too many subjects and attributes will produce massive output points (coordinates) which could dampen the visualization of coordinates in 2-D space map. Therefore, we propose a new way to visualize the MDS output presentation in 2-D space map by reclassifying the results according to attributes with different shapes and colours, and recalculate all the cases to view the similarity scaling according to height and distance ratio between all the output coordinates using Java programming. In this study, responses regarding preferences and reasons for choosing favorite colors were compared between subjects of different sample sizes. The purpose is to see the changes in the responses based on the new visualization of coordinates. The results had shown a marked improvement in the visual representation of the results based on different heights, shapes and colors.

Keywords: visualization, MDS coordinates, similarities, clustering, distance ratio algorithm, attributes, visual computing.

1 Introduction

In a survey study involving close-ended type of questions, respondents are normally allowed to provide only one answer to the multiple-choice type of questions. However, in some cases, respondents are also allowed to provide multiple answers to certain type of questions. With such data input, the challenge is to present similar information given by more than one respondent through certain visual representations for the purpose of identifying the relationship and pattern of similarities (or distances) between the responses. Multidimensional scaling (MDS) is one of the techniques that support the visualization of data in two dimensions. Its ability to present information in cluster-formed coordinates allows results to be interpreted easily according to the survey subjects and attributes. However, too many subjects and attributes will produce massive output points (coordinates) which could dampen the visualization

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 168–178, 2011.

results. So far, MDS has managed to produce results in 2-D space. However, the interpretation based on the display could be a daunting process due to the overcrowding and overlapping coordinates. Thus, several attempts were made to visualize the multidimensional data coordinates in several shapes [1-3] and colors [4] as well as in three dimensional (3-D) space [5]. However, with such visual improvement, we acknowledged that there are some limitations with regards to the viewer interface.

The limitations has led to the visual development of the MDS output coordinates by reclassifying the results according to color based attributes and recalculate all the cases (or survey subjects) to view similarity scaling according to height based on distance ratio [5] between all the output coordinates. A Java programming was written to read all the MDS output coordinates and visualize the output in 2-D and 3-D space maps with flexible views in producing colour-attribute coordinates for easy identification.

Section 2 discusses the survey results and the process of data visualization based on the MDS technique. Section 3 shall present the methodology and Section 4 illustrates the results and analysis before the concluding remarks in Section 5.

2 The Survey Results

In any survey research, data is normally presented descriptively and presented in tabular format and/or visualized in one-dimensional graphical space. However, if the collected data is of multiple response type, then visualization of data would require more clustering based on the similarities of the responses given by the subjects. This can be illustrated using Multidimensional Scaling (MDS) space map such that those attributes that are perceived to be very similar to each other are placed nearby, and those attributes that are perceived to be very different from each other are placed far away from each other on the map [6]. Thus, the results can be analyzed based on the visualization.

2.1 Visualization

Visualization is a graphical presentation of information with the purpose of providing the viewer with a qualitative understanding of information contents. It includes the process of transforming objects, concepts and numbers into picture or illustration [7]. Visual technologies have proven crucial for helping people to understand and analyze large amount of data. It helps by leveraging the human visual system [8]. Visual data analysis combines the efficiency of computers with the cognitive strength of human users [9]. The main challenge faced by information visualization is the need to represent abstract data and the relationship within the data [10]. Data visualization is a way to present and display information in a way that encourages appropriate interpretation, selection and association. It shifts the load from numerical reasoning to visual reasoning and exploits the capabilities of human eye to detect information from pictures or illustrations. Finding trends and relations in the data from a visual representation is much easier and far more time-saving compared to looking through text and numbers [7].

The two-dimensional (2-D) views are good for viewing details of a specific part and navigating or measuring distance precisely. Whereas, three-dimensional displays are good for gaining an overview of a 3-D space and understanding 3-D shape [11]. Other advantage of 3-D is its ability to show the relationships between variables [12]. It is possible to reduce the multidimensional data in a 2-D space as implemented by the multidimensional scaling (MDS).

2.2 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is a basic mapping technique for dimensionality which takes 2-D to illustrate the similarity relations among objects [13]. MDS is one of the methods for multivariate statistical analysis. It is commonly used in social science, psychology, and marketing. MDS focuses on finding the similarities or distances among objects. When all the measurements for dissimilarities among observed entities are transformed into a 2-D plane, users can observe the dissimilarity or dissimilarities through eye-browse. The information that lies beneath the data can be revealed and the representation of pattern proximities in two or three dimensions can help researchers to visually understand the structure of the data [14].

There are many different MDS techniques that can be used to analyze proximity data that include metric and nonmetric scaling, as well as deterministic and probabilistic MDS. In metric scaling, the dissimilarities between objects will be in numerical and distance. The objects will be mapped in metric space and the distance will be calculated [14]. On the other hand, nonmetric scaling uses the ranks of the data only [15]. Therefore, it can only be used when the dissimilarities between objects have meaning. The well known method to that uses nonmetric scaling is Kruskal [13]. For the deterministic MDS, each object is represented as a single point in multidimensional space [16], while the probabilistic MDS uses a probability distribution in multidimensional space [17].

In fact, it is beneficial to use MDS in marketing for determining the consumer attitudes and preferences to position the product [18]. According to Buja *et al.*[17], MDS is invented to analyze proximity data which arises in the areas of social sciences, archaeology and classification problems. In social sciences, proximity data takes the form of similarity ratings for pairs of stimuli such as tastes, colors and sounds. For archaeology, the similarity of two digging sites can be quantified based on the frequency of shared features in artefacts found in the sites, while in classification problems, the classification of large number of classes, pair wise misclassification rates produce confusion matrices that can be analyzed as similarity data. An example is confusion rates of phonemes in speech recognition.

3 Methodology

In order to illustrate the visualization of MDS output, an initial survey comprises of 50 subjects and 10 attributes was conducted to obtain the multiple dichotomy responses on the subjects' preferences and reasons for their choice of primary colours. The 2-D coordinates were transformed to produce case similarity in heights and reclassify the attributes in colours. The testing on the algorithm was performed on another set of 200 subjects in order to visualize further changes in the pattern of proximities. The new results have demonstrated a marked improvement in terms of its quality of visualization, thus able to help researchers to visualize and interpret the survey results in flexible views.

The processes of transformation involved recalculating the MDS 2-D coordinates for the purpose of adding colours to the visual output and represent the results in several different views for analysis. Using Java programming, the output is produced with flexible choice of information clustering. The program was written where it reads the MDS output coordinates before performing calculations on all the coordinates involving the distance between a point to all other points in the 2-D space.

There are three main steps in the implementation phase. Firstly, the input data (MDS coordinates) will be read and classified into two categories; cases and attributes. Then, colors will be applied to the appropriate class according to the attributes in the input data and finally all coordinates are presented by suitable colored objects according to the type of response. The results can also be viewed in 2-D but within a 3-D space map, the height is used to indicate the degree of similarities. The 2-D MDS points are presented in X and Y coordinates and the third Z-coordinate is added for the 3-D space. The Z coordinate is produced based on a distance ratio algorithm [5]. Figure 1 shows the flow of work implementation.

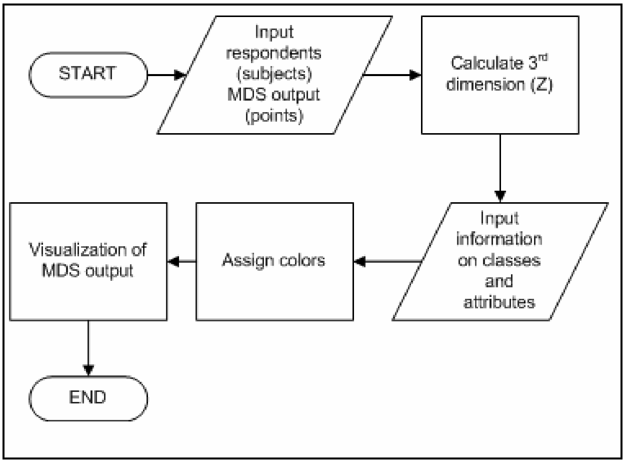

START
Input respondents (subjects) MDS output (points)
Calculate 3rd dimension (Z)
Input information on classes and attributes
Assign colors
Visualization of MDS output
END

Fig. 1. Work flow of the implementation

In order to make sure that every data has its own distance to every other data point in the 2-D space, a linear search algorithm is applied to find and calculate such distances. Throughout the process of calculating third dimension and assigning colors to the specified coordinates, it is important to include the maximum distance calculation so that the distance ratio is determined based on such maximum distance. Thus, with the value presented as ratio, the result is always between *0* and *1*. For the color enhancement to the output, the nearest attributes of selected class needs to be identified. It is performed by getting the minimum distance from a point to all attributes in a class. The flow of work for color enhancement is shown in Figure 2.

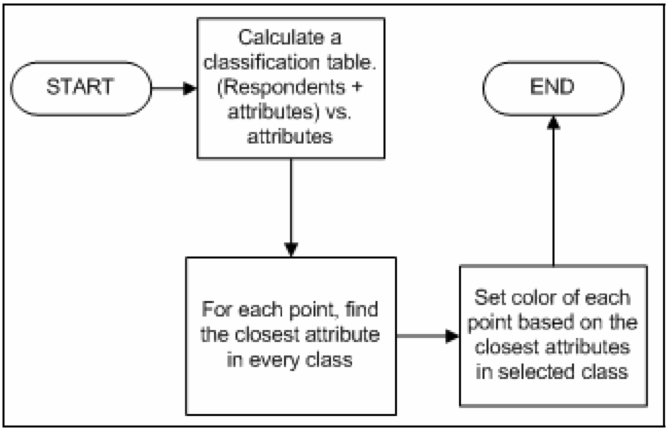

START
Calculate a classification table. (Respondents + attributes) vs. attributes
END
For each point, find the closest attribute in every class
Set color of each point based on the closest attributes in selected class

Fig. 2. The flow of work for color enhancement

The pseudo codes of color enhancement process are listed in Figure 3 below. It is important to identify the attribute values in the dataset. The algorithm for calculating distance of a particular point in relative to other points in space is still significant and part of the process in applying colors for the output. The algorithm implies that different colors are assigned to different attributes and all the points are calculated and colored according to the nearest attribute class.

```
for i<-0 to number of respondents + attributes
 attInClass = number of attributes in 1st class;
 Initialize min value;
 minIndex = 0;
 for j<- 0 to number of attributes
  --attInClass;
  dist = distance between this point to attribute j;
  if min > dist
   min = dist;
   minIndex = current index;
  if no more attribute in current class
   store minIndex for this class;
   if has more class
    attInClass = get number of attributes in next class;
   reset min and minIndex value
 store classification result for this point
```

Fig. 3. Pseudo codes for color enhancement

In the 3-D space map, the heights and proximities of the MDS coordinates indicate the degree of similarities among the subjects in the study. This programming has the flexibility to manipulate the output data and the ability to interact with the scene such as rotation and zooming. Furthermore, the visualization results in 3-D space can also be effectively visualized in two dimensional space for the height presentation.

4 Results and Analyses

The information visualization is performed on two different datasets. Both datasets are based on the same set of survey questions. The purpose of the survey is to analyze the respondents' feedback regarding their preferences and reasons for choosing their favorite colors. Respondents are allowed to give one or more possible answers for each question. The data is analyzed by comparing the visualization of subjects with two different sample sizes (50 and 200 subjects, respectively). The main contribution is in the way the visualization of results is viewed that include colors and flexible control. The classification of color coordinates for attributes include gender, favorite colors and reasons.

4.1 MDS Data with Small Sample (50 Subjects)

For 50 subjects, the visualization of MDS output is presented based on all cases (subjects that are represented by numbers) and attributes (gender, colors and reasons) as depicted in Figure 4. The output contains the proximity scaling on all the cases and attributes.

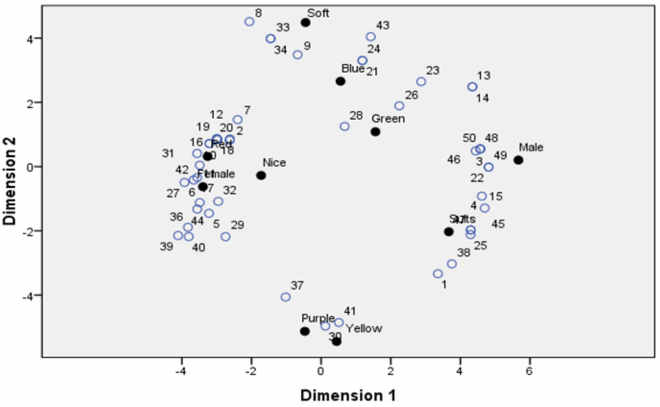
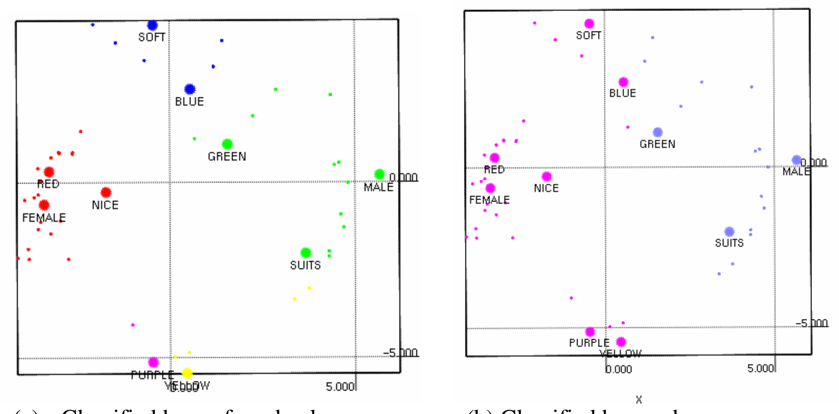


Fig. 4. MDS output for 50 respondents

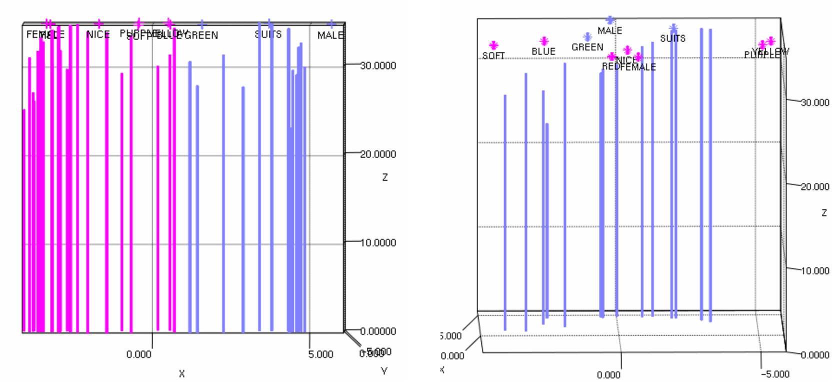
As the coordinates are read and executed by the implemented visualization tool, the output can be visualized in colored distribution based on the response attributes. Figure 5 shows the representation of output with colors for classification of the subjects' preferred colors and gender. When colors are applied to the output, the similarities in the proximity scaling among subjects can be determined. For example, in Figure 5(a), more female subjects respond similarly to red and nice where it is illustrated by the close clustered of red points among the subjects. On the contrary, more male subjects prefer green color although they do not give similar reasons for choosing the color. Hence, the cluster points are further apart.



(a) Classified by preferred colors (b) Classified by gender

Fig. 5. Proximity classified by preferred colors and gender

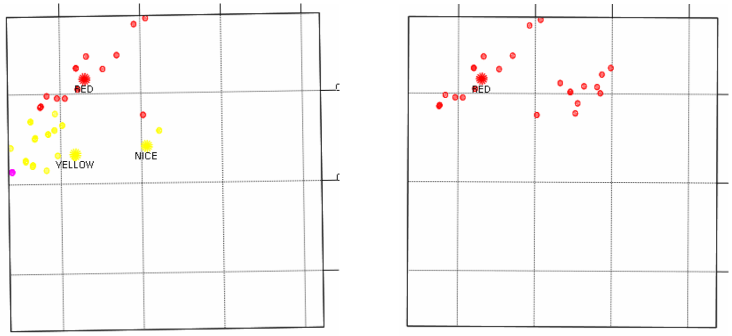
Another technique for analyzing the output is by calculating the distance of the points relative to all other points. The lesser distance indicates more similarities between the subjects. Then, a third dimension is produced to indicate the height of each point. Figure 6 shows the 3-D side view output on 2-D plane. The heights correspond to the points representing the subjects' gender. It is also possible to view only certain attribute such as male subjects (Figure 6(b)). The results show that the attribute *suits* is frequently stated but by less male subjects who had also stated green as their favorite color.



(a) Classified by gender (b) Results of male respondents

Fig. 6. Side view of 3-D output

In addition to visualizing the results according to attributes classification and 3-D side view, users are able to view the output according to four display regions: (1) upper left, (2) upper right, (3) lower left, and (4) lower right. Figure 7(a) shows the cluster view for red and yellow in region 1 and Figure 7(b) shows the cluster view of subjects for one or more attributes.



(a) Results for region 1

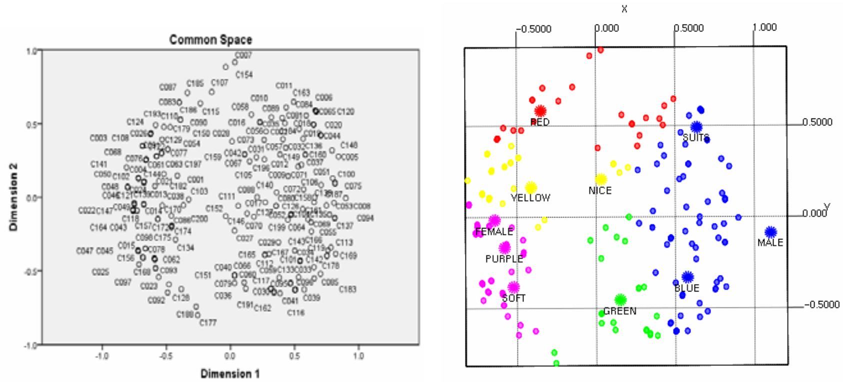
(b) Results for preferred *red* color

Fig. 7. Cluster view according to region and attribute

With all the flexibilities provided by the implemented visualization tool, it is important to check for its feasibility and correctness. Therefore, responses from another set of 200 subjects are tested.

4.2 MDS Data with 200 Subjects

The same set of survey questions is distributed to the 200 subjects. Figure 8(a) shows the output produced by the MDS proximity scaling.



(a) MDS output for 200 respondents

(b) Visualization by preferred colors

Fig. 8. Results of 200 subjects

With similar representation of results, it is difficult to analyze the similarities in the subjects' responses. However, the results are easier to view when colors are applied. Figure 8(b) illustrates the visualization of the multiple responses with respect to the preferred colors. The label for every respondent is optional and it is displayed as per user's request. When the colored representation attributes are displayed in 3-D side view, the density of the clustered bars based on the similarities of attributes can be differentiated. The results are depicted in Figure 9.

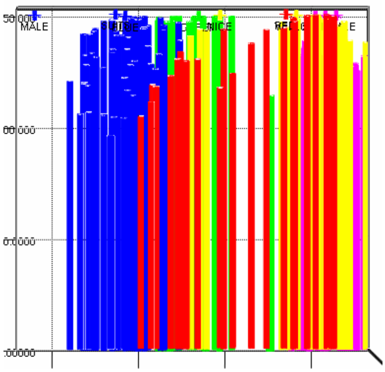


Fig. 9. 3-D side view (all points)

In order to check the usability of this work, a quick survey is conducted on 30 students and most of them are familiar with MDS proximity scaling. The output of 200 respondents as in Figure 8(a) and Figure 8(b) are shown to them together with the survey questions. Ninety percents of them agree that colors enhance the visualization and analyses. Furthermore, ten of them amaze with the changes since they feel that the normal MDS output is quite difficult to interpret in details.

5 Conclusion

In this paper, we have presented an attractive new method for visualization of multiple responses survey results based on different colors attributes via a flexible 2-D view mode. This transformation of output view is implemented using a tool which is written in Java programming language. The tool reads data that presents the X and Y coordinates of the MDS proximity scaling. The purpose of the tool is to enhance the 2-D visualization of survey results by providing users with the choice of viewing the output in colors which is categorized according to attributes of data. The tool has also demonstrated the possible transformation from the 2-D output into a 3-D representation by adding another dimension that portrays the degree of similarity based on heights. In this study, the 3-D side view is used to present the visualization in 2-D space. The enhancement of results using this visualization tool can be clearly seen when the output based on 50 and 200 subjects are compared. For usability test, a quick survey is carried out to ensure that the new visualization tool is able to help users to analyze multiple response survey results in a more practical way.

Acknowledgements

The authors wish to thank Universiti Teknologi MARA Malaysia for the financial support through the Excellence Fund. We also thank the anonymous reviewers for their valuable comments, which led to the improvement of this paper.

References

1. Paulovich, F.V., Oliveira, M.C.F.: The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization. In: XX Brazilian Symposium on Computer Graphics and Image Processing (SIBIGRAPI) 2007. IEEE Computer Society Press, Belo Horizonte (2007)
2. Noirhomme-Fraiture, M.: Visualization of Large Data Sets: The Zoom Star Solution. International Electronic Journal of Symbolic Data Analysis (2002)
3. Bentley, C.L., Ward, M.O.: Animating Multidimensional Scaling to Visualize N-Dimensional Data Sets. Proceedings of the IEEE Information Visualisation 96, 72–73 (1996)
4. Kamsiran, R.: Flexible Attribute Classification for MDS Points. Bachelor degree final year report, UiTM, Malaysia (2010)
5. Abidin, S.Z.Z., Hamid, N., Mahmud, Z.: Visualization of Multiple Response Data via 2-D and 3-D Colors Presentation. In: Proceedings 2010 International Conference in Information Retrieval and Knowledge Managemen. IEEE, Los Alamitos (2010)
6. Hair, J.F., Black, W.C., Babib, B.J., Anderson, R.E., Tatham, R.L.: Multivariate Data Analysis, 6th edn. Pearson, Prentice Hall (2006)
7. Kaidi, Z.: Data Visualization (2000), `http://www.cs.uic.edu/~kzhao/Papers/00_course_Data_visualization.pdf` (retrieved August 22, 2010)
8. Heer, J., Hellerstein, M.J.: Data Visualization and Social Data Analysis. In: Proceedings of the VLDB Endowment, pp. 1656–1657. VLDB Endowment, Lyon (2009)
9. Keim, D.A.: Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics, 1–8 (2002)
10. Chi, E.H.-h., Riedl, J., Shoop, E., Carlis, J.V., Retzel, E., Barry, P.: Flexible Information Visualization of Multivariate Data from Biological Sequence Similarity Searches. In: Proceedings of the 7th IEEE Visualization Conference, p. 133. IEEE Computer Society Press, California (1996)
11. Tory, M., Moller, T.: Human Factors in Visualization Research. IEEE Transactions on Visualization and Computer Graphics 10(1), 72–84 (2004)
12. Overbye, T.J., Weber, J.D.: Visualization of Power System Data. In: Proceedings of the IEEE Symposium on Information Visualization 2000, Hawaii, p. 7 (2000)
13. Tsumoto, S., Hirano, S.: Visualization of Rules Similarity using Multidimensional Scaling. In: 3rd IEEE International Conference on Data Mining, pp. 339–346. IEEE Computer Society, Florida (2003)
14. Steyvers, M.: Multidimensional Scaling. Encyclopedia of Cognitive Science (2002)
15. Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H.: XGvis: Interactive Data Visualization with Multidimensional Scaling (March 24, 2004), `http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-mds-jcgs.pdf` (retrieved September 26, 2010)

16. Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer, Heidelberg (2005)
17. Mackay, D.B.: Probabilistic Multidimensional Scaling: An Anisotropic Model for Distance Judgements. Journal of Mathematical Psychology 33(2), 187–205 (1989)
18. Chiesl, N.E.: The Stochastic Generation of a Multidimensional Scaling Technique Utilized in the Teaching of Marketing Management. In: Proceedings of the 11th Conference on Winter Simulation, vol. 1, pp. 117–124. IEEE Press, California (1979)

User Interface and Interaction Design Considerations for Collaborative Learning Using Augmented Reality Learning Object

T. Makina and Sazilah Salam

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
tarisa.makina@gmail.com, sazilah@utem.edu.my

Abstract. Most education is too often about teaching and not enough about learning. It is because students are forced to take whatever it is given to them without considering what they think about it, in other words, they passively take the given knowledge. This paper presents early investigation about interface and interaction design considerations for effective collaborative learning by taking account individual learning preferences and collaborative learning characteristics of engineering students. In our investigation, we follow Felder Silverman Learning Style Model and conducted a test measured using Index Learning Style. As a result, we discovered that engineering students tend to be active, sensory, visual, and sequential. Therefore, we implement augmented reality views to satisfy students' learning preferences toward content presentation (visual learner). It is also because augmented reality can give rich information toward real objects/environment. For collaborative characteristics, we studied past research on collaborative learning regarding its characteristics that affects learning effectiveness. Besides, our proposed design also considered the user interface principle which provides a guidance to effectively implement our consideration into an interface.

Keywords: user interface, learning style, collaborative learning.

1 Introduction

Most faculties rely on teaching their students about knowledge rather than guiding students to learn on their own. This creates a passive learning environment where students' knowledge is based on whatever knowledge given to them. Since learning is an active and constructive process, collaborative learning approach is suitable to be implemented in learning environment where the center of learning is on the students' exploration not simply the teacher's presentation [1]. By applying collaborative learning, students should exchange ideas with peers, analyze other opinions, and synthesize their understanding so that the knowledge is build from a combination of information they already had and information they get from communication with others. Furthermore, students' level of understanding are higher and information will

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 179–187, 2011.

retain longer when students work collaboratively compared to them who work individually [2].

There are several elements should be considered in designing collaborative learning application. First is diversity of students within a learning group which may lead to learning disturbance. There are many elements of diversities but the most important element is learning style [1]. Second is environment and tool that support students' exploration. This includes requirements supported for collaborative learning [3] and the way information should be delivered (interface design). Interface design principle is important as a guidance to implement our finding on engineering students' learning style and characteristic of collaborative learning at our university.

In current environment, which is in UTeM's engineering faculty, learning material is not supplemented with clear visual explanation and this, however, affects students' understanding. Based on engineering students' learning style, this study implements an augmented reality view which will satisfy visual students' needs. It also extends what students see in physical world and Kaufmann [4]found the usage of augmented reality in education is very powerful.

Our design will be implemented on mobile devices by seeing the fact that mobile devices are very popular among students and should be considered for a learning medium. Litchfield [5] found that the use of mobile devices can enhance learning experience and learning outcomes because of its ability to change the approach of learning and learning perceptions. This can give a new feeling to students, hence can increase their learning interests, learning experiences, and learning outcomes.

In the next section, we will talk about the learning style model called Felder Silverman Learning Style Model (FLSM) and collaborative learning characteristic and its elements. Next, we discussed about user interface and how the learning style and collaborative learning combined together to create a learning instrument. Finally, we summarize and conclude this paper.

2 Learning Styles

Learning style is a unique characteristic of individual skills and preferences which affect how a student receives, collects, and processes the learning material. There are many studies regarding learning style models. The reasons of using Felder Silverman are:

1. There is a free instrument for Felder Silverman learning style model called ILS [6],and some researchers have found the effectiveness of using ILS [7-11].
2. The objective of Index Learning Style questionnaire is to determine dominant learning styles of each student and intended for engineering education [12].
3. This model represents the characteristic of cognitive style and social interaction [13].
4. ILS data is informative and easy to be translated into specific design guidance [9].

FLSM is based on students' preferences on perception, retrieval, process, and understanding of information. FLSM characterized learners based on four dimensions: *active learners* prefer to learn within a group whereas *reflective learners* learn individually or in pair. *Sensing learners* easy to learn concrete material (facts and data) whereas *intuitive learners* easy to learn abstract material (principles and theory). *Visual learners* remember best what they see (picture, diagram, and demonstration) whereas *verbal learners* prefer discussions either spoken or written. *Sequential learners* learn in sequential manner, mastering the material as it is presented whereas *global learners* learn in large gaps. Table 1 shows students' learning preferences and its corresponding teaching styles.

Table 1. Learning Preferences and the Corresponding Teaching Styles

Learning Style		Teaching Style	
Processing		Student participation	
Active	Reflective	Active	Passive
- Let's try it out - Process information by physical activity - Learn by working with others	- Let's think it through - Process information introspectively - Learn by working alone or in pairs	- Providing discussion area - Reminding student to guess several possible questions - Emphasizes on problem-solving method	- Think before going ahead - Stop periodically to review what have been learning - Writing summaries - Emphasizes on fundamental understanding
Perception		Content	
Sensory	Intuitive	Concrete	Abstract
- Practical and observing - Prefer concrete: facts and data - Prefer repetition	- Imaginative and interpretive - Prefer abstract: theory and modeling - Prefer variation	- Example first and followed by the exposition - Hand-on work, such as practicing in the applying environment - Provide concrete information (facts, data, experiment's result)	- Exposition first and followed by the example - More concept and abstract (principles, theories)
Input		Presentation	
Visual	Auditory	Visual	Verbal
- Prefer picture and diagram - Show me how	- Prefer written and spoken explanation - Tell me how	- More picture, graphs, diagram - Animation demonstration - Color important concepts	- Text - Audio
Understanding		Perspective	
Sequential	Global	Sequential	Global
- Understand in continual and increment steps - Linear reasoning process - Convergent thinking and analysis	- Understand in large leaps - Tactit reasoning process - System thinking and synthesis	- Step by step to present material - Constrict links	- Give big picture of the course - Provide all the links

We conducted preliminary analysis toward engineering students in our faculty using Index Learning Style (ILS). The ILS was given to 21 random students and results from our preliminary analysis shown that engineering students tend to be more active (19:90%) rather than reflective (2:10%), sensory (18:86%) rather than intuitive (3:14%), visual (15:71%) rather than verbal (6:29%), and sequential (12:57%) rather than global (9:43%). Figure 1 shows the results of our study.

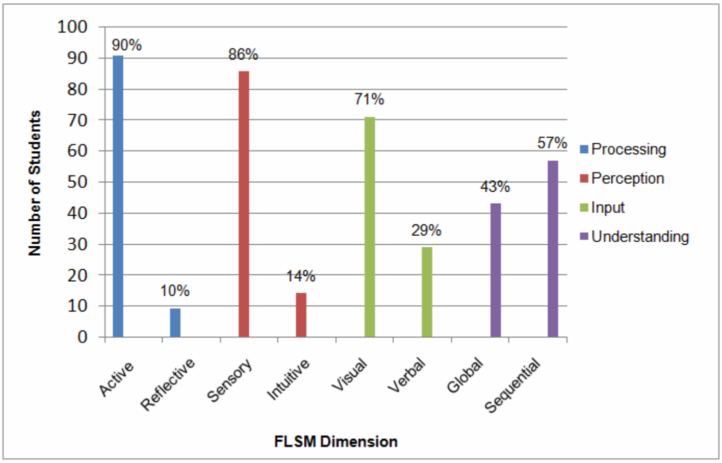


Fig. 1. Engineering Students' Learning Styles

3 Collaborative Learning

Collaborative learning is a situation in which students actively interact with each other to share knowledge and experiences in order to learn something together. Collaborative learning differs from traditional work group by the additional value within the group such as *interdependency* which is group success is based on everyone contribution; *accountability* which is every individual accountable to share his knowledge and group accountable to achieve its goal; and the development of *social skills* which every individual try to accept other's opinion, tolerate or resolve differences, make decision that agreed by all group members, and care what others doing. It is proved that student level of understanding is higher and information retain longer when students work on collaboratively compared to them who work individually [1-2, 14]. The characteristic of collaborative learning in a classroom based on Smith and MacGregor [1] and Soller et al. [15] are:(1)*communication* which is the way students share, exchange and grow the knowledge; (2)*social interdependence* which is students dependency toward each other to discover the knowledge and *participation* is important for every students; (3)students' *exploration* which is students' contribution to discover the knowledge not only depends on teacher's presentation; (4)*promote interaction* which can be done by *begin the activity with problems*; (5)*diversity between group* which means each group consist of different level of learning ability; (6)*assessment* which is not only based on group performance but also individual performance.

Table 2. Characteristics of Collaborative Learning

Collaborative Learning Characteristic
Communication
Social interdependence and participation
Students exploration
Promotive interaction
Diversity between group
Individual and group assessment

4 Interface Design

Interface design is a combination between system and users by providing interaction based on goals users trying to achieved, and tasks they should perform. User interface should concern about users, tasks, and context[16-17]. From preliminary analysis, we knew that the users characteristic are active, sensory, visual, and sequential. The context implemented in this study is collaborative learning environment whereas the tasks for students are to finish group and individual assessment. User interface was designed based on these three elements therefore the interface will implement augmented reality as learning instrument and functions to support collaborative learning. Table 3 shows the implementation of learning style consideration on user interface. Table 4 shows the implementation of collaborative learning characteristic in user interface.

Table 3. The Implementation of Learning Style in the Interface

Learning Style	User interface consideration	Explanation
Perception		
Sensory	Provide the overall pictures on generator and then determine each part and explain the functions.	Students tend to do observation and patient with details.
	Provide real usage of generator and motor.	Prefer facts.
	Provide animation to be played again and again.	Prefer repetition.
	Provide understandable marker for augmented reality	Dislike surprises and do not like complication.
Input		
Visual	Implement augmented reality.	Prefer picture and diagram because students remember best what they see
	Provides step by step explanation.	Show me how
Understanding		
Sequential	Provide step by step explanation.	Understand in sequential manner
Processing		
Active	Provide assessment to do.	Let's try it out (Do it)
	Support collaborative learning activity	Process information by physical activity and by working with others.

Table 4. The Implementation of Collaborative Characteristic in the Interface

Collaborative Characteristics	User interface consideration	Explanation
Communication Social interdependence Students' exploration Begin with problems	Provide group assignment	Group assignment allows students to take participation. It also acts as promotive interaction hence communication is irresistible.
Diversity between group Participation Promotive Interaction Performance analysis and group processing	Group and individual assessment	Group and individual assessment are different. Every student in a group will not get the same result. The result is not only from group performance but also individual understanding towards the assessment.

To cater sensory learner, interface provide "the big picture" to be observed. Animation is implemented to explain how things working which can be play over and over and brings advantage to them who like repetition. To cater visual learner, we implement augmented reality learning object. AR object allows students to make interaction with it in order to see the object clearly. To cater sequential learner, the interface provide sequential process in order to give clear explanation. To cater active learner, the interface provide assessments for individual and group. The interface also supports effective collaborative learning activity by providing centralized assessment. Through centralized assessment, students may submit their answers faster and lecturers can assess their assessment directly.

5 Augmented Reality Learning Object

Augmented reality is a term in which virtual object is imposed to real world so a person will see a virtual world as well as a real world. This characteristic brings advantages on learning to support visualization and collaborative activity. The way AR enables users to see both added information and real world brings advantages in collaborative learning environment. Students aware of others and communication happens without disruption [4, 18-19]. AR does able to added additional information into the real world. This gives advantage in learning because not all learning object can be display properly for example due to size (molecular object, big machine).

This study implemented augmented reality learning object to enhance students' visualization towards the learned material. The AR brings benefits not only for visual learner but also for active, sequential, and sensory learners. When the virtual object is displayed, sensory learner can explore and observe the objects' components. Sequential learners will also get the benefits because the virtual object can display a process of how something is done in a sequential manner in a form of animation and can be played over and over again.

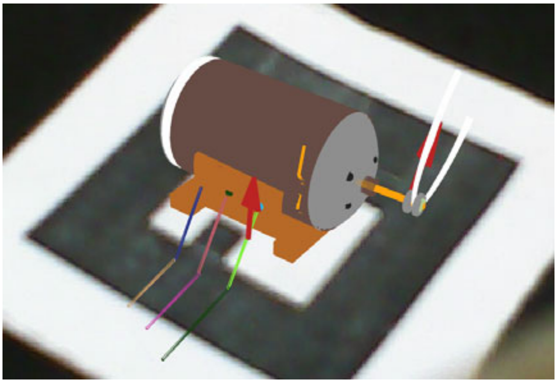


Fig. 2. The Implementation of Augmented Reality

6 Pilot Test

We conducted a pilot test to 24 engineering students. Students were asked to create a group of 3 students and they were assigned random questions. Then students had to separate from their group to meet students from other group who has the same questions. In the new group, students discuss about the given questions and create an artifact/document. After finished on this group, students were asked to back to their original group and presented the new knowledge to other group's member and created an artifact/document. Figure 3 shows group creation during pilot test. The artifact

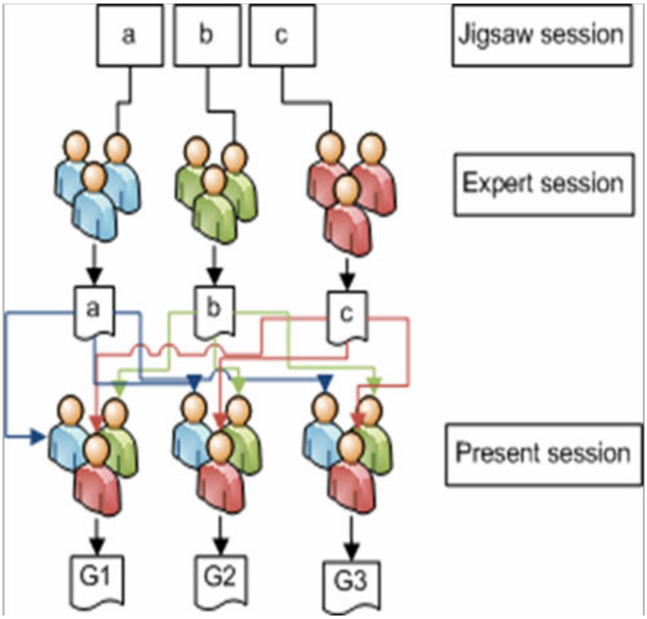


Fig. 3. Group Creation Based on Jigsaw Technique

created for each session were used to analyzed students' performance and satisfaction questionnaire were distributed to determine students' satisfaction towards the interface.

From this pilot test, the result shows that students who used the interface had higher performance compared to them who used traditional collaborative learning in expert session and present session. This was concluded based on the comparison on artifact/documents from expert session and present session between students who used the interface and them who used traditional collaborative learning. Students were satisfied with the interface in term of collaborative learning functions that it helped them to created and reused notes taken from other sessions. But there were negative comments on the interface regarding the redundant text and server connection. This was because the interface was not fully completed. For augmented reality display, students found it new and exciting thing but negative comments were received such as virtual object placement is not on its default view that students need to move and rotate the object to get the front view (correct view), it takes time to load the object because the objects contains too many meshes and object, unstable image tracking that sometimes virtual object was displayed and sometime was not.

7 Conclusion and Future Work

This paper presents early investigation about interface and interaction design considerations for effective collaborative learning by taking account individual learning preferences and collaborative learning characteristics of engineering students. In our investigation, we follow Felder Silverman Learning Style Model and we discovered that engineering students tend to be active, sensory, visual, and sequential hence in the interface, we implement centralized assessment for active learner and augmented reality views to cater visual learner. For collaborative characteristics, we studied past research on collaborative learning regarding its characteristics that affects learning effectiveness. Our next work is to assess the usability of interface on both collaborative functions and augmented reality view and conduct another test to assess students' performance after using the interface.

Acknowledgments. The research reported here is supported by Universiti Teknikal Malaysia Melaka under PJP No. PJP/2010/FTMK (13E) S787.

References

1. Goodsell, A., Maher, M., Tinto, V., Smith, B.L., MacGregor, J.T.: 'What is collaborative learning': 'Collaborative learning: A sourcebook for higher education'. (National Center on Postsecondary Teaching, Learning, and Assessment (NCTLA), pp. 10–29 (1992)
2. Gokhale, A.A.: Collaborative learning enhances critical thinking. Journal of Technology Education 7(1), 22–30 (1995)
3. Yu, L.: Principles for collaborative learning platform design. In: Proc. 1st International Conference on Information Science and Engineering (ICISE 2009), Nanjing, December 26-28 (2009)
4. Kaufmann, H.: Collaborative augmented reality in education. In: Proceeding of Imagina 2003 Conference, Imagina 2003 (2003)

5. Litchfield, A., Dyson, L.E., Lawrence, E., Zmijewska, A.: Directions for m-learning research to Enhance Active Learning. In: Proceeding of ICT: Providing Choices for Learners and Learning (Ascilite 2007), pp. 587–596 (2007)
6. http://www.engr.ncsu.edu/learningstyles/ilsweb.html (accessed May 21, 2010)
7. Graf, S., Viola, S.R., Kinshuk, Leo, T.: Representative characteristics of felder silverman learning styles: An empirical model. In: Book Representative Characteristics of Felder Silverman Learning Styles: An Empirical Model, pp. 235–242 (2006)
8. Kinshuk, T.L.: Application of learning styles adaptivity in mobile learning environments (2004)
9. Kirkham, P., Farkas, D.K., Lidstrom, M.E.: Learning styles data and designing multimedia for engineers. In: Proc. International Professional Communication Conference, 2006 Saratoga Springs, NY, October 23-25 (2006)
10. Kolomos, A., Holgaard, J.E.: Learning style of science and engineering students in problem based and project based education. In: Book Learning Style of Science and Engineering Students in Problem Based and Project Based Education. Sense Publishers (2008)
11. Viola, S.R., Graf, S., Kinshuk, Leo, T.: Analysis of felder-silverman index of learning styles by a data-driven statistical approach. In: Proc. Multimedia, ISM 2006, San Diego, CA (2006)
12. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. Engineering Education v78(7), 10 (1988)
13. Tobar, C.M., Luís, R.: Using learning styles in student modeling (2004)
14. Coleman, M.R., Gallagher, J.J., Nelson, S.M.: Cooperative learning and gifted students: Report on five case studies. In: Book Cooperative Learning and Gifted Students: Report on Five Case Studies. The University of North caroline (1993)
15. Soller, A., Goodman, B., Linton, F., Gaimari, R.: Promoting Effective Peer Interaction in an Intelligent Collaborative Learning System. In: Goettl, B.P., Halff, H.M., Redfield, C.L., Shute, V.J. (eds.) ITS 1998. LNCS, vol. 1452, pp. 186–195. Springer, Heidelberg (1998)
16. Parsons, D., Ryu, H., Cranshaw, M.: A design requirements framework for mobile learning environments. Journal of Computers 2(4), 1–8 (2007)
17. Weinschenk, S., Jamar, P., Yeo, S.C.: GUI design essentials. Wiley Computer Pub. (1997) (illustrated edn.)
18. Henrysson, A., Billinghurst, M., Ollila, M.: Face to face collaborative ar on mobile phones. In: Book Face to Face Collaborative ar on Mobile Phones, pp. 80–89 (2005)
19. Wagner, D., Pintaric, T., Schmalstieg, D.: The invisible train: A collaborative handheld augmented reality demonstrator. In: International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH 2004 Emerging Technologies, p. 12 (2004)

Augmented Reality Remedial Paradigm for Negative Numbers: AVCTP

Elango Periasamy[1] and Halimah Badioze Zaman[2]

Fakulty of Technology and Information Sciences, National University Malaysia
[1]surensutha@yahoo.com, [2]hbz@ftsm.ukm.my

Abstract. The aim of this study was about integrating augmented reality as visualization interface tool in mathematical remedial works. This study was based on the incorrect thinking process in solving negative numbers subtraction operation involving two integers then created an algorithm for visualization of correct thinking process in solving it. The respondent of this study were 124 students aged 14 years old from two secondary schools in Malaysia. The findings were an algorithm for visualization of correct thinking process (AVCTP) and its flowchart which was created in the process of assisting software engineering process of AR visualization based remedial works in subject domain.

Keywords: Visual Informatics, Augmented Reality, Visualization, Negative Numbers, Remedial.

1 Introduction

Augmented Reality (AR) technology is not new but its potential in education is just beginning to be explored [1]. Research in this area of AR is still maturing with few papers available on AR in education [1] and [2]. The efforts in developing an AR system is for the improvement of spatial abilities [2] and [3] then maximization of transfer of learning but now the need to apply them to real educational work have arrived [2]. In such, there is no simple one answer to guide specific practice in learning and teachers must provide a wide variety of methods through their diverse repertoire of class room practices in their lesson planning, the topic presented, the instructional experiences and activities incorporated in the learning session and their responses to children's questions [4]. Meanwhile, [5] say that a central function of the mind is to process the information, sort them in a meaningful way is determined by the rules and principles employed, thus learning is then perceived as appropriating these rules and principles, and being able to apply or process information according to these rules. Even though, such strategy is that part of the problem solving process that provides direction the problem solver should take in finding the answer but strategies are not as a problem specific as are algorithm and strategies are often used in combinations [6]. In such, [5] say that by analysing the way the experts think and by teaching students these expert ways of thinking, cognitivist hope to instruct students in order to emulate expert thinking and develop the students' expertise is a particular

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 188–198, 2011.

domain of knowledge. Meanwhile, in teaching mathematics, a rather different approach was to concentrate on studying thinking process involved in “doing mathematics” and researchers sought to characterized problem-solving heuristics, which included pattern recognition techniques, visual, spatial and logical reasoning by analogy with related or simpler problem [7].

In conjunction, AR as a visualization tool which can convey either static virtual objects or dynamic animation at the same time and it is not a panacea to all conceptual representation but instructional designers should be cautious in determining how to integrate AR into a curriculum properly [8]. With that, the aim of this study was about integrating AR as visualization interface tool in mathematical remedial works. In order to do so, this study investigated the incorrect thinking process (ITP) in solving negative numbers subtraction operation involving two integers [9] based on the finding in [10], then created an AVCTP in solving negative numbers subtraction operation involving two integers. Thus this paper is about the flowchart created for AVCTP which will be implemented in software engineering process of AR visualization based remedial works in subject domain.

2 Related Works

According to [11], their findings suggest that adults’ representations of operation with negative numbers are not as well established as their representations of operations with positive numbers. Furthermore, in operation involving negative numbers, [12] says that some students assume many mathematical things to be universally true and because of this they are at times, amazed to realize their assumptions have been false [12], [9] and [13]. For example, some students are not aware that the commutative property for addition operates in sets other than the counting number. A series of questions or problems like $^{-}3 + {}^{+}7 =$ and $^{+}7 + {}^{-}3 =$ could help lead to the appropriate conclusions and can be amplified with problems involving subtraction where ‘*commutativity*’ does not generally hold, sometimes that same students assume to be true ($^{-}5 - {}^{+}8 =$ and $^{+}8 - {}^{-}5 =$) [12] and how students hold to a incorrect thinking process when solving such sentence questions for more information can be found in [10]. Moreover, according to [14], qualitative analysis in their study also showed that the difficulties with having to make explicit the negative numbers involved in the problems could be overcome when children marked positive and negative numbers differently; when negative numbers were differentiated from the operation of subtraction; and when children correctly interpreted the results obtained from operating on the explicit representations they had generated. Furthermore, they added that the explicit representation could be either in writing or by use of manipulative material chosen amongst those made available (coloured cards, marbles, rulers or sticks). However, Amstrong used a *PowerPoint* presentation involving building sandcastles and digging holes was found to illustrate a direct teaching as in Table 1 and through formative assessment during the lesson, it became apparent that lower ability students found the model easy to understand where the weakest student in the class, found the lesson particularly accessible and was highly enthused at the pictures and explanations being used as examples when that weakest student typically struggles in maths however higher ability students became confused [13].

Table 1. Calculation Analogy

-1 + 1	A hole plus a sandcastle gives a level surface, since the sand from the sandcastle fills the hole, so answer = 0
3 + -3	3 sandcastles plus 3 holes gives a level surface, so answer = 0
3 – 5	3 sandcastles take away 5 sandcastles, which is the same as flattening 3 sandcastles, then taking away 2 more sandcastles, thus making 2 holes, so answer = - 2
4 + -2	4 sandcastles plus 2 holes. 2 of the sandcastles fit in 2 of the holes, leaving 2 sandcastles remaining, so answer = 2
3 + -5	3 sandcastles plus 5 holes. The 3 sandcastles fill 3 of the holes, leaving 2 holes remaining, so answer = - 2
-2 + -3	2 holes plus another 3 holes. This gives 5 holes in total, so answer = - 5
-7 – - 4	7 holes take away 4 holes. Taking away a hole means filling in the hole with the sand of one sandcastle. Filling in 4 holes leaves 3 holes, so answer = - 3
-2 - - 3	2 holes take away 3 holes. Having taken away (or filled in) 2 holes, we have enough sand to make one sandcastle, so answer = 1

Although different strategies were used by various researchers in helping remedial students gain the knowledge of solving negative numbers subtraction operation, nevertheless, we have been given absurd rules to apply to this weird concept such as: *a negative number multiplied by a negative number equals a positive number,* and questioned that how can it be that a negative number, which by the definition mathematicians have given us, is less than zero, when multiplied by another number that is less than zero, become a positive number? It has to be pure, unadulterated nonsense and it is clear that real objects manipulation for subtraction operation of negative numbers is an illusion as negative numbers are imaginary numbers claimed by [15]. Such phenomena is explain by Naylor as a situation whereby in many parts of the world, students learn a subtraction algorithm different from our own and this algorithm makes a great puzzle for students [16].

Moreover, Brumbaugh and Rock claimed that it is important for students to determine what things are as well as what they are not, if we are to help them avoid arising at incorrect assumptions, conclusions, thought processes and generalization and they suggested that assistance is provided to the discovery process through a carefully developed set of problems that guide the student to appropriate responses [12]. Nevertheless, according to [17], teaching is a complex endeavour that requires teachers to meld knowledge about the nature of learners, pedagogical strategies and discipline content. Thus, we would like to introduce a new algorithm in the process of learning negative numbers subtraction operation involving two integers for remedial activity which were based on the research by [9] and [10] which produced the meld knowledge about the nature of content and incorrect thinking process on subject domain.

3 Method

This study is a continuity research conducted by [9] and [10]. The demographic information of that research was 124 respondents aged 14 years old and among them were 53 boys and 71 girls. The number of respondent achieved a grade A is 26 (20.97%), grade B 58 (46.77%) and grade C 40(32.26%) for their Primary School Evaluation Examination (UPSR) in mathematics subject. The focus of this study was to create an AVCTP in guiding negative numbers subtraction operation involving two integers and its flowchart, so that an appropriate technology can be integrated in pursue to develop visualization of correct thinking process based remedial works. In such, this research was divided into four stages as follows:

i. First stage: The first stage was to identify the incorrect solution produced by respondent for each items tested is found in [10].
ii. Second stage: The second stage was to predict the ITP based on the incorrect solution from the first stage in [9].
iii. Third stage: The third stage was a scaffolding process to analyzing and synthesizing the ITP predicted from second stage with CTP by students and teachers for each item with respect to its frequencies. Then, to identify an AVCTP to help remedial students emulate CTP and develop their expertise in negative numbers subtraction operation of two integers.
iv. Fourth stage: The fourth stage was to create a specific flowchart for negative numbers subtraction operation involving two integers based on the AVCTP from the third stage.

This paper is to share the findings related to the third and fourth stage of the research.

4 Findings

The first part of this study finding was a nine step AVCTP for negative numbers subtraction operation involving two integers as which is as in Table 2. There are four types of different sentence questions for negative numbers subtraction operation involving two integers which are categorized as type A, B, C and D and illustrated in Table 3. Type A refers to negative numbers subtraction operation involving two positive integers, Type B refers to negative numbers subtraction operation involving negative with positive integers, Type C refers to negative numbers subtraction operation involving positive with negative integers and Type D refers to negative numbers subtraction operation involving two negative integers. Table 3 also shows the types of question and its respective steps needed to solve it correctly with respect to AVCTP in Table 1. The second part this study finding was a flowchart for AVCTP for negative numbers subtraction operation involving two integers as illustrated in Fig. 1 (Appendix A).

Table 2. Algorithm Visualization of Correct Thinking Process (AVCTP)

Step	AVCTP
1	Identify any continues symbols and solve it, such as (– , – = +) or (–, + = –) in between two numbers, then rewrite the sentence question.
2	Identify and draw circle onto the negative number.
3	Identify and draw square onto the positive number.
4	Make a negative/ positive number group table as shown. - | +
5	Move magnitude value of negative numbers into the negative group.
6	Move magnitude value of positive numbers into the positive group.
7	Sum all number in each group if more than one numbers in each group.
8	Move the smaller number into the bigger number group. Then, subtract the smaller number from the bigger number.
9	Write the answer and put the positive/negative symbol with reference to the group where the answer is.

Table 3. Algorithm Steps for Question Type

Subtraction Operation Involving	Type	a, b>0	Algorithm Steps								
Two Positive Integers	A	a – b	🟦	2	3	4	5	6	🟦	8	9
Negative with Positive Integers	B	-a – b	🟦	2	🟦	4	5	🟦	7	🟦	9
Positive with Negative Integers	C	a – (-b)	1	🟦	3	4	🟦	6	7	🟦	9
Two Negative Integers	D	-a – (-b)	1	2	3	4	5	6	🟦	8	9

5 Discussion

The AVCTP was created to be used consciously and effortlessly by remedial students in solving negative numbers subtraction operation involving two integers. Lee and Robling went through papers on algorithm visualization (AV) published in the last couple of years, they observed a trend suggesting that the academic discussion on AV has been changing gears and the focus used to be on whether AV has any pedagogical value for learners learning about algorithms but now the focus appears to center on how educators and researchers can improve the pedagogical efficacy of AV, and congruently assess its pedagogical value [18]. Then the most challenging aspect is to bring the AVCTP into the mind of remedial students. In such, an AR visualizing technique with animation will be integrated in the pursuance of developing a technology based remedial works in subject domain. Furthermore, researchers believe that in AR, despite the exciting possibilities of the new media, educational content creation for an interactive system is at least as difficult as authoring good textbooks and will require a substantial amount of time and work [19] and [2]. So, the AR as a communication tool in transferring CTP into the mind of remedial students consciously depends on how it will provoke them to adapt it as their own thinking process. Furthermore, [20] study shows that computer science students regarded software animation demonstrations (SADs) as tools that initiate learning by providing a specific range of learning paths, with explicit goals and expectations, by means of a persuasive and intriguing presentation and after two months of usage, the majority students claimed that they would recommend SADs to their friends and characterized them as a very efficient way of learning compared to other means. Thus, they concluded that SADs once more emerged as an attractive learning medium due to their authenticity, their multimedia affordances and the feeling of personal contact that they engender. Meanwhile, [21] reviewed the task characteristics that suit SADs as

i. when the task involves continuous changes not easily inferred by the user
ii. when it is based on direct manipulation interface
iii. when it is clearly segmented into procedural steps
iv. when it will be practiced immediately after watching SADs

Those characteristics were seen in the process of respondent solving negative numbers subtraction operation involving two integers. However, learners have challenged SADs capacity to present complex skills that interweave with deep domain knowledge and claimed that experts would have difficulty in overcoming the pacing deficiencies and the redundant information presented, as well as their exploration deficiencies [22]. Thus, the focus of this study was to animate and demonstrate the CTP for remedial students in solving negative numbers subtraction operation involving two integers via AR technique. In conjunction, an AVCTP and its flowchart was created to help remedial students aware of the mental analysis and synthesis process needed when they are confronted with negative numbers subtraction operation involving two integers and help guide the software engineering in process of developing AVCTP system. As the ultimate goal of educational algorithm visualization systems is to aid students to learn more effectively and therefore, the success of a particular system can only be proved by evaluation even though many systems are available, but there have been fewer evaluations of these systems than one might have expected [23]. Moreover, history has shown us that as new technologies evolve, there is a need to carry out user studies as early as possible, to identify and address usability and usefulness issues [24]. Furthermore, [18] suggest that AV be analyzed by its symbol system (composed of many subsystems such as texts, graphics, sounds and animations), its interactivity (functions that require user input and engagement), and its didactic structure (system design and organization based on pedagogical considerations). So, usability test need to be carried out for AVCTP via AR interaction system.

6 Conclusion and Future Works

The rapid advancement and accessibility of technology have opened up literally worlds of possibilities for mathematics education [25]. Technology will continue to improve further and nations have to grow with these advancement and excellence in teaching and learning cannot be achieved through technology alone [26] and the needs to properly addressing the integration aspects of technology into teaching and learning process have arrived. Furthermore, with this increased use comes the challenge to provide young children with developmentally appropriate programs that meet their unique needs [27] and as technology integration in teaching and learning continues to evolve, giving new interaction tools means bringing new teaching and learning experiences that would diminish without that technology, is a challenge for researcher. Thus, this study was intended to help remedial students needing assistance in the area of negative numbers subtraction operation involving two integers, moreover, it is a part of a process of creating an AR interaction as a remedial paradigm for subject domain. Nevertheless, this AVCTP and its flowchart can be extended or modified for solving negative numbers addition and subtraction operation involving two or more integers and others such as decimal, fraction, numbers and algebra. The next challenging aspect will be identifying how to scaffolding the material so as to allow remedial students adapt that AVCTP effectively then effortlessly via AR remedial paradigm.

References

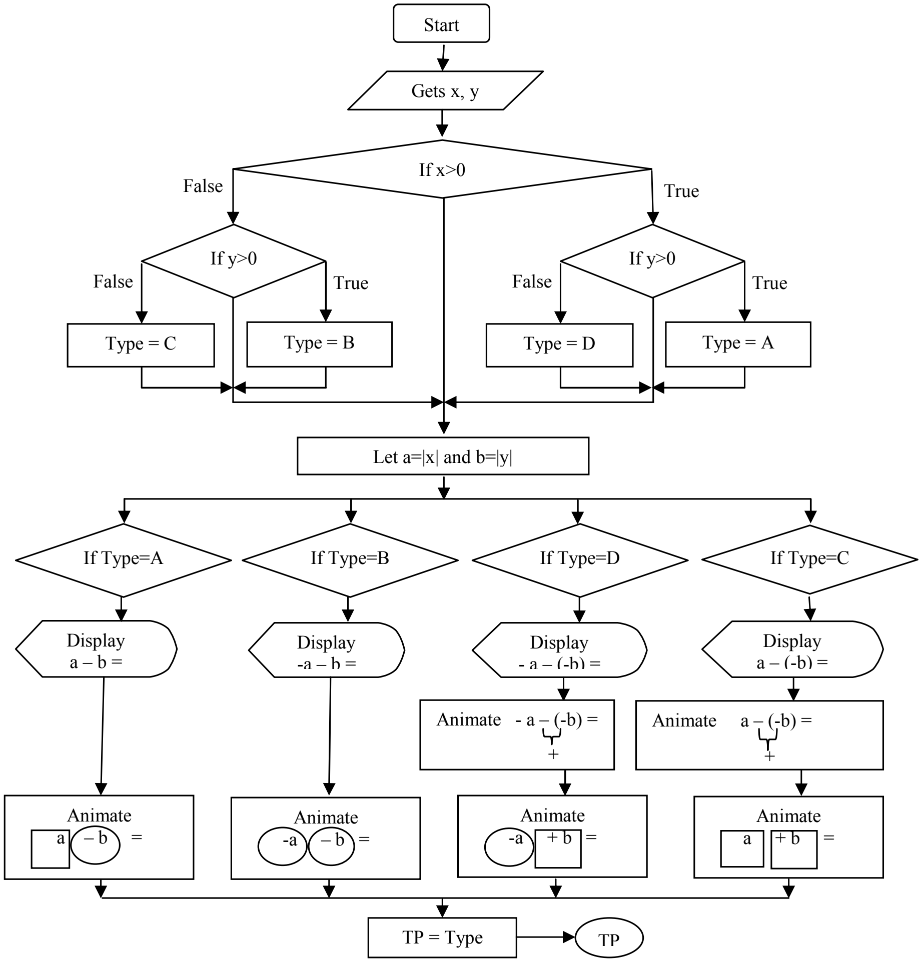
1. Billinghurst, M.: Augmented Reality in Education, New Horizons for Learning (2002), `http://www.newhorizons.org/strategies/technology/billinghurst.htm` (January 20, 2009)
2. Kaufmann, H., Schmalstieg, D.: Mathematics and geometry education with collaborative augmented reality. In: ACM SIGGRAPH 2002 Conference Abstracts and Applications, July 21-26. SIGGRAPH 2002, San Antonio, Texas. ACM, New York (2002)
3. Dünser, A., Steinbügl, K., Kaufmann, H., Glück, J.: Virtual and augmented reality as spatial ability training tools. In: Proceedings of the 7th ACM SIGCHI New Zealand Chapter's International Conference on Computer-Human Interaction: Design Centered HCI, CHINZ 2006, Christchurch, New Zealand, July 06 - 07, pp. 125–132. ACM, New York (2006)
4. Carnellor, Y.: Encouraging Mathematical Success for Children with Learning Difficulties. Social Sciences Press, Australia (2004)
5. Chen, V., Hung, D.: Learning Theories and IT in Instruction. In: Chee, T.S., Wong, A.F.L. (eds.) Teaching and Learning with Technology: Theory and Practice, pp. 82–100. Pearson Prentice Hall, Singapore (2003)
6. Krulik, S., Rudnick, A.J.: The New Sourcebook for Teaching Reasoning and Problem Solving in Junior and Senior High School. Allyn and Bacon, United States of America (1996)
7. Shteingold, N.: Young children thinking about negative numbers. In: Diss, D. (ed.) Dissertations & Theses: Full Text [Database on-line]. Rutgers The State University of New Jersey, New Brunswick (2008), `http://www.proquest.com.newdc.oum.edu.my`
8. Chen, Y.: A study of comparing the use of augmented reality and physical models in chemistry education. In: Proceedings of the 2006 ACM international Conference on Virtual Reality Continuum and Its Applications, VRCIA 2006, Hong Kong, China, pp. 369–372. ACM, New York (2006)
9. Periasamy, E., Zaman, H.B.: Predict Incorrect Thinking Process: Negative Numbers Subtraction Operation Second Category. In: International Conference on Advanced Science, Engineering and Information Technology, International Science Conference 2011, pp. 145–149 (2011)

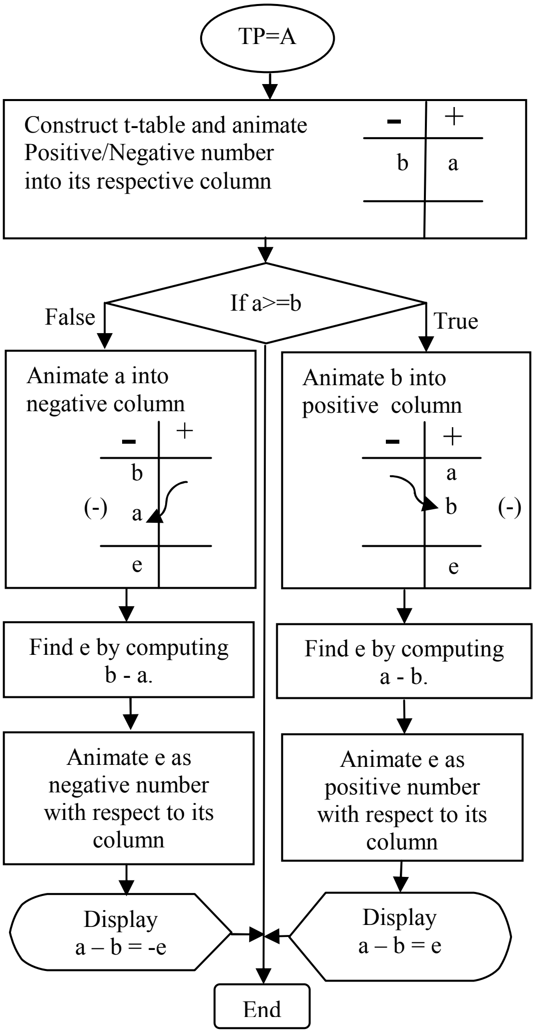
10. Periasamy, E., Badioze Zaman, H.: Augmented Reality as a Remedial Paradigm for Negative Numbers: Content Aspect. In: Badioze Zaman, H., Robinson, P., Petrou, M., Olivier, P., Schröder, H., Shih, T.K. (eds.) IVIC 2009. LNCS, vol. 5857, pp. 371–381. Springer, Heidelberg (2009)
11. Prather, R.W., Alibali, M.W.: Understanding of Principles of Arithmetic with Positive and Negative Numbers (2004), `http://www.cogsci.northwestern.edu/cogsci2004/ma/ma297.pdf` (September 4, 2008)
12. Brumbaugh, K., Rock, D.: Teaching secondary Mathematics. 3rd edn. Lawrence Erlbaum Assosiates, New Jersey (2006)
13. Armstrong, E.: Directed Numbers. Mathematics Teaching, ProQuest Education Journals (217), 3–5 (2010)
14. Borba, R., Nunes, T.: Teaching Young Children to Make Explicit their Understanding of Negative Measures and Negative Relations (2001), `http://www.bsrlm.org.uk/IPs/ip21-3/BSRLM-IP-21-3-2.pdf` (retrieved January 23, 2009)
15. Stanford, S.: Negative Numbers and Other Frauds (2003), `http://www.xeeatwelve.com/articles/negative_numbers.htm` (January 23, 2009)
16. Naylor, M.: More Fun with Algorithms. Teaching Pre K – 8. ProQuest Education Journals 37(8), 38–40 (2007)
17. Holmes, K.: Planning to teach with digital tools: Introducing the interactive whiteboard to pre-service secondary mathematics teachers. Australasian Journal of Educational Technology 25(3), 351–365 (2009)
18. Lee, M.H., Rößling, G.: A Little of that Human Touch in the Algorithm Visualization Construction Approach. In: Sanchez, J., Zhang, K. (eds.) Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2010, pp. 1400–1404 (2010)
19. Kaufmann, H.: Construct3D: an augmented reality application for mathematics and geometry education. In: Proceedings of the Tenth ACM International Conference on Multimedia, MULTIMEDIA 2002, Juan-les-Pins, France, December 01 - 06, pp. 656–657. ACM, New York (2002)
20. Palaigeorgiou, G., Despotakis, T.: Known and Unknown Weaknesses in Software Animated Demonstrations (Screencasts): A Study in Self-Paced Learning Settings (Screencasts): A Study in Self-Paced Learning Settings. Journal of Information Technology Education v(9), 81–98 (2010)
21. de Souza, J.M.B., Dyson, M.: Are animated demonstrations the clearest and most comfortable way to communicate on-screen instructions? Information Design Journal 16(2), 107–124 (2008)
22. Despotakis, T., Palaigeorgiou, G., Tsoukalas, I.: Students' attitudes towards animated demonstrations as computer learning tools. Journal of Educational Technology & Society 10(1), 196–205 (2007)
23. Rössling, G., Velázquez-Iturbide, J.: Editorial: Program and Algorithm Visualization in Education. Trans. Comput. Educ. 9(2), 1–6 (2009)
24. Wei, L., Cheok, A.D., Mei-Ling, C.L., Theng, Y.-L.: Mixed Reality Classroom: Learning from Entertainment. In: Proceedings of the 2nd International Conference on Digital Interactive Media in Entertainment and Arts. ACM, Perth (2007)
25. Wachira, P., Keengwe, J., Onchwari., G.: Mathematics preservice teachers' beliefs and conceptions of appropriate technology use. AACE Journal 16(3), 293–306 (2008)

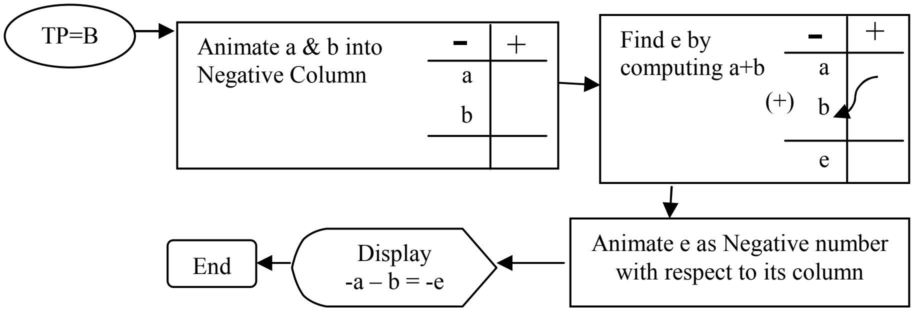
26. Zaman, H.B., Bakar, N., Ahmad, A., Sulaiman, R., Arshad, H., Yatim, N.F.M.: Virtual Visualisation Laboratory for Science and Mathematics Content (Vlab-SMC) with Special Reference to Teaching and Learning of Chemistry. In: Badioze Zaman, H., Robinson, P., Petrou, M., Olivier, P., Schröder, H., Shih, T.K. (eds.) IVIC 2009. LNCS, vol. 5857, pp. 356–370. Springer, Heidelberg (2009)

27. Varol, F., Colburn, L.K.: Investigation of critical attributes of mathematics software intended for use by young children. AACE Journal 15(2), 159–181 (2007)

Appendix A




TP=A
Construct t-table and animate Positive/Negative number into its respective column
-
+
b
a
If a>=b
False
True
Animate a into negative column
-
+
b
(-)
a
e
Animate b into positive column
-
+
a
b
(-)
e
Find e by computing b - a.
Find e by computing a - b.
Animate e as negative number with respect to its column
Animate e as positive number with respect to its column
Display a – b = -e
Display a – b = e
End


TP=B
Animate a & b into Negative Column
-
+
a
b
Find e by computing a+b
-
+
a
(+)
b
e
Animate e as Negative number with respect to its column
Display -a – b = -e
End

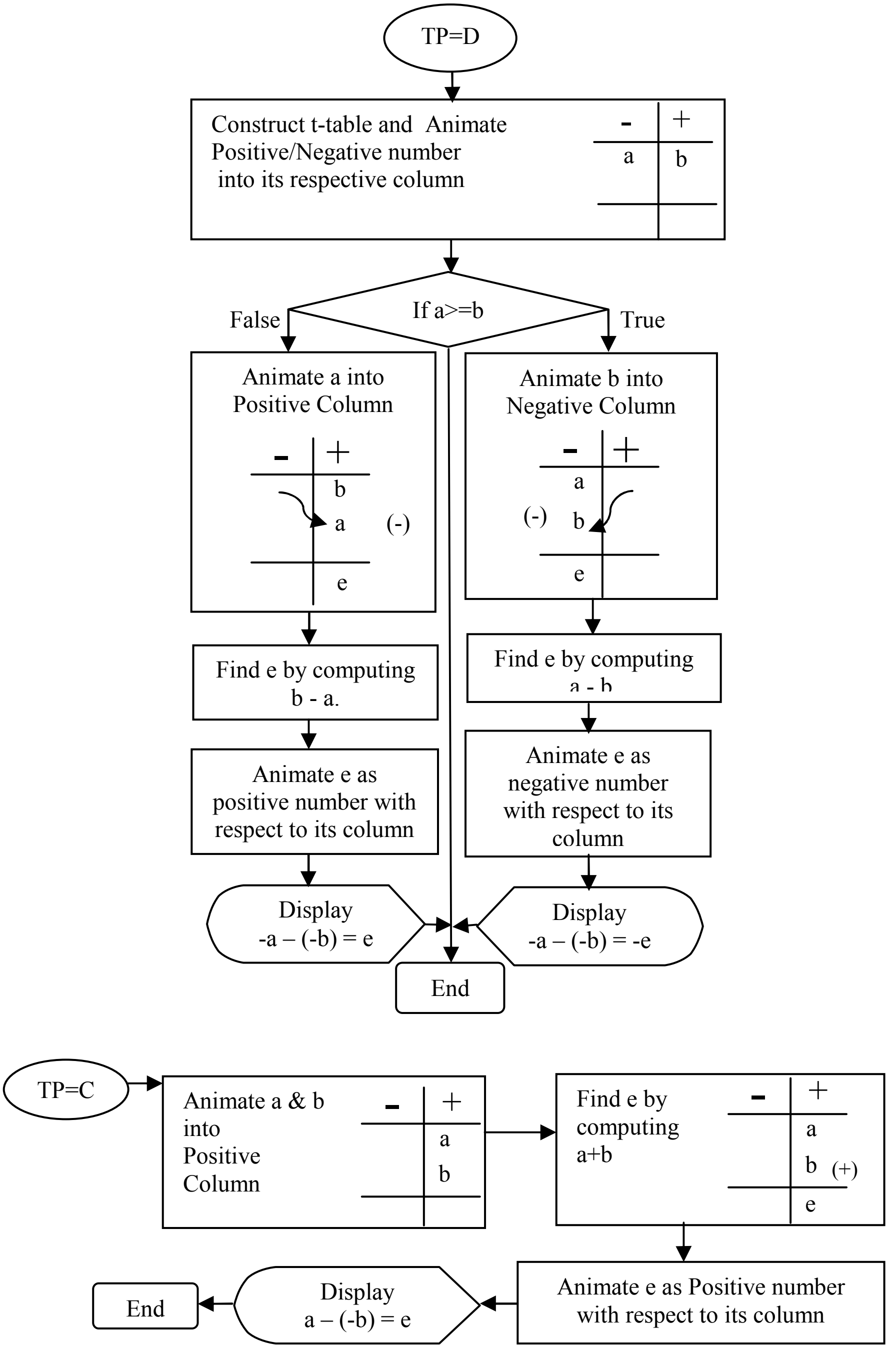

TP=D
Construct t-table and Animate Positive/Negative number into its respective column
-
+
a
b
If a>=b
False
True
Animate a into Positive Column
-
+
b
a
(-)
e
Animate b into Negative Column
-
+
a
(-)
b
e
Find e by computing b - a.
Find e by computing a - b
Animate e as positive number with respect to its column
Animate e as negative number with respect to its column
Display -a – (-b) = e
Display -a – (-b) = -e
End
TP=C
Animate a & b into Positive Column
-
+
a
b
Find e by computing a+b
-
+
a
b (+)
e
Animate e as Positive number with respect to its column
Display a – (-b) = e
End

Fig. 1. Flowchart of AVCTP

A Framework for Defining Malware Behavior Using Run Time Analysis and Resource Monitoring

Mohamad Fadli Zolkipli and Aman Jantan

School of Computer Science, Universiti Sains Malaysia, USM
Penang, Malaysia
fadli@ump.edu.my, aman@webmail.cs.usm.my

Abstract. Malware analysis is the process to investigate malware operation in order to learn and understand that malicious intent. Two common techniques that can be used to analyze malware are static analysis and dynamic analysis. Nowadays, many malware writers try to avoid security checking by implement techniques such as anti-reverse engineering, packing and encryption. It was make static analysis difficult to be implemented. In this paper, we propose a new framework to analyze malware by using dynamic approach. This framework will define malware behavior through run time analysis and resource monitoring. The contribution of this study is the new framework for defining malware behavior based on operation and target operation of the malware.

Keywords: malware, dynamic analysis, behavior analysis, run time analysis, resource monitoring, secure environment.

1 Introduction

Malicious software or malware is software that has malicious intent [1-3] to create harm to the computer or network operation. Malware can be describing as a collective term for any malicious software which enters system without authorization of user of the system[4]. Malware attack was become worldwide issues nowadays because of the number of malware growth was dramatically. McAfee reported that in 2010 an average of 60,000 of new malware threats was identified each day, which is nearly four times the 16,000 detected per day in 2007 [5].

Investigating malware goals and characteristics is very important task because that information is very useful in designing and implementing prevention mechanism on the computer systems as well as data network. In order to maintain computer operation, learning and understanding malware is the best practices to minimize threats from malware writers. This opportunity was not provided by signature-based technique because the detection process is only based on the string matching without knowing the goals and characteristics of the malware.

Learning and understanding malware can be done by using specific analysis process. Two common techniques that can be used to analyze malware are static analysis and dynamic analysis. Basically, both analysis technique are complementary

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 199–209, 2011.

in order to completely define malware [6]. Static analysis is the best technique on analyze malware compare to the dynamic analysis. Main advantage of static analysis is provide complete analysis because it not just jump to a specific execution of a program but can give guarantees that apply to each executions of the malware program [6]. However, there are some issues that related to current situation that divert the analysis process to the dynamic analysis technique. Most of the malware often uses anti-reverse engineering techniques to escape from security checks and to make to make static analysis process difficult [7]. It also has a limitation to deal with polymorphism, metamorphism [8], code encryption and packing techniques that implemented by malware writer.

In this paper, we propose a new framework for dynamic malware analysis using real time analysis and resource monitoring. The purposes of this propose frameworks are:

a. To analyze malicious program using run time analysis by focusing to the operation of the malicious program.
b. To observe malicious activities through resource monitoring based on the target operation during run time analysis.
c. To define malware behavior based on the result from both processes.

The proposed framework has three main processes such as run time analysis, resource monitoring and behavior definition. The research only focuses on the host-based malware attack which is end user's computer.

The rest of the paper is organized in the following way. Section 2 is devoted to a concept of malware behavior analysis. Section 3 discusses the background and related work on static and dynamic analysis. Section 4 presents our framework for dynamic malware analysis using real time analysis and resource monitoring. Section 5 presents experimental result and the workability of our framework. Finally, a few remarks and a discussion of future research are stated as a conclusion in Section 6.

2 Malware Behavior Analysis

Malware behaviors analysis is the process to understand the actions and characteristics of malicious software. The behavior describes the purpose and the function of the malware attacks. Understanding malware behaviors is very important and critical tasks in order to use it for definition and classification of the new malware. It is different from the signature-based detection method as the detection process will be done without knowing the behavior of the malware.

Normally, each malware samples has only one major behavior. However, it also can have multiple of behaviors based on the complexity of the program. Let B be the set of behavior that malware sample M can perform. The behavior $b \in B$ is done by M as one of the behavior. It mean that $M = \{b_1, b_2, b_3,\ldots,b_n\}$. Malware samples that belong to the same family often shared the same behavior because it have the similar purpose and function [9]. For example, two pieces of malware M_1 and a piece of malware M_2 were executed. Malware in the same family F will have the same behaviors based on the b. In other words malware family can be explained as shown below.

$M_1 = \{b_1, b_2, b_3, \ldots, b_n\}$
$M_2 = \{b_1, b_2, b_3, \ldots, b_n\}$

If $M_1 \subset F$ and $M_2 \subset F$, it mean that both malware in the same family (M_1 U M_2) because it shared the same behavior b_1, b_2, b_3.

Each malware types have specific characteristic and can be identified based on the behaviors and spread manner. Malware is a program with malicious intent designed to damage the computer on which it executes or the network over which it communicates [10]. Table 1 shows the general description for each malware types. Although all types of malware have their specific objective, the main purpose is to break the computer operation.

Table 1. General description for five common malware types

Type	Description
Virus	A self replicating program that infects a system by attaching itself to another program for the purpose to create damage
Worms	A self replicating program that uses a network to send copies of itself to other computers on the network just to spread and don't attempt to alter the systems
Trojan Horse	A deceptive program that appears harmless but have the ability to deliver destructive payloads and unload viruses, worms or spyware
Spyware	A sneaky program that tracks and reports your computing activity include sudden modifications to your application
Rootkits	A single program or collection of programs designed to take complete control of a system

There are basically three characteristics associated with these bad malware types such as self-replication, population growth and parasitic [1]. Selfreplicating malware actively attempts to propagate actively or passively by creating new copies of it. The population growth of malware describes the overall change in the number of malware instances due to self-replication. Parasitic malware requires some other executable program code in order to be executed.

3 Related Works

Nowadays, many researchers focus on the behavior-based technique that applies current malware detection strategies. Interest on using those strategies is to overcome the limitation of the static malware detection approach. There have been several works related to the behavior-based detection. Basically, there are two main approaches of malware analysis that commonly used such as static analysis and the other is dynamic analysis. Previous works was applying both analysis approaches in order to analyze the complex characteristics and behaviors of malware that harm the computer operation.

Static analysis utilizes the information in suspected executable programs without running it [11]. This approach also known as white-box analysis where malware sample is disassembled in order to understand the workflow, function and code structure of the program. The process was done by examining the program files without running the source code. It can be performed in a static way by using tools such as disassembler, decompiler or generic unpacker. Static analysis is the best approach that can catch malware sample before it execution and can cause the damage [11]. However, the main problem of this approach is the technique to disassemble the program because most of the malware codes are obfuscated by great variety of packers [12].

Bergeron *et al.* proposed a new approach for the static detection of malware code in executable programs [6]. This approach carried out directly on binary code using semantic analysis based on behavior of unknown malware. The reason for targeting binary executables is that the source code of those programs that need to detect malicious code is often not available. The primary objective of the research is to elaborate practical methods and tools with theoretical foundations for the static detection. The experiment was done in three steps such as generating an intermediate representation by using IDA32 Pro, analyzing the control and data flows and doing static verification by using their own prototype tool.

Purui *et al.* also proposed a static analysis method of studying the multiple execution paths encountered during malware analysis [8]. The method extracts paths to identify typical behaviors of the malware that focusing the analysis on the interaction between malware and environment. The method generates different inputs for the malware, based on the reverse analysis of path selection conditions. Prototype of the system was developed by modifying QEMU's translation and execution module. The result showed that the system could identify typical behaviors of malware without exploring all its possible paths.

Dynamic analysis or black-box approach utilized to monitor malware execution and to analyze a malware's behavior [8]. The malware sample is executed in an environment that is designed to closely observe its internal activities in detail [12]. Analysis process can be done by using analysis tool such as Anubis [13], CWSandbox and Cobra without disassemble or unpack the malware program. It not provides sufficient insight into the program because this approach only concentrates on external aspects of malware as it interacts with its environment. Normally, multiple paths should be triggered in order to increase the coverage of analysis [8] because it only observes a single execution path [12].

Syed Bilal *et al.* (2009) proposed a real-time malware detection scheme by using dynamic analysis that known as IMAD [14]. It analyzes the system call sequence of a process to classify it as malware or benign program. IMAD used Genetic Algorithm to optimize system parameters of the scheme to detect in-execution zero-day malware attack. Basically, IMAD not just the detection because it also have it own ngc786 classifier which can classify on the basis of variable length feature vector that useful in many critical realtime control systems.

Ulrich *et al.* (2010) presented an approach to improve the efficiency of dynamic malware analysis systems [15]. It is to overcome the huge number of new malicious files currently appears is due to mutations of only a few malware programs. The proposed system avoids analyzing malware binaries that simply constitute mutated instances of already analyzed polymorphic malware. It can drastically reduce the amount of time required for analyzing a set of malware programs. The limitation of this approach is due to the changes of the behavior after the analysis process that cause by the limitation of dynamic analysis.

4 Proposed Framework

As mentioned before, the goal of this paper is to present a proposed framework for defining malware behavior using dynamic approach. Compare to the code analysis, this approach is more acceptable in term of cost and time consuming. The main reason is it does not require a complicated process for unpacking and program disassembly. Although the code analysis is the best way to analyze malware, it suffer from the fact that it cannot analyze the majority of malicious file due to the well-protected by malware writer [15].

Using the framework that has shown in figure 1, we can analyze malicious file by identifying it operation through run time analysis. Resource monitoring will identify the target operation of the malware sample. The results from both processes will be used to define the behavior of that malware.

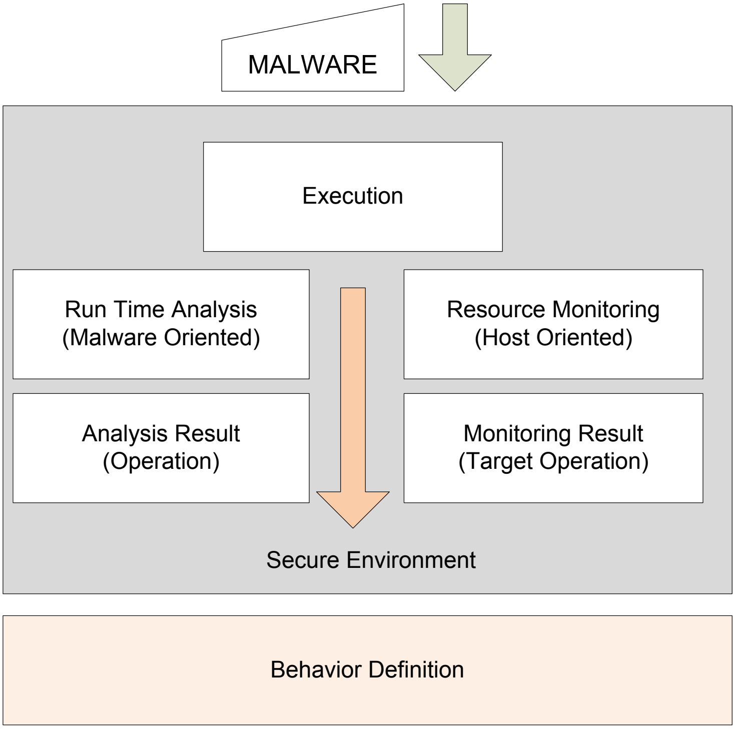


Fig. 1. Framework for defining malware behavior using run time analysis and resource monitoring

4.1 Secure Environment

Malware writers are one of the first users that recognize the useful of virtual machine. They normally used virtual machine on their product testing. Several environments will be created in order to test the workability of their new malware program. The successful of the new malware program then can be measure by observing the way of the program exploit. It is similar to the monitoring monkey behaviors in the zoo by creating an environment same like the actual habitat in the jungle.

In malware analysis, the purpose of virtual environment is to create secure environment in order to minimize damage to the actual computer resources when the malware sample executed. It is to avoid damage to the real operating system and computer resources if the malware executed. Once the process of malware analysis completed, the infected environment can be destroyed without contaminate the actual recourses. The virtual machine tools that normally used in dynamic malware analysis are VMware[16] and Microsoft Virtual PC[17].

Virtual machine operating system is the common solution on doing malware behavior analysis because malware often pose strong threat to the computer system. Virtual machine also can provide a tightly controlled set of resources because untrusted process cannot run out of the virtual machine [9, 18]. It can provide the better observation facilities because it has a total access of the memory space to gaining firm information about what a piece of program does. However, there are some malware samples that try to prevent against malware analysis that used virtual environment tool [19].

In our propose framework, we decided to improve secure environment infrastructure that use real Windows XP operating system with virtual machine tools. Running the malware sample to extract it behaviors in secure environment can reduce the potential damage to the real operating system and resources. Both analysis processes in our framework will be implement using specific host that apply secure environment setup.

4.2 Run Time Analysis

Run Time Analysis is the process to observe malware operation during execution. Malware process will be observed by using the existing kernel callback mechanism of the host kernel. These callbacks invoke functions inside of a kernel driver and pass the actual event information as analysis result. It is a possible way to observe malware operation such as read, write, or delete registry entries and files in real time. The tool that will be used for this run time analysis is Capture BAT [20] and Wireshark [21].

Capture BAT was originally not design for malware behavior analysis tool for Window environment. It is a client honeypot tool that developed in open source projects to find malicious servers on a network [22]. Capture BAT not only working on user level, it also observe the malware process at kernel level by using call-backs technique. It analyzes the state of the operating system and applications that execute on the system by monitoring the file system, process monitor and registry and generating reports for any events. Wireshark is a network monitoring tool that computer network event. In conducting run time analysis process, Wireshark is useful if the malware samples have ability to establish connection with the remote hosts in order to request objects. It can capture the information such as remote IP address, URL and port of target host.

From the analysis report, information about operation of that malicious code will be extract from registry events. Operation is referring to the malware actions that react to the host resources. That information will be use as input information in the run time analysis process.

4.3 Resource Monitoring

Resource monitoring is the process to monitor the changes in computer resources that cause by malware execution. This process can observed malware directly in windows environment and allows malicious action to be monitor effectively. All the program actions are recorded from the start of execution until end of the execution process. However, if the malware infection reached the target, it will totally infect that computer. It has the possibility to restore the computer environment into the normal state as soon as the monitoring process completed. Based on that reason, this process will be done under the secure environment that has been implemented.

This process also relies on kernel level inspection which is using API hooking and function call. Kernel mode API hooking applies a similar approach with user mode API hooking but working at different level. This technique insert the monitoring code into the kernel level itself in order to monitor state changes on host. Function call is the process of reference to a piece of executable code in order to pass an argument to other subroutine of the program. This allows a user mode software layer to call a function that defined in a kernel layer. It operates as interface between kernel and user mode to supports kernel level to notify an application at user mode about the state changes on the system. Available tool that apply this technique are, Process Monitor [23] and API Monitor[24].

This process will observed the executed malware by focusing to the specific target in Window systems. The operation of file creation and process at specific location will be recorded as a resource monitoring result.

4.4 Behavior Definition

This process will be used to define malware behavior based on the result from both processes which are run time analysis and resource monitoring. Behavior definition involves three stages of process such as features extraction, filtering and defining.

Features extraction is the process to extract the analysis results from previous processes (run time analysis and resource monitoring). All malicious activities that have been monitored will be extracting as features that need to be filter. Filtering process will select possible features by exclude the noise and unwanted information. In this framework, features about malware operation will be filter from run time analysis and features about target operation of malware will be filter from resource monitoring. Then, all that features will be select as attributes to define malware behaviors. The number of behaviors for each malware samples is based on the complexity and operation of that sample.

5 Experiment

Process flow in figure 2 describes about step by step process in this framework. Each malware samples must go through all that process in order to be analyzed and defined the behaviors. The malware sample must be executed at least twice in two different processes which are run time analysis and resource monitoring. Then, features from both results will be extracting for the filtering process. Malware features from both processes will be filter together in order to match the operation with it target. Match features will become an attribute for defining malware behavior.

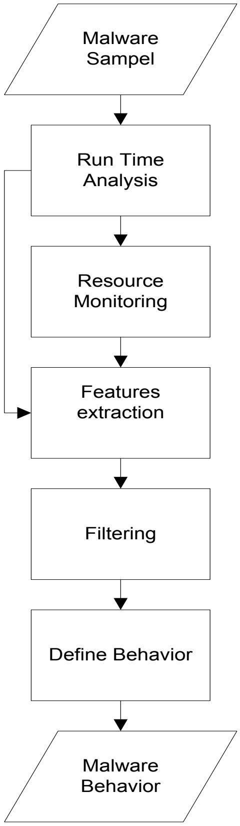


Fig. 2. Process flow for defining malware behavior

Five PE file format of malware samples were chosen to be explored in this study. These simple malwares were chosen in order to describe workability of our approach. Table 2 shows the file name of selected malware samples and also sizes of those samples. Kaspersky naming scheme was chosen to describe the types of that malwares.

Table 2. Selected malware samples

Malware Sample	Saiz (bytes)	Types	Kaspersky Naming
init.exe	293,684	Backdoor	Backdoor.Win32.Agent.bekw
jojo.exe	387,584	Trojan Horse	Trojan.Win32.Autoit.cm
winsyssrv.exe	43,373	Worm	Worm.Win32.AutoRun.ejk
nvsvc32.exe	81,920	Worm	IM-Worm.Win32.Yahos.gm
syschk.exe	237,568	Worm	P2P-Worm.Win32.Agobot.b

The behaviors were defined according to the active processes during execution and specific target location that malware samples executed. Information that related to network activities also defined according to the network monitoring result. Samples init.exe, winsyssrv.exe and syschk.exe only show one behavior each that related to host. Sample nvsvc32.exe has one behavior in host and one behavior on network activities. Sample jojo.exe shows four behavior in host and also one behavior on network activities. Table 3 shows the number of defined behavior that generated by using our proposed framework.

Table 3. The number of defined behavior for those selected malware samples

Malware Sample	Number of Defined Behaviors
init.exe	1
jojo.exe	5
winsyssrv.exe	1
nvsvc32.exe	2
syschk.exe	1

6 Conclusion

In this study, we have reviewed and analyzed the existing malware analysis techniques. From the analysis we have proposed a new framework for defining malware behavior using run time analysis and resource monitoring. Both processes in dynamic analysis shall complement in the process of defining malware behavior based on operation and target of malware attack. This research is a preliminary worked for malware behavior identification. This framework is a part of our proposed work on behavior-based malware identification and classification. This will contribute ideas in malware detection field especially in dynamic analysis in order to overcome the current issues by generating specific behavior definition.

Acknowledgments. The authors would like to thank the members in the Computer Security and Forensic Lab and also Security Research Group for their helpful discussions and suggestions. This work was supported by Short-term Grant

No.304/PKOMP/639021 and No.1001/PKOMP/822126, School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia.

References

1. Aycock, J.: Computer Viruses and Malware. Springer, Heidelberg (2006)
2. Christodorescu, M., Jha, S., Seshia, S.A., Song, D., Bryant, R.E.: Semantics-Aware Malware Detection. In: IEEE Symposium on Security and Privacy, pp. 32–46 (2005)
3. Jian, L., Ning, Z., Ming, X., YongQing, S., JiouChuan, L.: Malware Behavior Extracting via Maximal Patterns. In: 1st International Conference on Information Science and Engineering (ICISE), pp. 1759–1764 (2009)
4. Idika, N., Mathur, A.: A Survey of Malware Detection Techniques. In: Technical Report SERC-TR-286. Department of Computer Science, P.U., SERC, ed. (2007)
5. McAfee: McAfee Threats Report: Third Quarter 2010. Threats Report (2010), http://www.mcafee.com/Q3_Threat_Report
6. Bergeron, J., Desharnais, M., Desharnaias, J., Erhioui, M.M., Lavoie, Y., Tawbi, N.: Static Detection of Malicious Code in Executable Programs. Int. J. of Req. Eng. (2001)
7. Gérard Wagener, R.S.a.A.D.: Malware Behaviour Analysis. Journal in Computer Virology 4, 279–287 (2008)
8. Purui, S., Lingyun, Y., Dengguo, F.: Exploring Malware Behaviors Based on Environment Constitution. In: International Conference on Computational Intelligence and Security, CIS 2008, vol. 1, pp. 320–325 (2008)
9. Hengli, Z., Ming, X., Ning, Z., Jingjing, Y., Qiang, H.: Malicious Executables Classification Based on Behavioral Factor Analysis. In: International Conference on e-Education, e-Business, e-Management, and e-Learning, IC4E 2010, pp. 502–506 (2010)
10. Preda, M.D., Christodorescu, M., Jha, S., Debrey, S.: A Semantics-Based Approach to Malware Detection. ACM Transactions on Programming Languages and Systems 30 (2008)
11. Tzu-Yen, W., Chin-Hsiung, W., Chu-Cheng, H.: A Virus Prevention Model Based on Static Analysis and Data Mining Methods. In: IEEE 8th International Conference on Computer and Information Technology Workshops. CIT Workshops, pp. 288–293 (2008)
12. Inoue, D., Yoshioka, K., Eto, M., Hoshizawa, Y., Nakao, K.: Malware Behavior Analysis in Isolated Miniature Network for Revealing Malware's Network Activity. In: IEEE International Conference on Communications, ICC 2008, pp. 1715–1721 (2008)
13. Ulrich, B., Imam, H., Davide, B., Engin, K., Christopher, K.: A View on Current Malware Behaviors. In: Proceedings of the 2nd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More. USENIX Association, Boston (2009)
14. Syed Bilal, M., Ajay Kumar, T., Muddassar, F.: IMAD: In-execution Malware Analysis and Detection. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation. ACM, Montreal (2009)
15. Ulrich, B., Engin, K., Christopher, K.: Improving the Efficiency of Dynamic Malware Analysis. In: Proceedings of the 2010 ACM Symposium on Applied Computing. ACM, Sierre (2010)
16. VMware (2010), http://www.vmware.com/
17. Microsoft, Microsoft Virtual PC (2010), http://www.microsoft.com/windows/virtual-pc/

18. Vasudevan, A.: MalTRAK: Tracking and Eliminating Unknown Malware. In: Annual Computer Security Applications Conference, ACSAC 2008, pp. 311–321 (2008)
19. Willems, C., Holz, T., Freiling, F.: Toward Automated Dynamic Malware Analysis Using CWSandbox. IEEE Security & Privacy 5, 32–39 (2007)
20. Capture BAT, http://www.honeynet.org/project/CaptureBAT (2010)
21. Wireshark (2011), http://www.wireshark.org/
22. Seiferta, C., Steensona, R., Welcha, I., Komisarczuka, P., Endicott-Popovsky, B.: Capture – A Behavioral Analysis Tool for Applications and Documents. Digital Investigation 4, 23–30 (2007)
23. Microsoft, Process Monitor (2010), http://technet.microsoft.com/en-us/sysinternals/default
24. API Monitor, http://www.apimonitor.com/ (2010)

Success Factors in Cost Estimation for Software Development Project

Zulkefli Mansor[1], Saadiah Yahya[2], and Noor Habibah Hj Arshad[3]

[1] Faculty of Information Technology Industry
Universiti Selangor, Jalan Timur Tambahan, 45600 Bestari Jaya, Selangor, Malaysia
Tel: 6019-3937795; Fax: 0355238733
kefflee@unisel.edu.my
[2] Computer Technology and Networking Studies,
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
Tel: 6055211150; Fax: 603-55435501
saadiah@tmsk.uitm.edu.my
[3] System Science Studies, Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
Tel: 60355211241; Fax: 603-55435501
habibah@tmsk.uitm.edu.my

Abstract. Cost estimation process becomes a crucial factor in any software development project. There are many previous researches discussed the success factors that influence in software development project. This paper aimed to discuss factors that influences to the successful of cost estimation process in software development project. Literature survey is carried out from the past researches. Then, this paper presents the success factors that bring to the effectiveness of cost estimation in software development project. From the review, a conceptual model was developed to show the influence factors in cost estimation process. Realisation these factors will help software development communities contribute positively to the success of cost estimation process in software development project.

Keywords: Project Management, Success Factors, Software Development Communities, Cost Estimation.

1 Introduction

The accuracy of cost estimation result is important in any software development project. It becomes the most popular and interesting issues to be discussed when people wanted to develop any project from time to time. Information system is defined as interaction between people, process, data and technology [1]. The crucial question in software development project is how to complete a project in specific time, budget and resources. In order to measure these three attributes are achieved, person who involved in estimation process especially project manager need to

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 210–216, 2011.

measure all requirements are considered and well-defined [2],[3],[4] added project schedule overrun is one of the main contributors to project failure. Proper and complete resource identification is needed in order to start estimation process.

With consideration of all the requirements, the cost estimation process becomes easier and will produce accurate result. In common practices, cost estimation process is crucial at the early phases of software development. However, it can be in other phases for example in development and deployment process [5]. This is due to changes of needs and requirements from time to time.

The accuracy of cost estimation is depending on how software development communities defined the resources needed and the quantity of the resource. Meaning that, they need to thoroughly analyze these two elements during the project planning phases. Zhang et al [6] has listed the important of cost estimation accuracy such as (i) it can help to classify and prioritize development projects with respect to an overall business plan, (ii) it can be used to determine what resources to commit to the project and how well these resources will be used, (iii) it can be used to assess the impact of changes and support re-planning, (iv) projects can be easier to manage and control when resources are better matched to real needs and (v) customers expect actual development costs to be in line with estimated costs.

This paper discusses on cost estimation process and the existing and new success factors in software development project.

2 Research Method

In this paper, literature survey is carried out order to gather related information. Then from the discussion, the conceptual model is developed to show the relation between existing influence factors with new factors that contribute to success software cost estimation process.

3 Software Cost Estimation Process

Software cost estimation process begins during the planning phases in Software Development Life Cycle (SDLC). The process starts with identifying the types of resources and the quantity needed. The resources include list of hardware and software, training session, testing activities, infrastructure and few more. When project manager is assigned to manage certain project, he or she needs to understand the entire project first in order to start planning the tasks. Then, he or she will start identify and investigate the types of resources and quantity needed. Then, the identified resources will be listed and estimated.

For example in developing a personal website, in order to develop the modules, project manager needs to list the type of hardware and software needed. Furthermore, the number of team member also needs to be considered. In this case, who need to be hired: the web developer or web designer or web programmer or others? If the web designer and web programmer are needed, the next thing the project manager needs to think is how many of them to be hired. This scenario is one of the examples on how resources identification is conducted. Then, project manager will start estimating the project cost from the list of resources needed. Wrong identification of resources at early planning stages could cause incorrect results in estimation process and may lead to over budget. Due to this scenario, project manager needs some kind of tool to help them in resources planning and cost estimation process.

Many techniques can be implemented in both processes (identifying the resources and estimating process) such as expert judgement, parametic model, top-down approach, bottom-up approach, price-to-win, analogy and few more. However until the date, researchers are still investigating the best technique to be applied in both processes. Based on the literature, many researchers try to combine few techniques which are called the hybrid model in estimation process in hoping to produce accurate results. But, this matter is still under investigation phase. However, [7] found expert judgment and COCOMO II are the best techniques that can be applied in web-based application. From the literature, the most technique used to estimate the cost is by expert judgement [6], [8],[9],[10],[11] and COCOMO II [6],[7],[12]. Expert judgment is based on historical experience owned by the project manager may be or the person who are directly involved in estimation or budgeting cost. Normally, the estimation result is more accurate. However, some researchers felt that it can be bias since expert judgement is human [10].

4 Success Factors in Cost Estimation

Standish CHOAS Report in 1994 reported that five most factors in influencing a successful project are user involvement, executive management support, clear statement of requirements, proper planning and realistic expectations. Previous research [13] and team member's technical skills are the top three factors that contributed to the successful of any software development project [5,.[8]. Project manager also plays important roles and become important factor in the successful of any project [3],[8], [14],[15]. However, there are few more factors need to be considered such as entertainment, role of sponsors, changes of company policies, proper tool selection, suitable estimation technique and suitable software development methodology that contribute to the successful of cost estimation.

4.1 Entertainment

Overlook in entertainment cost is a serious matter in causing over budget in software development project. Entertainment cost is not only entertaining the client but includes having outside meeting with supplier or top management or other stakeholders. Some software development communities do not include entertainment cost in their estimation and budgeting process because they thought entertainment cost is not included in the expenses and they tend to use their own pocket money. However, if they look deeply into this matter, sometimes it causes them a big amount of money. As a result, they might claim these expenses under the project budget. Therefore, the actual cost is sometimes exceeded the allocation budget.

4.2 Role of Sponsor

Sponsors are commonly seen as providing or contributing to the allocation of resources and budget of a project. The main role of sponsors is to ensure a project is managed within the budget. Even though the sponsor is not fully involved in the project but somehow the project manager needs to report frequently to the sponsor. In common practice, the sponsor will influence the decision made. Therefore, the involvement of sponsor is needed in order to make sure the identification of resources is done properly. Proper identifying of the resources will ensure the cost estimation process is properly done and perhaps will result an accurate estimation. Without the involvement of sponsors, team members will simply define the types and quantity needed for the project. Therefore, the possibility of cost increment and over estimation of budget is high for unnecessary resources.

4.3 Changes of Company Policy

The changes of company policy can be one of the contributions for the success of cost estimation process in software development project. The changes include change of top management, change of technology, change of governs, change of environment and many more. All the changes are strongly affect to cost estimation process. In certain cases, an organization might have their own policy on resources, budget and duration of project. Therefore, software development communities need to consider the company policies in doing the estimation. Normally, the changes affect during the early stage, in progress or after the estimation process. However, the right time to consider any changes is at the early stage of the estimation process.

4.4 Proper Tool Selection

Choosing the right proper tool is important in cost estimation process. The right proper tool produce an accurate results. Studied by Zulkefli et al [7] showed that most of the cost estimation process was done manually. The traditional common tool used in estimating the cost is spreadsheet or Microsoft excel and Microsoft project. The biggest challenge in using the traditional method is the level of accuracy. Software development communities faced difficulty to achieve high accuracy in producing cost estimation result. Therefore, many studies have been done to develop automated tool for cost estimation process. However, no one claimed their proposed tool can produce accurate result.

Most of the researchers agreed that to produce accurate cost estimation is hard and crucial [12],[15],[16],[17]. From the literature, there is less accurate tool that can be used to estimate cost in any IS project. This is due to the different and various requirements and needs by the users or project developers [8],[18]. Even though, many researchers have tried to construct their own tool, until the date, nobody can claim their tool would produce good and accurate and widely acceptable estimation. Surprisingly, most of the process is done manually [12]. That shows, most of software development communities did not rely on automated tool.

Another reason is traditional tool does not provide a proper way in recording and tracking previous result of certain project. In some cases, the similarity from previous result can be applied for a new project. Therefore, a proper tool that provides good keeping and tracking system could help in order to estimate the cost for a new project. Common practice in using spreadsheet or Microsoft project is fragile. It might cause problem to track records that is missed place or the softcopy of the file is corrupted. However, a proper tool will provide high level of result availability. Therefore, proper tool plays an important role in ensuring the accuracy and availability in cost estimation process.

4.5 Suitable Estimation Technique

In software cost estimation process, there are few techniques that can be applied to estimate the cost. For examples, the expert judgment, top-down approach, bottom-up approach, price-to-win, rules of thumb, parametric model, analogy, time boxes and many more. However, until the date, there is no research that can ensure which technique is the most suitable in cost estimation process. Therefore, many researches have been done investigating the most suitable technique that can be applied. Choosing the right cost estimation technique is important to ensure the result is accurate. Different approach applied in the cost estimation techniques might produce different accuracy of result. There are few researches have been carried out to integrate more than one technique which is called the hybrid technique. Yet, no one can claim which technique is the best. Therefore, there is no right and wrong approach in choosing cost estimation technique in a project. The most important matter is to choose the right technique which is suitable to the project. Sometimes, few techniques might be applied in a single project but it caused the increment of cost and time [19].[20] However, this does not mean the result is better.

4.6 Suitable Software Development Methodology

Software development methodology plays an important role in cost estimation process. However, some software development communities just ignore it in the estimation process. The examples of software development methodology are agile, spiral, waterfall, Rapid Application Development, prototyping and many more. Each software development methodology provide different step which contributes in cost estimation process. For example traditional method (example waterfall) provides different process compared to agile methodology. The simplest process can affect the way the software development communities do the estimation. Less step or process might decrease the cost involved. Therefore, to secure the cost estimation process, the most suitable and correct software of development methodology is needed.

5 Results

From the discussion, the authors found other factors such as entertainment, role of sponsor, changes of company policy, proper tool, suitable estimation technique and suitable software development methodology that really contribute to the successful in cost estimation are influence in cost estimation process is success. All of these factors need to be well considered and defined in making sure of the success estimation process. Therefore, these factors are then added as a better value in the conceptual model shown in figure 1. The straight line of the rectangle indicates the existing factors were discussed by many researchers and the dot line indicates the new factors influences in cost estimation success.

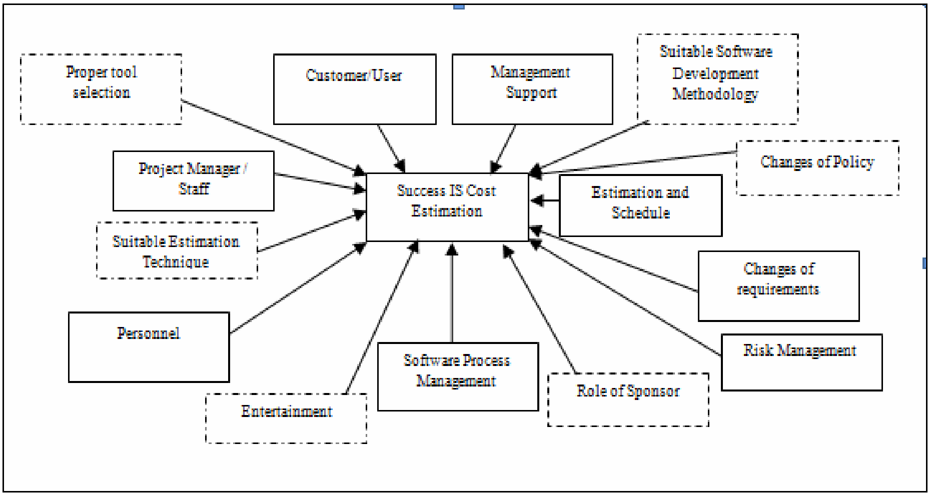


Fig. 1. Conceptual Model for Success Software Cost Estimation Process

6 Conclusion

In conclusion, research in cost estimation process has been carried out for decades with huge number of researches have to look into the various aspects in the estimation process. This study has looked at the success factors of cost estimation process in software development project and developed a conceptual model. Therefore, by considering these factors, perhaps cost estimation process can produce more accurate result. In the next steps, the determine factors will be evaluated through hypothesis test and a questionnaires survey.

References

[1] Schwalbe, K.: IT Project Management, 3rd edn. Thomson Course Technology, Canada (2004)

[2] Iman, A., Siew, H.O.: Project Management Practices: The Criteria for Success of Failure. Communication of the IBIMA 1(28), 234–241 (2008)

[3] Ostvold, K.M., Jorgensen, M.: A Comparison of Software Project Overruns – Flexible versus Sequential Development Model. IEE Transactions on Software Engineering 31(9), 754–766 (2005)

[4] O'Brien, R.: Critical Path Analysis: The Path to Project Enlightenment. TechRepublic, (March 29, 2004), `http://techrepublic.com/5102-6330-5175455.html` (retrieved November 10, 2010)

[5] Seetharaman, N., Senthilvelmurugam, M., Subramanian, T.: Budgeting & Accounting of Software Cost. Journal of Digital Asset Management 1(5), 347–359 (2005)

[6] Zhang, J., Lu, T., Zhao, Y.M.: Study on Top down Estimation Method of Software Project Planning. The Journal of China Universities of Posts and Telecommunications 13(2), 1–111 (2006)

[7] Zulkefli, M., Zarinah, M.K., Habibah, A., Saadiah, Y.: E-Cost Estimation Using Expert Judgment and COCOMO II. In: ITSIM 2010, vol. 3, pp. 1262–1267 (2010)

[8] Zulkefli, M.: Cost Estimation Tool for Web-Based Application. Master Dissertation. Department of Software Engineering, Universiti Malaya. Kuala Lumpur (2009)

[9] Xishi, H., Luiz, F.C., Jing, R., Danny, H.A.: Neuro-Fuzzy Model for Software Cost Estimation. In: Proceedings of the Third International Conference on Quality Software (QSIC 2003), Dallas, TX, USA, November 6-7, pp. 126–133 (2003)

[10] Zhang, F., Heraton, L.: Software Cost Estimation. Department of Computing. The Hong Kong Polytechnic University (2006)

[11] Jorgensen, M.: A Review of Studies on Expert Estimation of Software Development Effort. Journal of Systems and Software 70(1-2), 37–60 (2004)

[12] Yean, Z., Hee, B.K.T., Wei, Z.: Software Cost Estimation through Conceptual Requirement. In: Proceedings of the Third International Conference on Quality Software (QSIC 2003), Dallas, TX, USA, November 6-7, pp. 141–145 (2003)

[13] Noor Habibah, A., Azlinah, M., Zaiha, M.N.: Software Development Projects: Risk Factors and Occurrences. WSEAS Transactions on Computer Research Journal 2(2) (2007)

[14] Yang, D., Qing, W., Mingshu, L., Ye, Y., Kai, Y., Jing, D.A.: Survey on Software Cost Estimation in the Chinese Software Industry. In: ESEM 2008, October 9-10, Kaiserslautern, Germany (2008)

[15] Terry, C., Davies: The 'real' success factors on projects. International Journal of Project Management 20(3), 185–190 (2001)

[16] Guru, P.: What is Project Success: A Literature Review. International Journal of Business and Management 3(9), 71–79 (2008)

[17] Geethalakshmi, S.N., Shanmugan, A.: Success and Failure of Software Development: Practitioners' Perspective. In: Proceedings of the International MultiConference of Engineers and Computer Scientists 2008, IMECS, Hong Kong, vol. I (2008)

[18] Stellman, Jennifer: Applied Software Project Management, p. 322. O'reilly, Prentice Hall (2005)

[19] Little, T.: Schedule Estimation and Uncertainty Surrounding the Cone of Uncertainty. IEEE Software 23(3), 48–54 (2006)

[20] Min, X., Bo, Y.: A Study of the Effect of Imperfect Debugging on Software Development Cost. IEEE Trans. Software Eng. 29(5), 471–473 (2003)

Transactional Web Services Composition: A Genetic Algorithm Approach

Yong-Yi FanJiang[1], Yang Syu[1], Shang-Pin Ma[2], and Jong-Yih Kuo[3]

[1]Department of Computer Science and Information Engineering, Fu Jen Catholic University, Taipei, Taiwan
[2] Department of Computer Science and Information Engineering, National Taiwan Ocean University, Keelung, Taiwan
[3] Department of Computer Science and Information Engineering, National Taipei University of Technology, Taipei, Taiwan
yyfanj@csie.fju.edu.tw, a29066049@gmail.com, albert@mail.ntou.edu.tw, jykuo@ntut.edu.tw

Abstract. Service Oriented Architecture implemented by Web Services is one of the most popular and promising software development paradigm, however, it still has some challenging issues. One of that is how to automate web services composition at design time. Services composition reuses existing component services to provide composite service with more complexes, value-added functions that cannot be provided through any single component service; therefore it avoids constructing any new service from scratch. In this paper we propose an approach based on genetic algorithm to automatically composing web service without a workflow template beforehand and ensuring resulting service has reliable behavior (transactional properties). A composite service which is produced through our approach will be able to treat as a unit of work avoiding inconsistence and it does not ask user to define the workflow template manually. Experimental results are presented.

Keywords: Web services composition, genetic algorithm, transactional Web Service.

1 Introduction

Service Oriented Architecture (SOA) usually realized by Web Services (WSs) is one of the most promising paradigm for modern software development, since it allows well-formed and autonomous components to be repeatedly reused rather than creating new one from scratch. On SOA, services composition focuses on how to employ existing WSs that are offered from diverse organizations and providing different features including functional (e.g., booking, payment) and behavioral (e.g., retriable or compensatable), to quickly construct workable applications or software for the requirements which are requested by users and unable to fulfill through any single component service. By this way, the needed cost and time to construct new software or application will be shrunken extremely and finally enhancing the competition of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 217–231, 2011.

adopter. An application or software, which is the product of services composition called a Composite Web Service (CWS).

Generally, the life cycle of a CWS has three phrases including, design time, runtime and monitoring, reengineering [1]. At the first phase, design time regard to enacting specification of a CWS. Runtime and monitoring phase concerns with execution of the CWS, and it also detects violations and errors during entire execution duration. Ultimately, reengineering phase revises the specifications of the CWS depend on the information from previous phase.

In this paper, we only aim at the problems happened at design time. The entire process of design time can be divided as three steps: (1) workflow specification, which is consisted of activities, for a CWS; (2) discovery of available candidate services on Inter- or intra-net for each activity; and (3) selection of fittest service for each activity in candidate services set. At first step, user submits a specification of requirements including functional, non-functional (preference), and behavioral perspectives. Based on the requirements from user, designer must manually (or semi-automatically) generate a workflow, which is composed of activities, to provide an abstract service backbone for matching the functional request from user. After acquiring a workflow, the process enters step 2, which involves searching services for every activity among the workflow. Many researches omit this step because they already have a repository storing all information about available services. At this time, each activity can be really driven by many realistic services. Thus, in step 3, it is picks up a best one according to certain criteria (e.g., QoS, transactional behavior) in order to gain a CWS with optimal quality or/and reliable execution.

According to surveying current works, which have been done in design time service composition, we categorize existing approaches and researches into two classes: (1) dynamic workflow service combination; and (2) transaction-aware service selection. The approaches belonging to first class [2], [3], [4], [5], [6], [7], [8] mainly address the problem occurring at the first step in design time. Namely it automates the production of workflow. Approaches contained in second class [9], [10], [11], [12], [13], [14] deal with transactional-aware service selection problem (third step), these approaches guarantee an reliable CWS.

According to categorization, most approaches merely have one of the characteristics (dynamic workflow, transaction-aware). The approaches without a handcrafted workflow template do not assure reliable execution. Contrarily, the approaches guarantying reliable execution need a manual, handcrafted workflow template in advance. In other words, they are complementary to each other from the viewpoint of characteristics.

The objective of our research is to provide a one-stepped, automatic approach for service composition in design time. Dynamic workflow and transactional requirement (reliability) will be thoroughly considered in order to mitigate the workload and burden from manually designing a CWS. The problem modeled by us is a NP-hard problem. Therefore, it would not suitable to solve it via regular methods such as AI-planning or brute force since the searching space faced by the problem has very high dimensions and the space will enlarge quickly when more services available are in repository. For that reason, the core of our approach adopts genetic algorithm (GA) as its power in searching answer for problems with enormous solution space.

Our contribution is proposing an approach based on genetic algorithm to completely cope with tasks faced by designer and occurring in the design time of service composition. The approach will select and compose reasonable component services as a functional- and behavioral-correct, executable CWS. The process of composition does not need any human effort and intervention, indeed, it is automation. It guarantees that each selected component service of a CWS corresponding to user's requirements is globally fulfilling the functional and behavioral requests from user, and it also does not ask for pre-defined workflow template and acceptable termination states. The composed CWS can be viewed as an undividable working unit to ensure the consistence of states of its component services.

The rest of this paper is organized as follow. Section 2 discusses related work and the process about how to use our approach is in section 3. Section 4 is the definition of specification. Section 5 introduces how to synthesize component services as a CWS. The detail of genetic algorithm is described in section 6. Section 7 presents and discusses the results of experiments. Finally section 8 is conclusion.

2 Related Work

The approaches with dynamic workflow mean that it does not ask for a pre-defined workflow before the approach works. On the other words, the workflow is created by the approach. Aversano *et al.* [2] use genetic algorithm to combine services, and it is QoS-aware. However, the QoS types that are considered within this approach are degree of matchmaking between two services (a service's output and subsequent service's input). Silva *et al.* [3] proposed an architecture to support all of requirements that is required when a user composes a CWS, and most importantly a graph-based algorithm used for creating workflow. The algorithm requests that all material services are placed into a CLM (Causal Link Matrix) appropriately. Lécué and Léger [4] relied on an backward style algorithm to connect services. Fujii *et al.* [5], [6] exhibit a three layered architecture that helps user to compose web services as application depend on their semantic, and it also allows the requirement describe in nature language. Otherwise, Fujii *et al.* [6] analyzed the difference between methods that are able to build template (aka workflow) or not. Lajmi *et al.* [7] and Soufiene *et al.* [8] proposed exploiting CBR (Case-Base Reasoning) to address service combination problem.

Bhiri *et al.* [9] come up with the concept of transaction pattern to help designer designing flexible and reliable CWS. Bhiri *et al.* [10] had a method to verify that whether a CWS satisfies the requested atomicity, and it will suggest how to revise if there has any fault within CWS. Li *et al.* [11] proposed the concept of safe connection point to inspect that whether the two services is able to view as a working unit. Portilla *et al.* [12] exploited contract, which is separated with business process, to assure that transaction property is achieved. Montagut *et al.* [13] had a iterative service assignment procedure to choose service according to the requested transaction property. In particular, Hadad *et al.* [14] proposed an algorithm which integrates QoS- and Transaction-aware service selection simultaneously. Table 1 shows the comparison of the related work.

Table 1. A comparison table for previous works

Author	Approach	Workflow	Transaction-aware
L. Aversano *et al.* [2]	Genetic algorithm	Produce by algorithm	No
E. Silva *et al.* [3]	Graph-based algorithm based on CLM	Produce by algorithm	No
F. Lécué *et al.* [4]	AI-planning based on CLM	Produce by algorithm	No
K. Fujii *et al.* [5], [6]	Semantic-based architecture	Produce by algorithm	No
S. Lajmi *et al.* [7] **L. Soufiene *et al.* [8]**	Case-based Reasoning	Produce by algorithm	No
J. El Hadad *et al.* [14]	Transaction-first selection	Given beforehand	Yes
S. Bhiri *et al.* [9]	Transaction pattern concept	Given beforehand	Yes
S. Bhiri *et al.* [10]	Rule based approach	Given beforehand	Yes
L. Li *et al.* [11]	Safe connection point concept	Given beforehand	Yes
A. Portilla *et al.* [12]	Contract based behavior model	Given beforehand	Yes
F. Montagut *et al.*[13]	Rule based approach	Given beforehand	Yes

3 Approach Process

This section depicts how to operate our approach. Fig. 1 illustrates entire process precisely through UML activity diagram. First of all, the user must provide the specification of required services. The specification used in the approach will be introduced in next section. Then, the approach exploits the information from user to construct a CWS. Before executing the core genetic algorithm, it matches with every single service in the services repository to confirm whether a single service

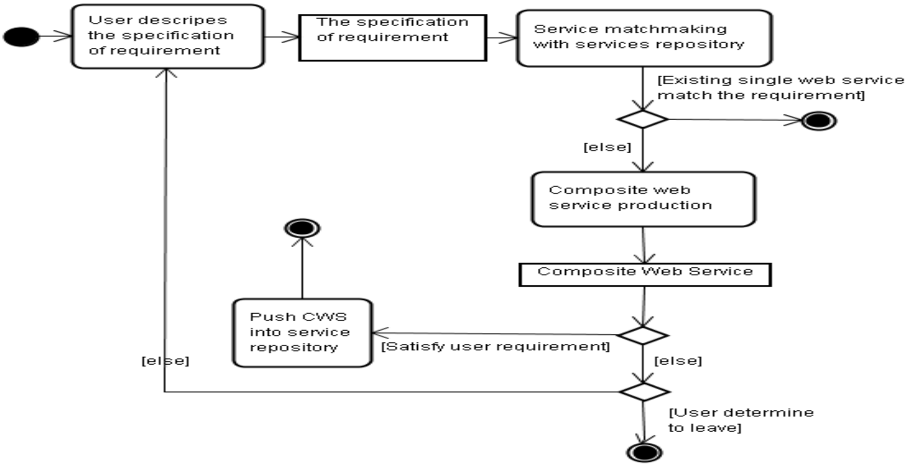

User descripes the specification of requirement
The specification of requirement
Service matchmaking with services repository
[Existing single web service match the requirement]
[else]
Composite web service production
Composite Web Service
Push CWS into service repository
[Satisfy user requirement]
[else]
[else]
[User determine to leave]

Fig. 1. The process of proposed approach

(composite or component service) is able to satisfy the user's demand existing. If a service in the repository matched with the user's demand is found, then the approach simply returns the result and saves a lot of times. However, if there is no any single service which meets the user's requirements, the point of process walks into the evolution of GA. The user will receive a CWS when the GA is completely finished. Afterward he/she decides whether accepting the CWS or not. If the user does not accept the CWS, he/she can drop and exit the approach process or redefine a specification of requirements. The user probably relaxes the original requirements since the original is hard to solve, even there is no solution existing. Finally the approach puts CWS into service repository as a component service.

4 Specification

This section defines the assumptions, notations, and symbols used by proposed approach. Section 4.1 defines the tuple of a service, which are the expression formula of a service. Section 4.2 introduces three kinds of service and explains their usage. How to model a CWS, including its order logic and structures of component services, is presented in section 4.3. Next, section 4.4 states the recognized three transactional properties (TPs) and their possible combination.

4.1 Tuple

Any type of service is uniformly represented by six tuple. These tuple $\langle ID, I, O, P, E, B\rangle$ semantically depicts the features of a service from diverse aspects. *ID* is the name or identity of a service. *IOPE* describe the functionality of a service. Input and Pre-condition (*I* and *P*) are material that a service needed. Output and Effect (*O* and *E*) represent the product of service. *B* assigns the transactional behavior (properties) of a service.

It is an assumption of our approach, an *object* described by Input and Output can be consumed only once, but a *condition* described by Pre-condition and Effect can be used infinitely. Different to other approaches, the process of service execution is stateful in our approach that means the product of a service can be viewed by not only immediate successor but also services following the producer.

4.2 Service Types

There are three types of service in our approach. Tuple with suffix "*required*" is the description of user's requirements; tuple with suffix "*provided*" is the answer built by our approach according to user's requirements; tuple with suffix "*component*" present an available component service in services repository, which is a storage storing all available resources. Below is the definition and usage of those three service types.

- $\langle ID, I, O, P, E, B\rangle_{required}$: It is used for a user to describe the specification of service that he/she wanted. $\langle O, E\rangle$ is the product that the user expected. $\langle I, P\rangle$ is the material that the user is able to provide. $\langle B\rangle$ presents the transaction properties asked by the user.

- $\langle ID, I, O, P, E, B\rangle_{component}$: It is an available resource service stored in services repository. Our approach exploits those resources to make a CWS for satisfying user requirement. The services contained in the services repository are real web services available on network or CWSs composed of component services.
- $\langle ID, I, O, P, E, B\rangle_{provided}$: Theses tuples together represent an executable, realistic CWS, which is composed of component services and created by our approach for the corresponding user's requirement. $\langle O, E\rangle$ is the product that the CWS create. $\langle I, P\rangle$ is the material that the user of CWS must enter. $\langle B\rangle$ is the transaction properties of this CWS.

4.3 Workflow and Activities

Workflow (business process) exactly describes the order logic and structure of a CWS. Currently, the mainstream to model a workflow is the WS-BPEL [15], which is a XML based orchestration language and its fundamental constitute is various activities. These activities can be divided into two categories, basic activities and structure activities. Each category has many types of activity for diverse usage, and we filter out non-vital activity type to keep simplification. In our approach, the only one basic activity is <invoke/>, which means invoking a service to complete certain work. Here we assume that a service has only one operation for simplicity. The adopted structure activities are <sequence/>, <flow/>, <while/>, <if/>, <Process/> and their exhausted semantic can be found in [15]. In order to facilitate XML operation and prevent syntax error in our approach, the WS-BPEL XML format will transforms as tree structure during the evolution of GA [2], [16], [17].

4.4 Transaction Properties

It is recognized that there are three transactional property (behavior) types for component service [9], [10], [11], [12], [13] [14]. Transactional property (TP) presents the facility of service when fault is occurred. A service with *Pivot* (*p*) means that it is able to rollback when it is not successfully complete yet. *Compensatable* (*c*) represents a service that can be undone even though it is successfully completed. *Retriable* (*r*) guarantees that a service will be repeatedly invoked until it is successfully completed. A service can have combination of TPs, and the set listing all possible property is $\{p, c, pr, cr\}$.

5 Synthesis

A CWS is a conglomeration of component services under reasonable order and structures. Consequently, there are many specifications representing component services within a CWS and these specifications have to be integrated as a single specification, which is able to precisely express a CWS. A service can be viewed from two facets including: functionality and transactional behavior. This section describes how to integrate them respectively.

5.1 Synthesis to Functionality

The tuple for convey the functionality of a service are ⟨*I*, *O*, *P*, *E*⟩, and they can be recursively synthesized. The integration of functionality specification under different structure activities has different rules, and we assume that workflow process is stateful. The order to integrate the specification starts from left to right at level 2 of tree. If the order encounters a node having subnodes (the node is a structure activity), those subnodes will be integrated recursively as single specification. When the synthesis is achieved, a functionality specification which is capable of to represent the functionality of CWS is acquired. From the CWS specification, user can realize that what material should be provided to the CWS in order to successfully execute the CWS and what product will be received after the CWS completely work. Fig. 2 is an example illustrating how to synthesize and the rules under different structure activities. For instance, the input and output of S1 and S2 are O1, O2, O3, and O2, O4, O5, respectively. The results after synthesizing under sequence structure are O1, O4 to input and O3, O5 to output, respectively.

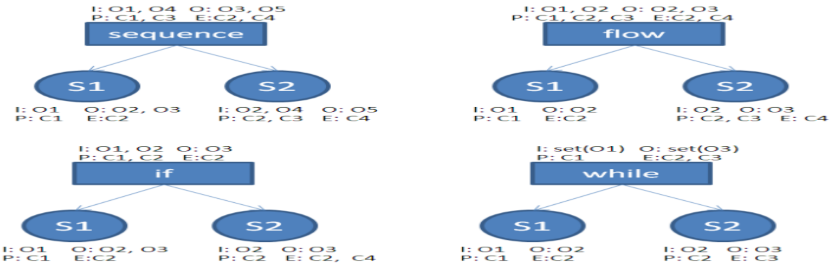


Fig. 2. The synthesis for functionality

5.2 Synthesis to Transactional Property

A CWS has one or two of the four composite service transactional properties (CTPs), where CTPs was decided by its component services TPs [14]. The CTPs adopted by our approach and how to synthesize CTPs under different structure activities are defined in [14]. The considered CTPs are atomic (ă), non-atomic (ã), compensatable (*c*), and retriable (*r*). A CWS with atomic is able to rollback when it encounters fault during its execution, but a CWS with non-atomic is unable to do so. CWS with compensatable is able to recover its effect even if its execution is over. Retriable ensures a CWS will execute repeatedly until it is success. A CWS may have two CTPs and all possible combination are, atomic ($\breve{a}$), non-atomic ($\tilde{a}$), compensatable (*c*), compensatable-retriable (*cr*), atomic-retriable ($\breve{a}r$). In [14], it define an automation to aid inferring the CTPs of a CWS under Sequence and Flow structures. But we further define the synthesis under While and If structures. The synthesis for While structure is identical with Sequence structure [11]. But synthesis for If structure is stricter, all of services under If must have same TPs, otherwise the If is with non-atomic.

6 Genetic Algorithm

This section exhausts the core of proposed approach, a genetic algorithm. Section 6.1 introduces GA and its process. Section 6.2 talks about the configuration of core GA including how to initiate first generation, constraints, and operators. Section 6.3 is formal definition to most crucial part in GA, the fitness functions.

6.1 Introduction to Genetic Algorithm

Genetic algorithm is originally proposed by John H. Holland in 1975 [2]. It is most popular approach in evolutionary computation, and usually it is used for problems having enormous searching space to look for an optimal or near-optimal solution. It applies the notion from biology principle to simulate the process of evolution in nature, namely, "survival of the fittest", or from another point of view, "the natural selection". Fig. 3 is a graphical illustration for genetic algorithm and detailed textual description can be found in [2].

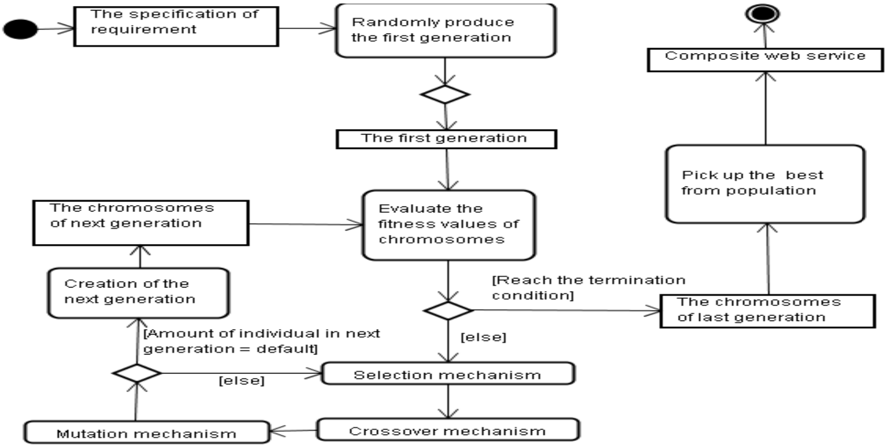

The specification of requirement
Randomly produce the first generation
Composite web service
The first generation
Pick up the best from population
The chromosomes of next generation
Evaluate the fitness values of chromosomes
Creation of the next generation
[Reach the termination condition]
The chromosomes of last generation
[Amount of individual in next generation = default]
[else]
[else]
Selection mechanism
Mutation mechanism
Crossover mechanism

Fig. 3. The process of genetic algorithm

6.2 Configuration of GA

Generally, the skeleton of GA is fixed, but each elaboration has diverse encoding style, initialization, operators, and fitness function. The encoding style of chromosome is tree structure transformed from original WS-BPEL. To WS-BPEL tree, there are limitations during the initialization and entire process of evolution presented in section 6.2.1. The working manner of each operator presents in section 6.2.2 including selection, crossover, and mutation.

6.2.1 Initialization and Limitation

The creation of chromosomes at first generation is entirely random, but there have some limitations applying to the chromosomes of every.

First limitation is the length of chromosome (i.e. the amount of services in a chromosome). We do not explicitly specify the length but stipulate a constraint, which is the length must more than one and less than the total amount of services in repository. The reason about why there is no explicit indication to the amount of component services is the size of each problem (user's requirements) are different, therefore, it is very difficult to foresee the amount of services that is required to deal with the problem.

Second limitation is the degree of tree. Within our setting, the deepest degree of tree is four in order to avoid the structure of CWS away from the reality (excessively nested structure). Because each tree already restrict its degree, a chromosome with longer length wills acquires relatively worse QoS level such as higher cost and longer response time and request more material (*I and P*). That leads the chromosome hard to survive in evolution. We anticipate using this implicit mechanism to automatically evolve appropriate length for chromosome. Consequently, the user does not have to judges the length of chromosome in advance that makes our approach as more flexible, powerful, and easy to use.

6.2.2 Operators

This subsection discusses selection, crossover, and mutation mechanism. There are various selection operators and we choose one of the most popular, binary tournament. The operator randomly picks up two chromosomes from current population and then chooses the winner from these two according to their fitness values. Winner will become one of parents. The characteristic of the binary tournament selection is that it retains both the randomness and the clarity at same time. After the selection operator works twice, there are two individuals (the parents). Next step is to crossover between the parents according to the crossover probability. If the two parents do not copulate decided by the crossover probability, the parents directly stream to next step. To do the crossover, the crossover operator randomly determine a node (may be a basic or structure activity) at level two of the parent trees, and then exchange the nodes. Fig. 4 is an illustration for crossover operators. The reason about why the chosen node is at level two is to prevent that the depth of tree deeper than four. Each chromosome from the result of crossover stage must consider about whether is needed to do mutate or not depend on the mutation probability.

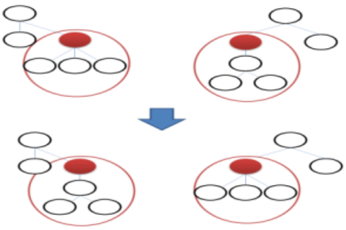


Fig. 4. The crossover operators

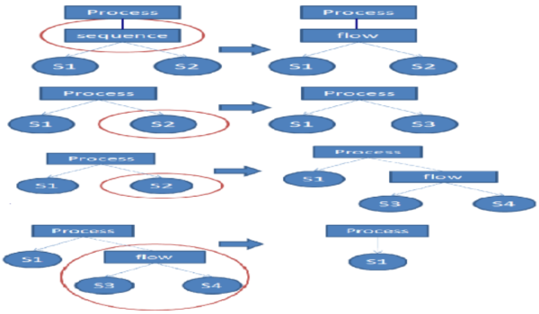


Fig. 5. The mutation operators

Commonly the crossover probability is much higher than the mutation probability. To do the mutation, the mutation operator determines a node at level two of tree and changes it as another activity randomly. There are four kind of possible mutating pattern: (1) a service to another service; (2) a service become a structure with services; (3) a structure to another structure; and (4) a service or structure disappears. Fig. 5 is an illustration to mutation operators.

6.3 Fitness Functions

In general, fitness function is most crucial element in GA. it greatly affects the performance of GA, even it is success or not. The difference between traditional and our GA is that there exist three fitness functions for evaluating chromosome. Each fitness value from one of the fitness functions assesses a chromosome from different facets and represents different meaning. Due to the heterogeneity between these fitness values, they do not be normalized as sole one but have priority order for them. The priority order is $F_1 > F_2 > F_3$ when the selection operator work. However, we will make an experiment, in which the fitness value is a summation of values from F_1 and F_2 and F_3, to confirm that the assumption of having priority order is correct.

Fitness Function 1 (F_1)**:** The value from this fitness function represents the similarity between the functionality of the solution $\langle ID, I, O, P, E\rangle_{provided}$ and $\langle ID, I, O, P, E\rangle_{required}$. Furthermore, the function assesses two factors: (1) the coverage about $\langle O, E\rangle_{provided}$ to $\langle O, E\rangle_{required}$, and (2) the usage about $\langle I, P\rangle_{provided}$ to $\langle I, P\rangle_{required}$. Because a CWS is consisted of component services, they must be synthesized as single specification before the fitness function work. The manner to synthesis is already mentioned in section 5.1.

$$S_I(\text{input similarity}) = \sum_{i=1}^{|I_R|} is\_Satisfied(I_{R_i}, I_p)$$

where

I_R is the set of required inputs from user,
$|I_R|$ is the number of elements of I_R,
I_{R_i} is ith element in I_R,
I_P is the set of needed inputs by CWS generated from our approach, and

$$is\_Satisfied(I_{R_i}, I_P) = \begin{cases} 0, & \text{iff } I_{R_i} \in I_P \\ -1, & \text{otherwise.} \end{cases}$$

S_O, S_P, S_E are calculated as same as S_I with replacing I to O, P, and E, respectively. Therefore, the fitness function 1 is defined as:

$$F_1 = (S_I + S_O + S_P + S_E)$$

Fitness Function 2 (F_2): The value acquired from this function exhibits the lack of material of the services among CWS. Initially there is an original value and then walk throughout the services which construct the chromosome, if a service requires materials (*I and P*) that cannot be fulfilled by currently available resources (user supply or generated from prior services), subtracts the value. The $\boldsymbol{F_2}$ is defined as below.

$$F_2 = \sum_{z=1}^{|CWS|} is\_Executable(Com_z, Obj_{resource}, Con_{resource})$$

where

$|CWS|$ is the number of component services in CWS,
Com_z is the *z*th component service in the CWS,
$Obj_{resource}$ is the set of currently available objects for the service, the initial elements are the same as I_R,
$Con_{resource}$ is the set of currently available conditions for the service, the initial elements are the same as P_R, and

$$is\_Executable(Com_z, Obj_{resource}, Con_{resource}) =$$

$$\begin{cases} 0\text{, iff } Com_{I_z} \subseteq Obj_{resource} \ \wedge \ Com_{P_z} \subseteq Con_{resource} \\ -\left(\left|Com_{I_z} - Obj_{resource}\right| + \left|Com_{P_z} - Con_{resource}\right|\right), \\ \text{otherwise.} \end{cases}$$

where

$Com_{I_z}$is the set of inputs from the zth component service in the CWS, and
Com_{P_z} is the set of pre-conditions from the zth component service in the CWS.

Fitness Function 3 (*F*$_3$): The value calculated by this function presents whether the TPs at CWS tree's level 2 belong to the set of legal TPs, which is a group including permitted TPs for assigned CTP and is inspired from the automation in [14]. How to synthesize the TPs under four structure activities is in section 5.3. Table 2 lists the elements in legal set for each requested CTP. The elements in braces are permitted TPs at first node of level 2 of tree.

Table 2. The legal TP set of requested CTP

Required TP	Legal Set
C	c, cr
$\overleftarrow{a}$	$\{p, \overleftarrow{\mathrm{a}}, \mathrm{c}, \mathrm{cr}\}$, pr, $\overleftarrow{a}r$, cr
cr	cr
$\overleftarrow{a}r$	pr, $\overleftarrow{a}r$, cr

$$F_3 = \sum_{i=1}^{|TPs|} is\_Belong\_To(\mathrm{TP_i}, \mathrm{LS})$$

where

$|TPs|$ is Transactional Properties at level 2 of CWS tree,
$\mathrm{TP_i}$ is the ith TP at level 2 of CWS tree,
LS is the set including legal TPs for requested CTP, and

$$is\_Belong\_To(\mathrm{TP_i}, \mathrm{LS}) = \begin{cases} 0 & \text{, iff } \mathrm{TP_i} \in LS \\ -(\text{amount of activities that construct } \mathrm{TP_i}), & \text{otherwise} \end{cases}$$

7 Experiments and Discussion

In order to evaluate the viability of our approach, experiments are conducted by implementing the program code of our approach on a PC Core 2 2.39 GHz with 0.98 GB RAM, Windows XP, and Java SE 6.0 platform. An primary API from one of Java API for XML Processing (JAXP) is Document Object Model (DOM) because it is able to bidirectional transforms XML document as in-memory tree structure (and vice versa). A set of services having 72 services is used in service repository, which totally has six kinds of functionalities listing in Table 3.

Table 3. Functionalities list of service repository

	S_1	S_2	S_3	S_4	S_5	S_6
I	O_1	O_2, O_3	O_5	O_4	O_6	O_9
O	O_2	O_4	O_6	O_8, O_9	O_9, O_{10}	O_{12}
P	C_1	C_1	C_1	C_1, C_2	C_1	C_1
E	C_2					

Each kind of functionality has four TPs combination (*p, c, pr, cr*). Thus, the number of services in repository is 6 x 4 = 24. The experiments are driven by several requirements. One of the requirements and the parameters and its corresponding answer CWS are shown in the follows. The user's requirement is $\langle\{requirement_1\}$, $\{O_1, O_3\}$, $\{O_{12}\}$, $\{C_1\}$, $\{C_2\}$, $\{atomic\}\rangle_{required}$. The amount of chromosomes is 500 and the number of generations is 350. The crossover and mutation probability are 0.91 and 0.085 respectively. Following is a corresponding solution CWS found in generation 46 for the requirement.

```
<Process F1="0" F2="0" F3="0">
  <flow>
  <invoke service="S1_pivot"/>
 </flow>
  <flow>
  <sequence>
   <invoke service="S2_pivot_retriable"/>
   <invoke service="S4_pivot_retriable"/>
   <invoke service="S6_pivot_retriable"/>
  </sequence>
 </flow>
</Process>
```

Fig. 6 shows the curve graphs displaying variety of chromosome's parameters along evolution. In Fig. 6, X axis means the generation number, and upper Y axis presents the average amount of services contained in the chromosome, and lower Y axis is the average fitness value of chromosomes. At first few generations, the selection operator works depend on F_1 and the services contained in chromosome are much more than the amount which is required to correctly solve this requirement (appropriate amount is four in this case), so the value from F_1 quickly closes to best (zero). However, the material from user is fixed and finite. A fact that the amount of services contained in chromosome is too much leads the value from F_2 and F_3 worse since a chromosome having more services asks for more material and including more component services with incorrect TPs. When F_1 value close to zero, the values from F_2 and F_3 are away from zero. After the value from F_1 reaches optimal (zero), the selection operator moves to F_2. In the term that the selection operator depends on F_2, the evolution eliminates the chromosomes with unneeded, plethoric services. The average amount of services in a chromosome descend also causes the value from F3 nearing perfect. When the amount of services in a chromosome decreasing, it is very difficult to observer that the value form F_1 slightly goes away from optimal because a CWS with less services is harder to cover the material and product from user requirement (the definition of F_1).

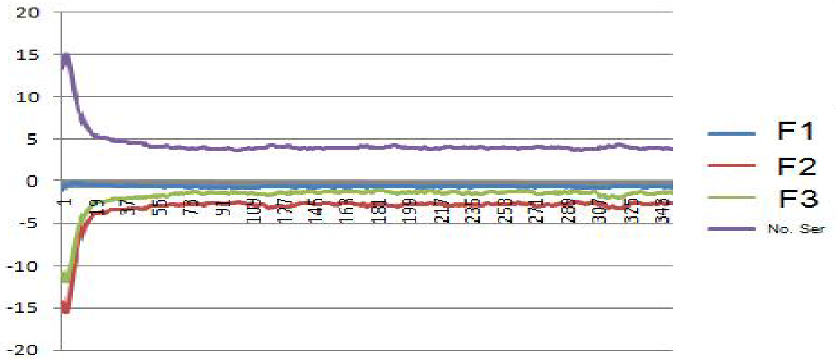


Fig. 6. The variety of average fitness values and amount of services during evolution

Another point is that the experiments based on a unique fitness value for each chromosome, which is the summation of all fitness values from three fitness functions, were done, and the consequence reveals that it is unable to find an appropriate answer. Indirectly, that was verified the necessity of priority order and the independence of three fitness functions.

Overall, the result shows that the evolution initially assures that the functionality of chromosomes is matched with requirement (F_1) and then finds out chromosomes they are executable by material from user (F_2). From these chromosomes, which are intended and workable, the evolution let chromosomes with wanted CTPs (F_3) survive.

8 Conclusion

In this paper, we have presented an automatic, design time service composition approach based on genetic algorithm. It does not need a handcrafted workflow template at design time, and it is transaction-aware during the composition, that assuring the reliability and consistence of resulting CWS. The resulting CWS will be stored into service repository as a component service for future use. Thus, a component service of CWS could be a CWS or pure WS. From the viewpoint of CWS designers, they are relaxed from the burden of design CWS concerning two facets (dynamic workflow, transaction-aware selection) simultaneously. In order to consider these facets, many things must be considered concurrently during the process of design time. In this paper, we would like to take care of and handle the monotonous things by the proposed approach rather than by human designer. Consequently, the problem which is faced by our approach is a multi-dimension problem having a huge searching space. We exploit genetic algorithm as core of the approach to do non-linear, jumping search in the vast space. In proposed GA, it has three fitness functions to score chromosomes from diverse aspects, and each fitness function has a priority order. The experiments driven by user's requirements indicate that the workability of proposed approach, also make sure the priority order and independence for fitness functions is vital. We also define the specification for represent service from different sides.

Acknowledgments. This research is partially sponsored by National Science Council (Taiwan) under the grant NSC99-2220-E-030-004 and by Fu Jen Catholic University (Taiwan) under the grant FJU 409931044037.

References

1. Gaaloul, W., Bhiri, S., Rouached, M.: Event-Based Design and Runtime Verification of Composite Service Transactional Behavior. IEEE Transactions on Services Computing 3, 32–45 (2010)
2. Aversano, L., Penta, M.D., Taneja, K.: A Genetic Programming Approach to Support the Design of Service Compositions. International Journal of Computer Systems Science & Engineering 21, 247–254 (2006)
3. Silva, E., Ferreira Pires, L., van Sinderen, M.: Supporting Dynamic Service Composition at Runtime Based on End-User Requirements. In: Dustdar, S., Hauswirth, M., Hierro, J.J., Soriano, J., Urmetzer, F., Möller, K., Rivera, I. (eds.) CEUR Workshop Proceedings of the 1st International Workshop on User-Generated Services, UGS 2009, located at the 7th International Conference on Service Oriented Computing, ICSOC 2009, Stockholm, Sweden (2009)
4. Lécué, F., Léger, A.: A Formal Model for Web Service Composition. In: Proceeding of the 2006 conference on Leading the Web in Concurrent Engineering: Next Generation Concurrent Engineering, pp. 37–46. IOS Press, Amsterdam (2006)
5. Fujii, K., Suda, T.: Semantics-Based Context-Aware Dynamic Service Composition. ACM Transactions on Autonomous and Adaptive Systems 4, 1–31 (2009)
6. Fujii, K., Suda, T.: Semantics-Based Dynamic Web Service Composition. International Journal of Cooperative Information Systems, 293–324 (2006)

7. Lajmi, S., Ghedira, C., Ghedira, K.: CBR Method for Web Service Composition. In: Damiani, E., Yetongnon, K., Chbeir, R., Dipanda, A. (eds.) SITIS 2006. LNCS, vol. 4879, pp. 314–326. Springer, Heidelberg (2009)
8. Soufiene, L., Chirine, G., Khaled, V., Djamal, B.: WeSCo_CBR: How to Compose Web Services via Case Based Reasoning. In: Proceedings of IEEE International Conference on e-Business Engineering (ICEBE 2006), pp. 618–622 (2006)
9. Bhiri, S., Perrin, O., Godart, C.: Extending Workflow Patterns with Transactional Dependencies to Define Reliable Composite Web Services. In: Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, p. 145. IEEE Computer Society Press, Los Alamitos (2006)
10. Bhiri, S., Perrin, O., Godart, C.: Ensuring Required Failure Atomicity of Composite Web Services. In: Proceedings of the 14th international conference on World Wide Web, pp. 138–147. ACM, Chiba (2005)
11. Li, L., Chengfei, L., Junhu, W.: Deriving Transactional Properties of CompositeWeb Services. In: Proceedings of the Conference Deriving Transactional Properties of CompositeWeb Services, pp. 631–638 (2007)
12. Portilla, A., Vargas-Solar, G., Collet, C., Zechinelli-Martini, J.-L., García-Bañuelos, L.: Contract Based Behavior Model for Services Coordination. In: Filipe, J., Cordeiro, J. (eds.) Web Information Systems and Technologies, vol. 8, pp. 109–123. Springer, Heidelberg (2008)
13. Montagut, F., Molva, R., Golega, S.T.: Automating the Composition of Transactional Web Services. International Journal of Web Services Research (IJWSR) 5(1), 24–41 (2008)
14. Hadad, E.J., Manouvrier, M., Rukoz, M.: TQoS: Transactional and QoS-Aware Selection Algorithm for Automatic Web Service Composition. IEEE Transactions on Services Computing 3, 73–85 (2010)
15. Alves, A.: OASIS Web Services Business Process Execution Language (WSBPEL) v2.0. OASIS Standard (2007)
16. Menascé, D.A., Casalicchio, E., Dubey, V.: A Heuristic Approach to Optimal Service Selection in Service Oriented Architectures. In: Proceedings of the 7th International Workshop on Software and Performance, pp. 13–24. ACM, Princeton (2008)
17. Menascé, D.A., Casalicchio, E., Dubey, V.: On Optimal Service Selection in Service Oriented Architectures. Performance Evaluation 67, 659–675 (2010)
18. Blake, M.B., Cummings, D.J.: Workflow Composition of Service Level Agreements. In: Proceedings of the IEEE International Conference on Services Computing, pp. 138–145 (2007)

Rapid Development of Executable Ontology for Financial Instruments and Trading Strategies

Dejan Lavbič and Marko Bajec

University of Ljubljana, Faculty of Computer and Information Science,
Tržaška cesta 25, 1000 Ljubljana, Slovenia
{Dejan.Lavbic,Marko.Bajec}@fri.uni-lj.si

Abstract. In this paper we employ Rapid Ontology Development approach (ROD) with constant evaluation of steps in the process of ontology construction for development of Financial Instruments and Trading Strategies (FITS) ontology. We show that ontology development process does not conclude with successful definition of schematic part of ontology but we continue with post development activities where additional axiomatic information and instances with dynamic imports from various sources are defined. The result is executable ontology as part of Semantic Web application that uses data from several semi structured sources. The overall process of construction is suitable for users without extensive technical and programming skills and those users are rather experts in the problem domain.

Keywords: ontology, semantic web, financial instruments, trading strategies, rapid ontology development.

1 Introduction

Semantic Web technologies are being adopted less than expected and are mainly limited to academic environment, while we are still waiting for greater adoption in industry. The reasons for this situation can be found in technologies itself and also in the development process, because existence of verified approaches is a good indicator of maturity. There are various technologies available that consider different aspects of Semantic Web, from languages for capturing the knowledge, persisting data, inferring new knowledge to querying for knowledge etc. Regarding the development process, there is also a great variety of methodologies for ontology development, as it will be further discussed in section 2, but simplicity of using approaches for ontology construction is another issue. Current approaches in ontology development are technically very demanding and require long learning curve and are therefore inappropriate for developers with little technical skills and knowledge. Besides simplification of the development process ontology completeness is also a very important aspect. In building ontology, majority of approaches focus on defining common understanding of a problem domain as a schematic model of the problem and conclude the development after few successful iterations. Post development

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 232–243, 2011.

activities that deal with defining instance data and employing developed ontology in Semantic Web application are usually omitted.

In this paper we apply Rapid Ontology Development (ROD) approach to construct Financial Instruments and Trading Strategies (FITS) ontology. The goal was to develop ontology by constructing schematic part of ontology including axiomatic information to fully support trading by employing reasoning. Furthermore this TBox part of ontology was combined to instance data (ABox) to construct knowledge base and therefore build mash up Semantic Web application to support financial instruments trading by applying various trading strategies. Target users of this approach are ones without extensive technical knowledge of data acquisition and ontology modeling but experts in financial trading. The main guideline in constructing ontology was to develop it to the level that enables direct employment in an application, which differs from majority of existing approaches where ontologies are mainly developed only to formally define the conceptualization of the problem domain.

The remainder of this paper is structured as follows. First we present some related work in section 2 with emphasis on ontology development methodologies and applications of financial ontologies. Next, in section 3, we introduce our approach for facilitating Semantic Web applications construction. The details of case study from the domain of financial instruments and trading strategies is further presented in section 4. First FITS ontology is presented, followed by semantic integration of data sources and then technological details about the prototype are depicted. Finally in section 5 conclusions with future work are given.

2 Related Work

Ontologies are used for various purposes such as natural language processing [1], knowledge management [2], information extraction [3], intelligent search engines [4], business process modeling [5] etc. While the use of ontologies was primarily in the domain of academia, situation now improves with the advent of several methodologies for ontology manipulation. Existing methodologies for ontology development in general try to define the activities for ontology management, activities for ontology development and support activities. Several methodologies exist for ontology manipulation and will be briefly presented in the following section. CommonKADS [6] is focused towards knowledge management in information systems and puts emphasis to early stages of software development for knowledge management. Enterprise Ontology [7] is groundwork for many other approaches and is also used in several ontology editors. METHONTOLOGY [8] enables building ontology at conceptual level and this approach is very close to prototyping. TOVE [9] is oriented towards using questionnaires that describe questions to which ontology should give answers. HCONE [10] is a decentralized approach to ontology development by introducing regions where ontology is saved during its lifecycle. OTK [11] defines details steps in two processes – Knowledge Meta Process and Knowledge Process. UPON [12] is based on Unified Software Development Process and is supported by UML language. DILIGENT [2] is focused on different approaches to distributed ontology development.

In the domain of finance several ontologies and implementations of Semantic Web based application exits. Finance ontology [13] follows ISO standards and covers several aspects (classification of financial instruments, currencies, markets, parties involved in financial transactions, countries etc.). Suggested Upper Merged Ontology (SUMO) [14] also includes a subset related to finance domain, which is richly axiomatized, not just taxonomic information but with terms formally defined. There are also several contributions in financial investments and trading systems [15-17]. Several authors deal with construction of expert and financial information systems [18-21].

3 Facilitating Semantic Web Applications Construction

3.1 Problem and Proposal for Solution

This paper describes semantic mash up application construction based on ontologies. The process is supported by continuous evaluation of ontology where developer is guided throughout the development process and constantly aided by recommendations to progress to next step and improve the quality of the final result. Our main objective is to combine dynamic (Web) data sources with a minimal effort required from the user. The results of this process are data sources that are later used together with ontology and rules to create a new application. This final result includes ontology that not only represents the common understanding of a problem domain but is also executable and directly used in the semantic mash up application.

Existing approaches for ontology development and semantic mash up application construction are complex and they require technical knowledge that business users and developers don't possess. As mentioned in section 2 vast majority of ontology development methodologies define a complex process that demands a long learning curve. The required technical knowledge is very high therefore making ontology development very difficult for non-technically oriented developers. Also majority of reviewed methodologies include a very limited evaluation support of developed ontologies and if this support exists it is limited to latter stages of development and not included throughout the process as is the case with our approach. Another problem that also exists is that the development process of ontology is completed after the first cycle and not much attention is given to applicability of ontology in an application.

3.2 Rapid Ontology Development

The process for ontology development ROD [22] that we follow in our approach is based on existing approaches and methodologies but is enhanced with continuous ontology evaluation throughout the complete process.

Developers start with capturing concepts, mutual relations and expressions based on concepts and relations. This task can include reusing elements from various resources or defining them from scratch. When the model is defined, schematic part of ontology has to be binded to existing instances of that vocabulary. This includes data from relational databases, text files, other ontologies etc. The last step in bringing ontology into use is creating functional component for employment in other systems.

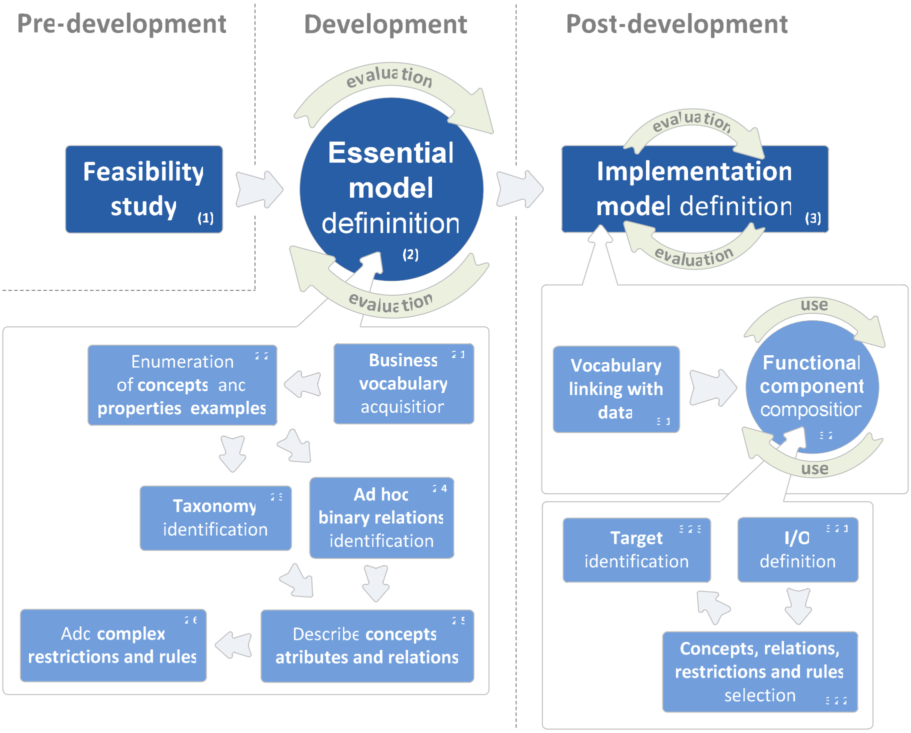
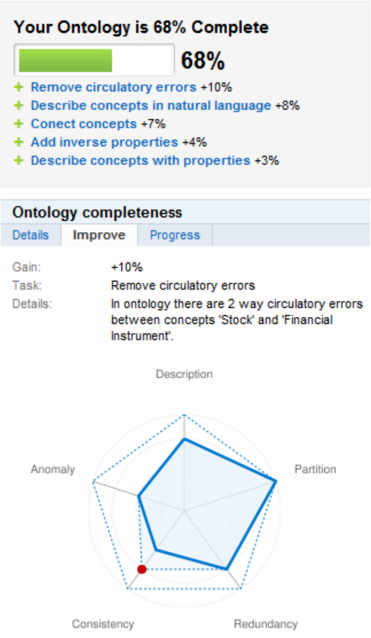

Pre-development
Development
Post-development
Feasibility study (1)
Essential model defininition (2)
evaluation
evaluation
Implementation model definition (3)
evaluation
evaluation
Enumeration of concepts anc properties examples 2 2
Business vocabulary acquisitior 2 1
Taxonomy identification 2 3
Ad hoc binary relations identification 2 4
Adc complex restrictions and rules 2 6
Describe concepts atributes and relations 2 5
Vocabulary linking with data 3 1
Functional component compositior 3 2
use
use
Target identification 3 2 3
I/O definition 3 2 1
Concepts, relations, restrictions and rules selection 3 2 2

Fig. 1. Process of Rapid Ontology Development (ROD)

The ROD development process can be divided into the following stages: *pre-development*, *development* and *post-development* depicted in Fig. 1. Every stage delivers a specific output with the common goal of creating functional component based on ontology that can be used in several systems and scenarios.

The role of constant evaluation as depicted in Fig. 1 is to guide developer in progressing through steps of ROD process or it can be used independently of ROD process. In latter case, based on semantic review of ontology, enhancements for ontology improvement are available to the developer in a form of multiple actions of improvement, sorted by their impact. Besides actions and their impacts, detail explanation of action is also available (see Fig. 2). When OC measurement reaches a threshold (e.g. 80%) developer can progress to the following step. The adapted OC value for every phase is calculated on-the-fly and whenever a threshold value is crossed, a recommendation for progressing to next step is generated. This way developer is aided in progressing through steps of ROD process from business vocabulary acquisition to functional component composition. Detail presentation of ontology completeness indicator is further presented in [22].



Your Ontology is 68% Complete

68%

- Remove circulatory errors +10%
- Describe concepts in natural language +8%
- Conect concepts +7%
- Add inverse properties +4%
- Describe concepts with properties +3%

Ontology completeness

Gain:	+10%
Task:	Remove circulatory errors
Details:	In ontology there are 2 way circulatory errors between concepts 'Stock' and 'Financial Instrument'.

Fig. 2. Ontology completeness and improvement recommendation

4 Case Study Implementation and Discussion

4.1 FITS Ontology

The problem domain presented in this paper is financial trading and analysis of financial instruments. As already discussed in related work section there are several financial instruments ontologies already present. The purpose of our work was to extend these approaches to the information system level, couple the ontology with reasoning capabilities, define inputs, outputs, dynamic imports and build fully executable Semantic Web solution for financial instruments analysis and trading strategies. For this purpose basic Financial Instruments (FI) ontology was developed following ROD approach (see Fig. 3). The FI ontology introduces basic concepts, including financial instrument, stock exchange market, trading day and analysis. Further details in form of taxonomy are provided for financial instruments, trading day and analysis.

While FI ontology defines elementary entities from financial trading domain, are ontologies that capture trading strategies more complex, including advanced axioms and rules. In our case we have define four different trading strategies: (1) simple

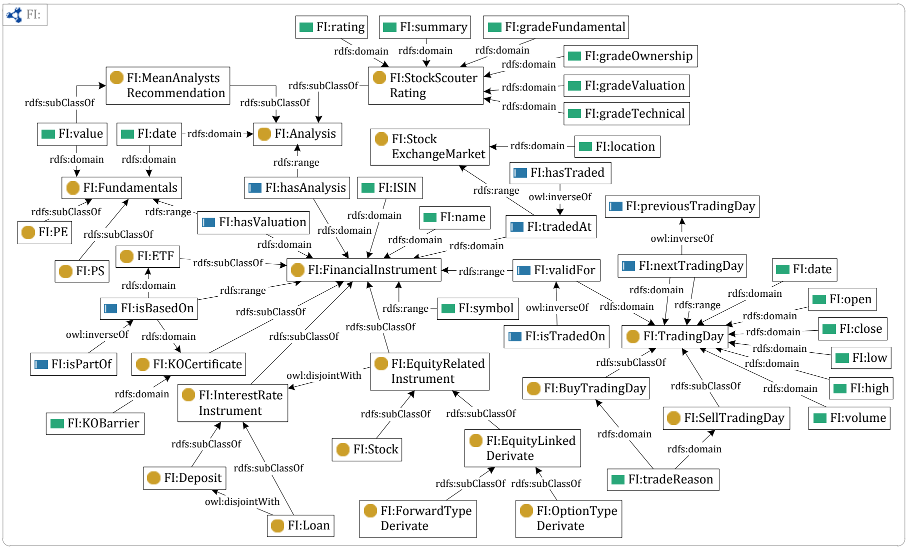


Fig. 3. Excerpt from FITS ontology

trading strategy (STs), (2) strategy of simple moving averages (SMAs), (3) Japanese candlestick trading strategy (JCTs) and (4) strategy based on fundamental analysis (FAs).

Every user has a possibility to define its own trading strategy whether from scratch or reusing existing ones. The main purpose of trading strategies is to examine the instances of *FI:TradingDay* concept and decide whether the instance can be classified into *FI:SellTradingDay* or *FI:BuyTraddingDay*. An example of this process can be found on Fig. 4 where and excerpt from JCTs is presented.

The JCTs is based on price movements which enable to identify patterns from daily trading formations. In this strategy price of a financial instrument is presented in a form of candlestick (low, open, close, high) and several patterns are identified (e.g. doji, hammer, three white soldiers, shooting star etc.). This strategy is rather complex but by following ROD approach (presented in section 3.2) domain experts can define it without being familiar with technical details of knowledge declaration and encoding.

After the selection of desired trading ontologies or composition of existing ones user can define the final ontology (see Fig. 5) which is than coupled with reasoning engine to allow the execution and performing trading analysis on real data available from several sources. At this point the schematic part of ontology (TBox component) is defined and further it still needs to be associated to instances (ABox component) by semantic integration of several data sources, which we will address in the following section 4.2.

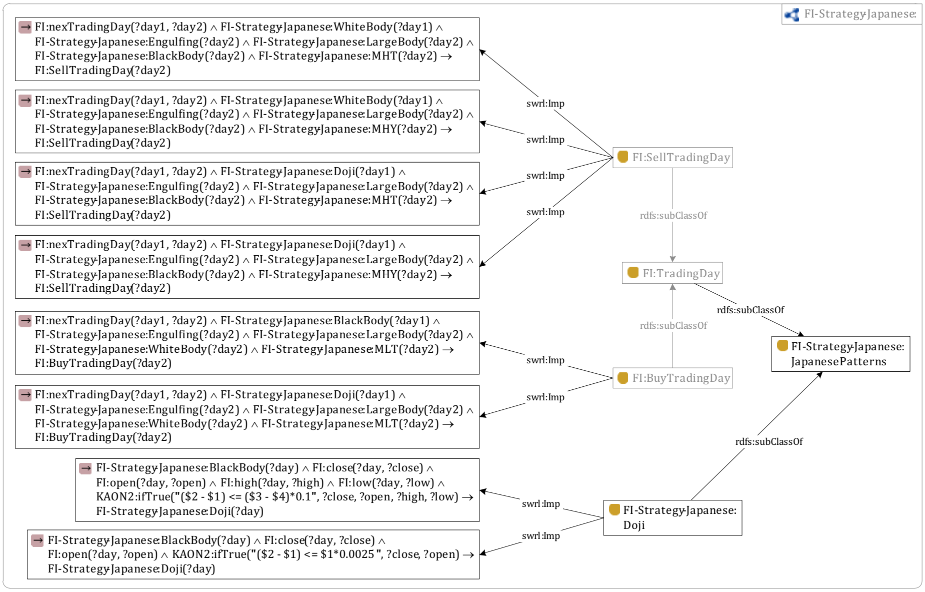

FI-StrategyJapanese:
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:WhiteBody(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:BlackBody(?day2) ∧ FI-StrategyJapanese:MHT(?day2) → FI:SellTradingDay(?day2)
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:WhiteBody(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:BlackBody(?day2) ∧ FI-StrategyJapanese:MHY(?day2) → FI:SellTradingDay(?day2)
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:Doji(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:BlackBody(?day2) ∧ FI-StrategyJapanese:MHT(?day2) → FI:SellTradingDay(?day2)
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:Doji(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:BlackBody(?day2) ∧ FI-StrategyJapanese:MHY(?day2) → FI:SellTradingDay(?day2)
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:BlackBody(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:WhiteBody(?day2) ∧ FI-StrategyJapanese:MLT(?day2) → FI:BuyTradingDay(?day2)
FI:nexTradingDay(?day1, ?day2) ∧ FI-StrategyJapanese:Doji(?day1) ∧ FI-StrategyJapanese:Engulfing(?day2) ∧ FI-StrategyJapanese:LargeBody(?day2) ∧ FI-StrategyJapanese:WhiteBody(?day2) ∧ FI-StrategyJapanese:MLT(?day2) → FI:BuyTradingDay(?day2)
FI-StrategyJapanese:BlackBody(?day) ∧ FI:close(?day, ?close) ∧ FI:open(?day, ?open) ∧ FI:high(?day, ?high) ∧ FI:low(?day, ?low) ∧ KAON2:ifTrue("($2 - $1) <= ($3 - $4)*0.1", ?close, ?open, ?high, ?low) → FI-StrategyJapanese:Doji(?day)
FI-StrategyJapanese:BlackBody(?day) ∧ FI:close(?day, ?close) ∧ FI:open(?day, ?open) ∧ KAON2:ifTrue("($2 - $1) <= $1*0.0025", ?close, ?open) → FI-StrategyJapanese:Doji(?day)
swrl:Imp
swrl:Imp
swrl:Imp
swrl:Imp
FI:SellTradingDay
rdfs:subClassOf
FI:TradingDay
rdfs:subClassOf
rdfs:subClassOf
FI-StrategyJapanese: JapanesePatterns
swrl:Imp
FI:BuyTradingDay
swrl:Imp
rdfs:subClassOf
swrl:Imp
FI-StrategyJapanese: Doji
swrl:Imp

Fig. 4. Excerpt from Japanese candlestick trading strategy

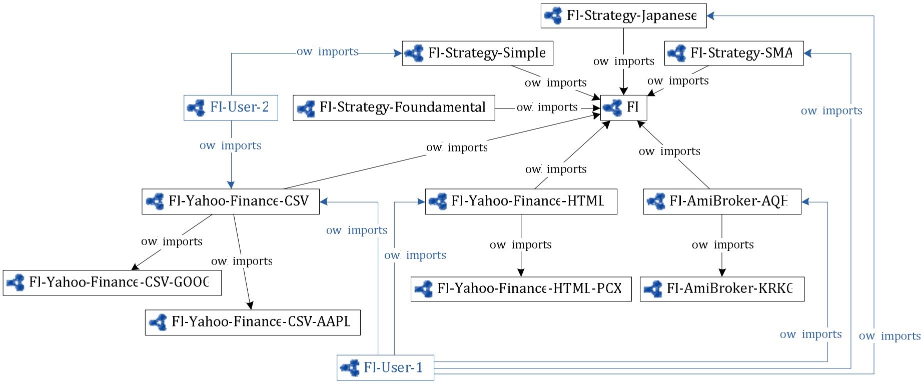
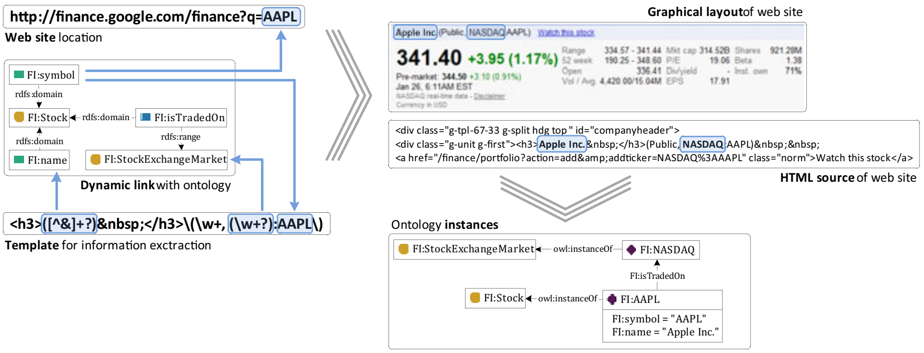


Fig. 5. Composition of final ontology for employment in Semantic Web application

4.2 Semantic Integration of Data Sources

In ROD approach there are several imports available: (1) existing ontologies, (2) relational or analytical databases, (3) CSV file and (4) semi structured data sources (e.g. HTML). In the process of creating FITS ontology the most prominent approach was reusing data from semi structured sources, mainly from HTML pages. When



```
http://finance.google.com/finance?q=AAPL
```

```
<h3>([^&]+?)&nbsp;</h3>\(\w+, (\w+?):AAPL\)
```

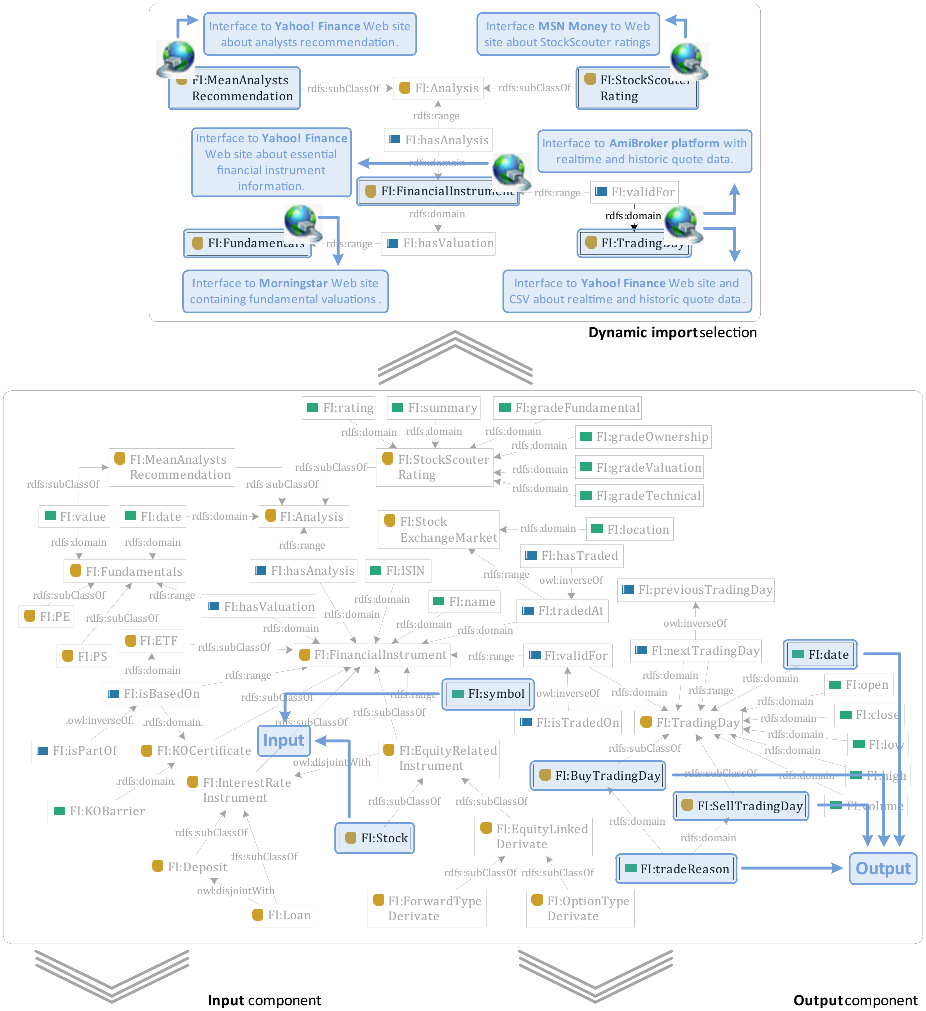
```
<div class="g-tpl-67-33 g-split hdg top " id="companyheader">
<div class="g-unit g-first"><h3>Apple Inc.&nbsp;</h3>(Public, NASDAQ:AAPL)&nbsp;&nbsp;
<a href="/finance/portfolio?action=add&amp;addticker=NASDAQ%3AAAAPL" class="norm">Watch this stock</a>
```

Fig. 6. Dynamic import of data property values related to financial instrument concept from Google Finance web data source

building executable ontology we relied on publicly available data about trading financial instrument, which are available on web pages and in vast majority in an unstructured form. Therefore linking wizard from ROD approach was used which incorporated the technology of regular expression and XQuery formulation for extracting data from semi structured data sources.

The role of semantic integration of data sources is to define wrapper to selected data sources and establish dynamic link between ontology entities (e.g. classes, properties etc.) and data source. An example of a simple web site wrapper is depicted in Fig. 6. This wrapper takes as an input financial instrument's symbol and uses Google Finance web page to extract information about financial instrument's name and stock exchange market where is being traded. As a result individuals are added or altered to the knowledge base with FITS ontology. These dynamic links can be defined for every selected entity as depicted in Fig. 7. For our case study there are 6 links defined. As analysis is concerned, mean analysts ratings are extracted from Yahoo! Finance web site, while stock scouter ratings are extracted from MSN money web site. All the essential data about the financial instruments are retrieved from Yahoo! Finance web site, while data about fundamental analysis are obtained from Morningstar web site. The quotes data are transferred from various sources, including historical data from Yahoo! Finance web site and real-time data from AmiBroker trading platform.

The last step in defining the Semantic Web application is to outline the input and the output component. The user can choose within the graphical interface which ontology entities will be used for input and which for output. In our case the input includes the symbol of stock that we want to trade and the outputs includes instances of trading days with buy or sell signals and trade reasons.



Input component

```
PREFIX FI: <http ://www .lavbic.net/SW/FI.owl>
INSERT DATA {
   ?FI FI:symbol ?symbol .
   ?FI rdf:type FI:Stock
}
```

Output component

```
PREFIX FI: <http://www.lavbic.net/SW/FI.owl>
SELECT ?symbol ?date ?type ?tradeReason {
   { ?day rdf:type FI:BuyTradingDay } UNION { ?day rdf:type FI:SellTradingDay } .
   ?day rdf:type ?type .
   ?day FI:date ?date .
   FILTER (?date >= ?dateFrom) .
   FILTER (?date <= ?dateTo) .
   ?day FI:tradeReason ?tradeReason .
   ?day FI:validFor ?instrument .
   ?instrument FI :symbol ?symbol
}
```

Fig. 7. Dynamic import selection with input and output definition

4.3 Technology

The selected language for ontology presentation is OWL DL, since it offers the highest level of semantic expressiveness for selected case study and is one of the most widely used and standardized language that has extensive support in different ontology manipulation tools. Besides OWL logical restrictions, Semantic Web Rule Language (SWRL) rules were also employed due to its human readable syntax and support for business rules oriented approach to knowledge management [23].

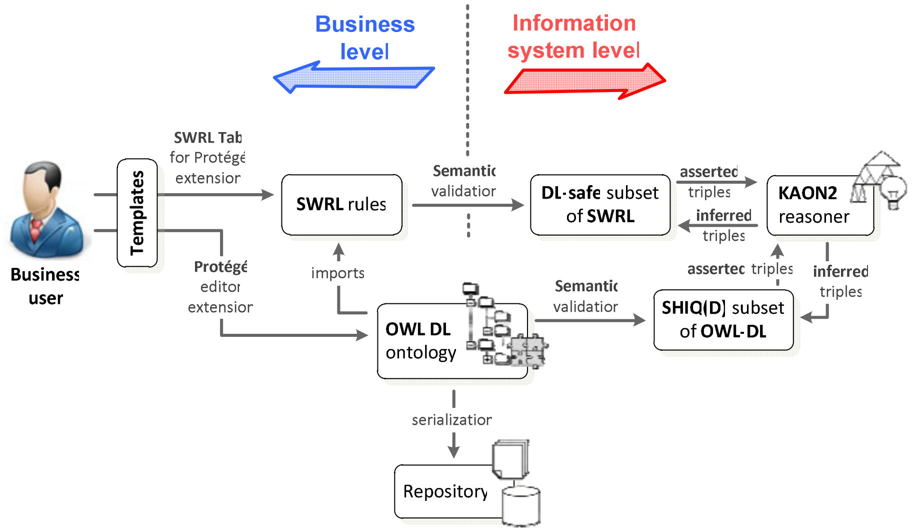


Fig. 8. Prototype of selected case study

The ontology manipulation interface for business users is based on Protégé Ontology Editor and Knowledge Acquisition System and SWRL Tab for Protégé. It enables entering OWL individuals and SWRL rules where a step further is made

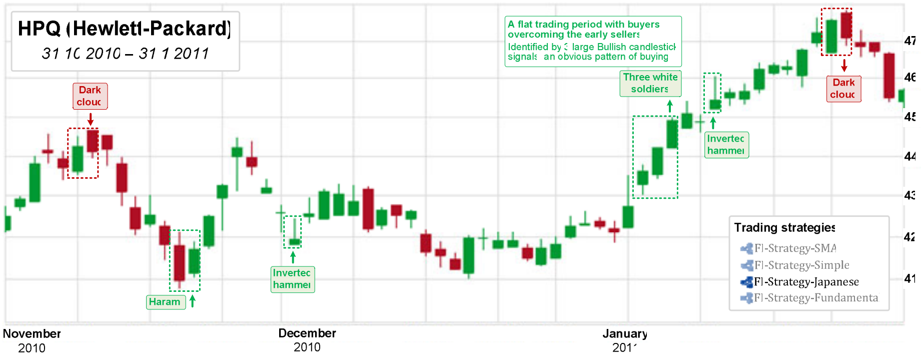


Fig. 9. GUI example of a Japanese trading strategy analysis on HPQ stock in the period from November 2010 to February 2011

towards using templates for entering information (see Fig. 8). At the information system level KAON2 inference engine is used to enable inference capabilities. Due to limitations of SHIQ(D) subset of OWL-DL and DL-safe subset of SWRL language, before inference is conducted, semantic validation takes place to ensure that all preconditions are met.

Fig. 9 depicts an example of firing trading rules on a real case scenario. The selected quote is HPQ (Hewlett-Packard) in the trading period of 3 months where several trading rules from Japanese trading strategy are being fired. From the GUI user can always select which subset of trading strategies is used (see section 4.1) and get details about the pattern found.

5 Conclusions and Future Work

Current approaches for ontology development require very experienced users and developers, while using the ROD approach for constructing FITS ontology is more appropriate for less technically oriented users. With constant evaluation of developed ontology that ROD approach offers, developers get a tool for construction of ontologies with several advantages: (a) the required knowledge for ontology modeling is decreased, (b) the process of ontology modeling doesn't end with the last successful iteration, but continues with post development activities of using ontology in a Semantic Web application and (c) continuous evaluation of developing ontology and recommendations for improvement. It has been demonstrated on a case study from financial trading domain that a developer can build Semantic Web application for financial trading based on ontologies that consumes data from various sources and enable interoperability. The solution can also be easily packed into a functional component and used in various systems. The results from using ROD approach is that the resulting artifact is executable ontology that is available in open format (e.g. OWL and SWRL language) and available for further inclusion. When reusing and building additional applications users have free selection of inference engines and also ontology manipulation tools. Added value is also defined in dynamic imports of data (instances in knowledge base) that can be acquired also at the runtime level.

The future work includes improvement of developed ontology and combining it with other approaches that mainly focus on schematic part of ontology and extend the possible use cases. One of the planned improvements is also integration with popular social networks to enable developers rapid ontology development based on reuse and therefore employ the community effort in curation process.

References

1. Staab, S., Braun, C., Bruder, I., Duesterhoeft, A., Heuer, A., Klettke, M., Neumann, G., Prager, B., Pretzel, J., Schnurr, H.P., Studer, R., Uszkoreit, H., Wrenger, B.: A system for facilitating and enhancing web search. In: Staab, S., Braun, C., Bruder, I., Duesterhoeft, A., Heuer, A., Klettke, M., Neumann, G., Prager, B., Pretzel, J., Schnurr, H.P., Studer, R., Uszkoreit, H., Wrenger, B. (eds.) International Working Conference on Artificial and Natural Neural Networks: Engineering Applications of Bio-Inspired Artificial Neural Networks, IWANN 1999 (1999)
2. Davies, J., Studer, R., Warren, P.: Semantic Web technologies - trends and research in ontology-based systems. John Wiley & Sons, Chichester (2006)

3. Wiederhold, G.: Mediators in the architecture of future information systems. IEEE Computer 25, 38–49 (1992)
4. Heflin, J., Hendler, J.: Searching the web with SHOE. In: Artificial Intelligence for Web Search, pp. 36–40. AAAI Press, Menlo Park (2000)
5. Ciuksys, D., Caplinskas, A.: Reusing ontological knowledge about business process in IS engineering: process configuration problem. Informatica 18, 585–602 (2007)
6. Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., van de Velde, W., Wielinga, B.: Knowledge engineering and management - The CommonKADS methodology. The MIT Press, Cambridge (1999)
7. Uschold, M., King, M.: Towards a methodology for building ontologies. In: Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI 1995), Montreal, Canada (1995)
8. Fernandez-Lopez, M., Gomez-Perez, A., Sierra, J.P., Sierra, A.P.: Building a chemical ontology using methontology and the ontology design environment. Intelligent Systems 14 (1999)
9. Uschold, M., Grueninger, M.: Ontologies: principles, methods and applications. Knowledge Sharing and Review 11 (1996)
10. Kotis, K., Vouros, G.: Human centered ontology management with HCONE. In: IJCAI 2003 Workshop on Ontologies and Distributed Systems (2003)
11. Sure, Y.: Methodology, Tools & Case Studies for Ontology based Knowledge Management. In: Institute AIFB, vol. 332, PhD. University of Karlsruhe (2003)
12. Nicola, A.D., Navigli, R., Missikoff, M.: Building an eProcurement ontology with UPON methodology. In: 15th e-Challenges Conference, Ljubljana, Slovenia (2005)
13. Vanderlinden, E.: Finance ontology (2011)
14. Farrar, S., Lewis, W., Langendoen, T.: A Common ontology for linguistic concepts. In: Knowledge Technologies Conference, USA, Seattle (2002)
15. Zhang, Z., Zhang, C., Ong, S.S.: Building an Ontology for Financial Investment. In: Leung, K.-S., Chan, L., Meng, H. (eds.) IDEAL 2000. LNCS, vol. 1983, pp. 308–313. Springer, Heidelberg (2000)
16. Mellouli, S., Bouslama, F., Akande, A.: An ontology for representing financial headline news. Journal of Web Semantics 8, 203–208 (2010)
17. Qin, H., Taffet, M.D.: Vocabulary use in XML standards in the financial market domain. Knowledge and Information Systems 6, 269–289 (2004)
18. Chen, Y., Zhou, L., Zhang, D.S.: Ontology-supported web service composition: An approach to service-oriented knowledge management in corporate financial services. Journal of Database Management 17, 67–84 (2006)
19. Cheng, H., Lu, Y.C., Sheu, C.: An ontology-based business intelligence application in a financial knowledge management system. Expert Systems with Applications 36, 3614–3622 (2009)
20. Castells, P., Foncillas, B., Lara, R., Rico, M., Alonso, J.L.: Semantic web technologies for economic and financial information management. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 473–487. Springer, Heidelberg (2004)
21. Ying, W., Sujanani, A., Ray, P., Paramesh, N., Lee, D., Bhar, R.: Design and Development of Financial applications using ontology-based Multi-Agent Systems. Computing and Informatics 28, 635–654 (2009)
22. Lavbič, D., Krisper, M.: Facilitating Ontology development with continuous evaluation. Informatica 21, 533–552 (2010)
23. Horrocks, I., Patel-Schneider, P.F., Bechhofer, S., Tsarkov, D.: OWL rules: A proposal and prototype implementation. Journal of Web Semantics 3, 23–40 (2005)

A Greedy Approach for Adapting Web Content for Mobile Devices

Rajibul Anam, Chin Kuan Ho, and Tek Yong Lim

Faculty of Information Technology, Multimedia University, Persiaran Multimedia,
63100 Cyberjaya, Selangor, Malaysia
rajibul.anam08@mmu.edu.my, ckho@mmu.edu.my, tylim@mmu.edu.my

Abstract. Mobile internet browsing usually involves a lot of horizontal and vertical scrolling, which makes web browsing time-consuming and in addition to this user may be interested in a section of a webpage, which may not fit to the mobile screen. This requires more scrolling in both dimensions. In this paper, we propose to address this problem by re-arranging the geometric sequence of the blocks from a large webpage while maintaining their semantics. Our proposed system, Web-adaptor, reduces unnecessary information by allowing its users to see the most relevant blocks of the page and provides the target contents. The Web-adaptor assigns profit to each object of the webpage according to the user preferences. It also assigns a weight to each object of the block by analyzing the object's elements. It uses greedy algorithm to select the profitable blocks, and delivers them to handheld devices. The proposed solution improves web content accessibility and delivers the target contents to the users.

Keywords: Content adaptation, Web-adaptor, Mobile browsing, Small display, Adaptive interface for small screens.

1 Introduction

Last few years, people start using network ready mobile devices like handheld computers, PDAs and smart phones to access internet. The term mobile device refers to a device specially designed for synchronous and asynchronous communication while the user is on the move [3]. Among the mobile devices, the mobile phone and PDA are the most popular and commonly used by the users and one of the facilities is access internet [3]. These kinds of devices provide good mobility but very limited computational capabilities and display size [7]. Mobile users don't feel comfortable to browse internet via mobile devices, because of small screen, limited memory, processing speed and slow network [1][7]. Since most of the existing Web contents are originally designed for display in desktop computers, so the content delivery without layout adjustment and content adaptation make the contents unsuitable for mobile devices. Users mobile devices need to scroll the screen horizontally and vertically to find the desired content. Moreover searching and browsing could frustrate, because most of the web site is designed for the standard desktop display. This means that most of the web contents are not suitable for mobile devices [4].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 244–258, 2011.

Content adaptation refers to techniques which dynamically adjust the contents in the direction of represent the contents according to the properties of the handheld devices for better presentation. The conventional way to provide Web contents to support various types of handheld devices is to create the same contents but different formats for different devices. This method is simple but the chances of making errors are very high with different handheld devices. To support new handheld devices, all the previous Web contents have to be reformatted for that handheld device. Sometimes, changes in the main contents require updating of all handheld contents format. So, this is neither a practical nor feasible solution for large volume of Web contents.

Since the screen size is limited, the content adaptation method needs to apply various content transformation processes including layout changes, content format reconfiguration and rearrange the context presentation [4][6][7]. However, simple content adaptation solution for changing multiple-column layout to a single-column layout for handheld screens also shows some disadvantages. Without the semantic analysis and relationship among the semantic objects, this kind of adaptation may cause an awkward organization of a webpage and gives different meaning to information. A tool or mechanism is needed to provide users opportunity to flexible Web contents on handheld devices.

Web contents are typically composed of many multimedia objects (text, images, audio and video) [12], which are semantically connected by various objects in a section. For example, an image can be illustrated by a section of a text article or a text title can abstract an article and some images. In other words, these related objects are integrated to help readers to understand what the sections intend to express. Improper arrangement of these kinds of objects and their relationship may lead to the misunderstanding or loss of information to the users. Therefore, it is very important for a content adaptation mechanism to maintain the original semantic relationship among the objects during the adaptation and content delivery process.

In this paper, we present a novel method which support dynamic webpage content adaptation for handheld devices. Our goal is to improve Web content accessibility, deliver the target information and multimedia contents according to the user preferences. This will help the users to reach the target information from a large webpage. To achieve this goal, we introduce Web-adaptor (Fig.2) as an automatic content adaptation system for dynamic webpages. Our algorithm automatically identifies the semantic relationship among the objects from the blocks, assigns profit to the objects from the user define preferences and weight from the object elements. The greedy algorithm selects the profitable blocks and delivers them to the users. The major contributions of this paper are as follows: Firstly, assign profit to each object of the html structure by using assignment profit algorithm. Secondly, re-arrange the sequence of the blocks by using greedy algorithm. The rest of the paper is organized as follow: Section 2, summarizes the related research work. Describe the proposed solution in section 3. Section 4, discussion and result about the system. Finally, we conclude in section 5.

We would like to draw your attention to the fact that it is not possible to modify a paper in any way, once it has been published. This applies to both the printed book and the online version of the publication. Every detail, including the order of the names of the authors, should be checked before the paper is sent to the Volume Editors.

2 Related Work

There are general purposes to content adaptation systems that have been developed. The CMo uses proxy-based architecture to adapt web contents for handheld devices. This system reduces the information to overload by allowing its users to see the most relevant fragment of the pages and navigate between other fragments if necessary. It captures the context of the links and applies a simple topic boundary detection technique and also uses the context to identify the relevant information in the next page with the help of Vector Machine [5].

Web Clipping is one of the methods that researchers are still working on. Web Clipping is a technique where the system extracts and represents some part of HTML document for mobile browser [8]. An annotation or parsers are general declares the properties that qualify a particular portion of a target document. The system annotates some parts of the webpage and it provides the annotation contents to the content adaptation engine. The Web Clipping methods modify the HTML structures. This system breaks the page into small parts, makes them a new separate page and adds title and header. Sometimes it removes extra unnecessary objects from the HTML. The concept of this method is to read the webpage, tag some parts of the page and regenerate the webpage for the mobile browser.

Xadaptor is a content adaptation system that consists of rule-based and fuzzy logic approach. The rule-based approach facilitates extensible, systematic and adaptive content adaptation and also integrates adaptation mechanisms for various contents types and categorizes them into rule-based approach. Rules are applied by the user define preferences. The HTML objects are transformed into content and pointer objects. Therefore the system uses content parser to separate the objects from the HTML tags and the system reformats the standard tables from the HTML objects. The table reformatting algorithm uses fuzzy logic and rule-based approaches. This system is not fully automated. The users need to assign some parameters to adapt the contents [2].

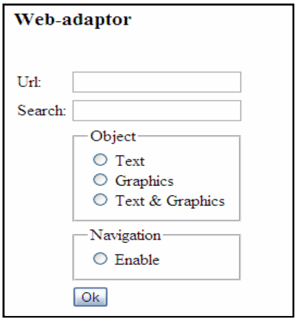
Some researchers use Vision Based Page Segmentation (VIPS) algorithm [1][4-6][9] that manages the web structure to find the interesting objects and restructure the web pages as a block. The Web page Tailoring System follows some rules to select the interesting blocks [9]. The system removes unnecessary information, creates a new title for the block and tries to summarize the block contents information. The VIPS identifies the interesting data, change the format of the webpage and also presents the information to the user with user's interest. User preferences are used by Page Segmentation and Pattern Matching to make the information interesting for every user.

Researches use Web Clipping, rule-based and VIPS to identify the objects and adapt the contents. However, users still need to scroll vertically. Sometimes users browse the contents but fail to reach the target contents because of overload information [13]. Still, it is a great challenge to achieve satisfactory precision for dynamic webpage segmentation which is based on HTML elements analysis.

3 Proposed Solution

This section presents the framework of the Web-adaptor (Fig.2). The architecture of Web-adaptor is based on proxy server, the users request for a webpage and Web-adaptor adapt the contents and delivers. We choose proxy server, because it acts as a cache of the mobile device and also process the data for the mobile devices. First, we describe constructing the tree for caching functionalities to improve our system performance and identification of important blocks. Second, assign profit to each objects of a block. Third, pre-process each object elements. Fourth, assign weight to each objects of a block. Finally, use greedy algorithm to select blocks to display on mobile.

A mobile user connects to the Web-adaptor by a web browser. The user types the URL of a webpage in the input box (Fig.1). Afterwards choose their preferences like search any particular info in the page or multimedia contents or navigations. After selecting the preferences, the users submit the request to the Web-adaptor and it delivers the subpages to the users. All the subpage contents are fit for the mobile screen. The users do not need to scroll. Users press next and previous buttons to navigate the subpages.



Web-adaptor

Url:

Search:

Object

○ Text

○ Graphics

○ Text & Graphics

Navigation

○ Enable

Ok

Fig. 1. Home page of the Web-adaptor with user preferences

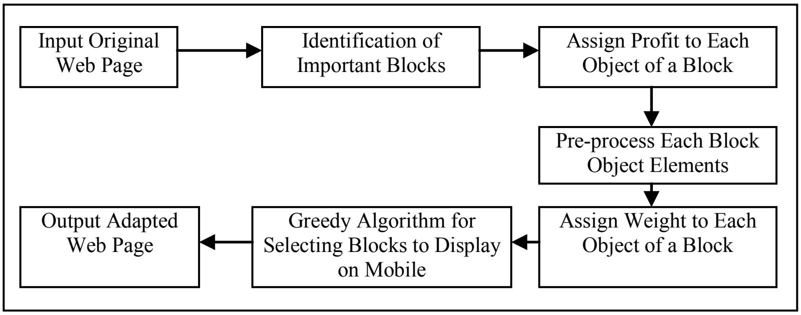
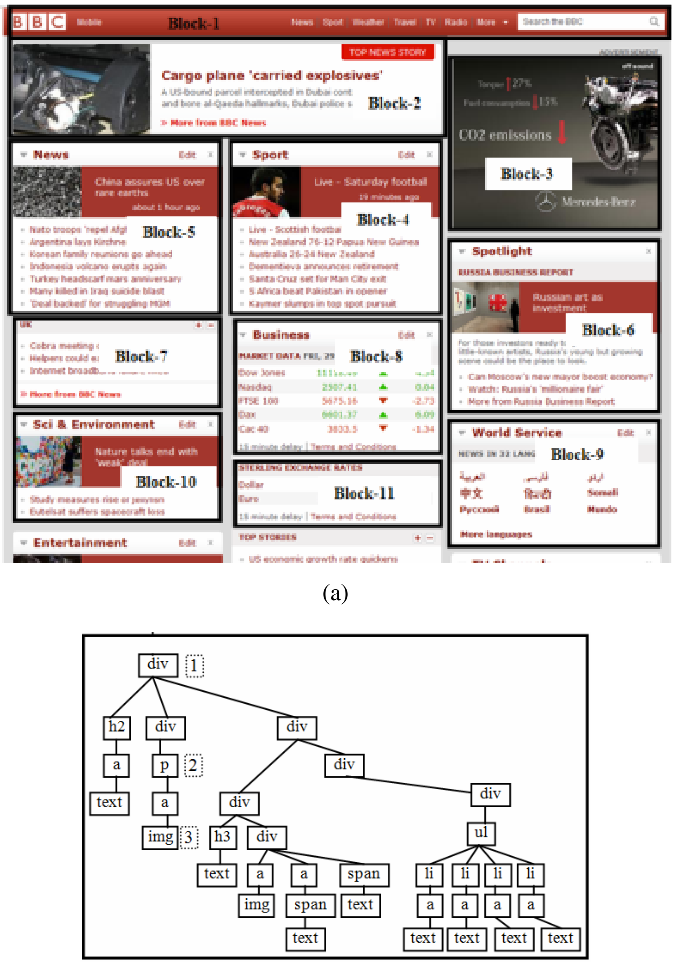

Input Original Web Page
Identification of Important Blocks
Assign Profit to Each Object of a Block
Pre-process Each Block Object Elements
Assign Weight to Each Object of a Block
Greedy Algorithm for Selecting Blocks to Display on Mobile
Output Adapted Web Page

Fig. 2. The framework of the Web-adaptor

3.1 Identification of Important Blocks

The Web-adaptor first transfers the HTML page to a tree. The root node is the top element of the tree. The root node has no parents. The internal node means node with one or more children [10]. Fig.3.(b), illustrates the tree presentation of the block-5 (Fig.3.a). This tree contains all the context of the block-5 HTML elements. The root of the tree is <div> (1), internal node is <p> (2) and leaf node is <img> (3).



(a)

(b)

Fig. 3. (a) Original BBC webpage categories as semantic blocks and (b) HTML Tree hierarchy presentation of Block-5

The system identifies the blocks according to the semantic relationship among the HTML elements. Since webpages are always presented in HTML format, the system traverses the tree and identifies the blocks in the tree according to the HTML elements, analyze their properties and tag the blocks. The system avoids objects like Java Script and CSS. These objects are not important for mobile view [3]. The system uses the Depth First Search to traverse the tree, it remove nodes contain elements like <script>, <style> to simplify the tree [3]. The semantic blocks are identifies by using structure functionality [7], which arrange the layout of information objects [13]. Each semantic block is an individual block content unit and it contains group of individual content unit. Fig. 3. (a), illustrates a typical page from "BBC", block 1 is the menu, Block-2 is the top news, Block- 4,5,7,8,10,11 are the relevant interesting blocks. Block- 6 and 9 contains some related links and Block-3 is an advertising Block.

Fig. 4. Illustrates the algorithm for identifying the blocks. The input is the original HTML tree T (line 2), output tree T contains only the important blocks (line 4), *v* and *w* is the node of the tree and *structure functionality* [7] is the set of HTML tags which points to the HTML blocks. The algorithm traverses the tree and searches for the *structure functionality* (line 8). If it matches, then it keeps the internal nodes and leaf nodes, if it does not match, it removes the internal nodes and leaf nodes from the tree. This algorithm keeps only the important block contents in the tree and removes other contents.

```
1.   Input:
2.     T is the Tree;
3.   Output:
4.     T is the Modified Tree;
5.   Start
6.     Visit(T,v)
7.       Perform visit of the node v;
8.       If v match with structure functionality Then
9.         If v is an internal node Then
10.          For all child w of v Do
11.            Visit(T,w);
12.            If w do not match with structure functionality Then
13.                Remove all the children of w;
14.      Else
15.        Remove all the children of v;
16.    End
```

Fig. 4. Algorithm for Identifying the Important Blocks

3.2 Assignment of Profit to Each Object of a Block

The system searches the individual content units from the semantic block (Fig.5) and sends it for different assignments. The profits are positive numeric values. The profit is assigned according to the tree element properties of the individual content unit. The assigned profit algorithm (Fig.6) uses users define preferences like User Parameter

Object, Decoration Object, Hyperlink Object and Multimedia Object [11]. With the user define preferences, algorithm uses rules to assign profits to the nodes of the tree.

The search keyword option (Fig.1) gives the users more specific target information. If the user keys in any prefer data, the system checks the keyword the user prefers to find and starts searching the keyword from the root to leaf node HTML elements of the tree. Fig.6.(line 7-8) algorithm shows, if it finds any keyword matches with the user prefer data, assign profits for each match data to the individual content unit.

```
Input:
  T is the Tree;
Output:
  T is the Modified Tree;
Start
  Traverse(T,v)
    Perform visit of the node v;
    AssignmentOfProfit(v);
    Preprocessing(v);
    AssignmentOfWeight(v);
    If v is an internal node Then
        For all child w of v do
            Traverse(T,w);
End
```

Fig. 5. Algorithm to identify individual content unit from Blocks

```
1.  Input:
2.    v is the node of the tree;
3.  Output:
4.    v with profit points;
5.  Start
6.    AssignmentOfProfit (v)
7.    If search keyword matches with v and preference enable Then
8.      Assign profit to $v_p = v_p + ( 500 \times \textit{search keyword} \bullet v \text{ set data})$ ;
9.    If decoration object matches with v and preference enable Then
10.     Assign profit to $v_p = v_p + 1$ ;
11.   If hyperlink object matches with v and preference enable Then
12.     If hyperlink is related to the other objects Then
13.       Assign profit to $v_p = v_p + 4 + \textit{text length}$;
14.     If hyperlink is not related to the other object Then
15.       Assign profit to $v_p = v_p + 1 + \textit{text length}$;
16.   If multimedia object matches with v data and preference enable and multimedia
      object pixels > 21 Then
17.     Assign profit to $v_p = v_p + ( \textit{object height} \times \textit{object width} / 300)$ ;
18. End
```

Fig. 6. Algorithm for assignment of profit to the individual content units

Some HTML elements use for formatting [7] and decoration [11] purposes. All these kinds of HTML elements are used to display the HTML contents in a more interesting manner to the readers and contain important information. If the users select text option (Fig.1), system searches these HTML elements inside the tree and assigns profit to the individual content unit. Fig.6.(line 9-10), illustrates that algorithm searches all the decoration and formatting elements and assigns profit to the individual content unit for each matching elements.

Most webpages use hyperlinks to navigate to other sections or another portal. If the users select navigation option (Fig.1), the system assigns profit according to the Hyperlinks and text lengths. There are two kinds of Hyperlinks, Independent and Dependent Hyperlink. The Independent Hyperlink means the one which navigates to another domain or has weak relationship with the objects around it [11]. These kinds of HTML elements are not so important for the readers. Fig.6.(line 14-15) algorithm illustrates for all kinds of Independent Hyperlinks and assigns low profit to the individual content unit. The Dependent Hyperlink means the one which navigates to the same domain and has strong relationship with the objects around it [11]. These kinds of navigation consider important for the readers. Fig.6.(line 12-13) algorithm shows for all kinds of Independent Hyperlinks and assigns high profit to the individual content unit. At the end, it counts the Hyperlink text lengths and assigns the text lengths as profit of the individual content unit.

Multimedia contents are very important to deliver information. If the users select multimedia option (Fig.1), system assigns profit to the multimedia content dimensions. Fig.6.(line 16-17) algorithm shows, when the multimedia contents dimensions are small, the profit gets small but profit points are high, if the pixel dimensions are higher. Moreover, when the image width is less than 21 pixels, it ignores the image to assign profit because, less than 21 pixels either use as a symbol of navigation or fill the blank space in the webpage. When the image width is more than 21 pixels, it means that the Multimedia content is important for the users and needs to assign profit.

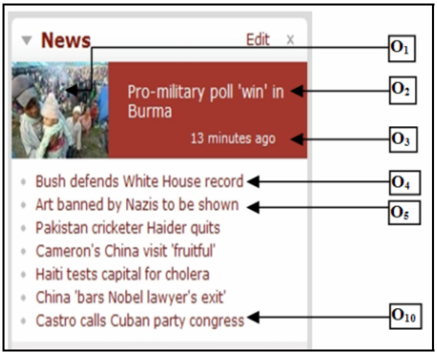

News
Edit x
Pro-military poll 'win' in Burma
13 minutes ago
- Bush defends White House record
- Art banned by Nazis to be shown
- Pakistan cricketer Haider quits
- Cameron's China visit 'fruitful'
- Haiti tests capital for cholera
- China 'bars Nobel lawyer's exit'
- Castro calls Cuban party congress

Fig. 7. The Geometric order of the objects from Block-5

After completing all the profit assignment processes, the system adds the total profits of all individual content units of a block to group profit and reassigns the profit of individual content units with this new value of group profit. So, all the individual content units of the block will carry the same profits. The profit calculation can be changed or updated just by adding or editing some rules.

The Geometric order of the webpage means the sequence of contents arrangement or display in the webpage [5]. The Geometric order is very important to keep the original semantic relationship among the individual content units of the semantic block. Therefore the system assigns profit according to the Geometric order. Fig.7. illustrates geometric order of the objects in the webpage where O_1 gets the high profit, because it appears at the beginning of the semantic block. O_{10} gets the lowest profit because it appears at the end of the semantic block. The system searches all the blocks and assigns profits to the individual content units by geometric order of the HTML structure.

3.3 Pre-processing of Blocks

Some HTML elements are used to format the data to highlight important information as title [7] and decoration [7] makes attractive to the readers. For mobile readers this kind of extra formatted data presentation is not necessary. Therefore, it is better to remove and edit the HTML elements. Fig.8.(line 7), algorithm illustrates *Iscreen* control the multimedia contents dimension. *Iscreen* contains the ratio of the multimedia contents which is suitable for the mobile screen. The algorithm (line 9) first, removes all the HTML decoration elements, changes the background color of the elements property. Second, it edits the multimedia content dimension properties, if the multimedia contents dimension is more than *Iscreen* (line 10-20). This algorithm modifies the original dimensions of the multimedia contents and reduces the dimensions which suits in the mobile screen.

```
1.  Input:
2.    v is the node of the Tree;
3.  Output:
4.    v with modified node of the Tree;
5.  Start
6.    Preprocessing(v)
7.      Iscreen =( DeviceWidth / DeviceHeight) × 200;
8.      If v is decoration or highlight element Then
9.        Remove the element from v and change the background color properties;
10.     If v is multimedia element and multimedia height, width more than Iscreen Then
11.       MaxHeight = Iscreen;
12.       MaxWidth = Iscreen;
13.       Ratio = height / width;
14.       If height > MaxHeight Then
15.         newheight = MaxHeight;
16.         newwidth = height / Ratio;
17.       Elseif width > MaxWidth Then
18.         newwidth = MaxWidth;
19.         newheight = width × Ratio;
20.     update v with the new width and height;
21.   End
```

Fig. 8. Algorithm to Pre-processing each object elements

3.4 Assignment of Weight to Each Object in a Block

The Web-adaptor assigns weights according to text length and multimedia objects dimension of the each individual content unit from the tree. The greedy algorithm uses the individual content unit weights to deliver limited contents to every subpage. The system considers Text and Multimedia objects to assign weights. It searches all the individual content units of the tree, if it finds any text content in the HTML element, the system counts the text length, converts the text length into pixels and assigns to the individual content unit as weight. Fig.9.(line 7-8), shows the algorithm converting the text lengths to pixels. Here v_w is the total number of the pixels, 10 is the height of the character and 30 is the threshold, $v_w = (30+10\times$ *number of text character*); [2].

The system searches all the individual content units of the Tree, if it finds any of the multimedia contents in the HTML element; it checks the properties of the multimedia contents. If the properties of the content dimensions are provided, it then grabs the information for the weight; otherwise system checks the multimedia content and gets the dimension by itself. Fig.9.(line 9-10), shows the algorithm where *height of the image* is the height of the multimedia content and *width of the image* is the weight of the multimedia content. The variable v_w is the area and weight of the multimedia object. v_w = *height of the image* × *width of the image*;

```
1.  Input:
2.    v is the node of the Tree;
3.  Output:
4.    v with weight value;
5.  Start
6.     AssingWeight(v)
7.     If v is text object Then
8.        Assign weight to $v_w = v_w + (30 + 10 \times \text{number of text character})$;
9.     If v is image object and image height, width more than 20pixels Then
10.       Assign weight to $v_w = v_w + (\text{height of the image} \times \text{width of the image})$;
11. End
```

Fig. 9. Algorithm of assignment of weight to individual content unit

3.5 Greedy Algorithm for Selecting Blocks to Display on Mobile Device

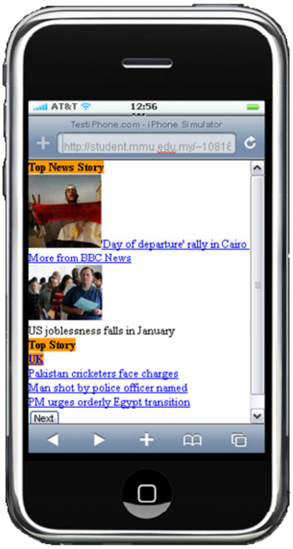
The greedy algorithm selects the highly profitable individual content unit from the tree and generates subpages with the weight capacity of individual content unit. The total weights of individual content units cannot be more than the mobile screen capacity for each subpage. The greedy algorithm selects the most profitable individual content units according to the user define preferences and the weighs must be less than or equal of the mobile screen capacity.

Fig.10. illustrates, let $contents_i$ is the tree individual content unit, w_i is the weight and p_i is the profit assigned by i_{th} item, C is the capacity base on the screen mobile dimension (*Height* × *Width*).

Fig.10, illustrates subpage generator scheme from the tree. The greedy algorithm selects the highest profitable objects and weight smaller or equal to the C capacity.

```
1.   Input:
2.     $contents$ is the tree individual content unit;
3.     $i$ is the number of offset;
4.     $w$ is the weight;
5.     $next$ is the next button;
6.     $previous$ is the previous button;
7.   Output:
8.     $subpage$ is HTML page contains individual content unit;
9.   Start
10.    while $weight < C$ Do
11.      If $next$ = TRUE Then
12.        Do $i$ = best new selected $contents$ according by $p_i$ maximum value;
13.      Elseif $previous$ = TRUE Then
14.        Do $i$ = best old selected $contents$ according by $p_i$ minimum value;
15.      If $weight + w_i \leq C$ Then
16.        $subpage = subpage$ append $contents_i$
17.        $weight$ = weight + $w_i$
18.      $i$ go to next offset until $n$;
19.  End
```

Fig. 10. Greedy algorithm for selecting blocks


AT&T
12:56
TestiPhone.com - iPhone Simulator
http://student.mmu.edu.my/~1081E

Top News Story

'Day of departure' rally in Cairo

More from BBC News

US joblessness falls in January

Top Story

UK

Pakistan cricketers face charges

Man shot by police officer named

PM urges orderly Egypt transition

Next

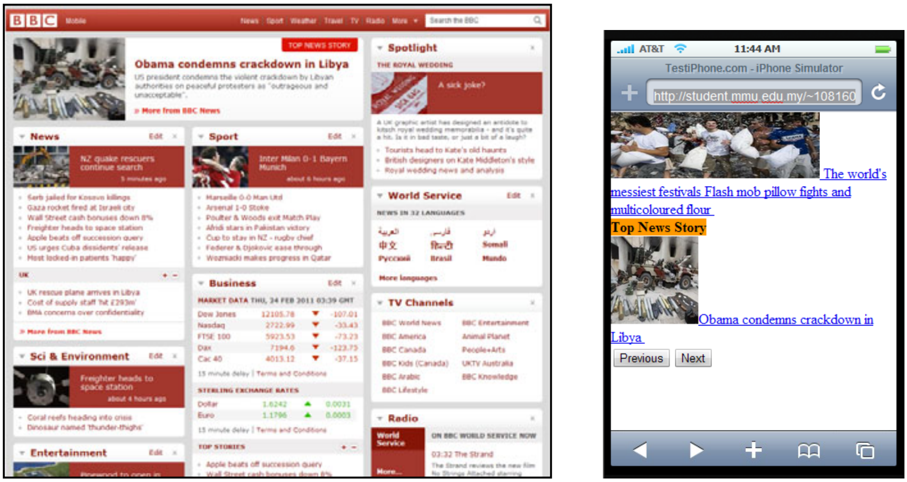
Fig. 11. Adapted BBC page by Web-adaptor on handheld device

When the weight becomes more than the capacity C, the then process stops (line 10). The algorithm delivers the selected contents in a subpage. The *next* and *previous* value is used to direct the next subpage or previous subpage and dominates the p. If the next is true then p chooses the maximum profit (line 12) and if the previous is true, then p contains the minimum profit (line 14). The Greedy algorithm creates small subpages from the tree. The first subpage contains the most important contents according to the user define preferences with the weight capacity. The second subpage contains less important contents compare to first subpage.

The Web-adaptor is a dynamic Web content adaptation system. Fig. 11 illustrates the BBC main page after adaptation. Any Web site can adapt by this system. Sometimes, there are few Web sites that can’t be adapted properly because they don’t follow the standard (W3C) format of the HTML.

4 Results and Discussion

A mobile user first opens a general web browser and connects to the Web-adaptor. The user types the URL of a webpage in the input box Fig.1. The users choose their preferences like searching for any particular info in the page or preferring to browse multimedia contents or navigations. After selecting the preferences, the user submits the request to the Web-adaptor and provides contents which fit the users screen. The system delivers the adapted contents to the users. Fig.12.(a) illustrates the desktop version of BBC webpage. After adapting, fig.12.(b) illustrates the system provides the BBC webpage for the mobile. The background color, font size and multimedia contents are modified to suite the mobile screen. If the user does not find the target contents, he can then navigate to next or previous pages. In the next section of this paper, we compare our proposed method with other existing methods; discuss potential implementation issues and other considerations.



Obama condemns crackdown in Libya

US president condemns the violent crackdown by Libyan authorities on peaceful protesters as "outrageous and unacceptable".

» More from BBC News

NZ quake rescuers continue search
5 minutes ago

- Serb jailed for Kosovo killings
- Gaza rocket fired at Israeli city
- Wall Street cash bonuses down 8%
- Freighter heads to space station
- Apple beats off succession query
- US urges Cuba dissidents' release
- Most locked-in patients 'happy'

- UK rescue plane arrives in Libya
- Cost of supply staff 'hit £293m'
- BMA concerns over confidentiality

» More from BBC News

Freighter heads to space station
about 4 hours ago

- Coral reefs heading into crisis
- Dinosaur named 'thunder-thighs'

Borewood to open in

Inter Milan 0-1 Bayern Munich
about 6 hours ago

- Marseille 0-0 Man Utd
- Arsenal 1-0 Stoke
- Poulter & Woods exit Match Play
- Afridi stars in Pakistan victory
- Cup to stay in NZ - rugby chief
- Federer & Djokovic ease through
- Wozniacki makes progress in Qatar

MARKET DATA THU, 24 FEB 2011 03:39 GMT

Dow Jones	12105.78	▼	-107.01
Nasdaq	2722.99	▼	-33.43
FTSE 100	5923.53	▼	-73.23
Dax	7194.6	▼	-123.73
Cac 40	4013.12	▼	-37.15

15 minute delay | Terms and Conditions

STERLING EXCHANGE RATES

Dollar	1.6242	▲	0.0031
Euro	1.1796	▲	0.0003

15 minute delay | Terms and Conditions

TOP STORIES + −

- Apple beats off succession query
- Wall Street cash bonuses down 8%

THE ROYAL WEDDING

A sick joke?

A UK graphic artist has designed an antidote to kitsch royal wedding memorabilia - and it's quite a hit. Is it in bad taste, or just a bit of a laugh?

- Tourists head to Kate's old haunts
- British designers on Kate Middleton's style
- Royal wedding news and analysis

NEWS IN 32 LANGUAGES

العربية فارسی اردو
中文 हिन्दी Somali
Русский Brasil Mundo

More languages

BBC World News	BBC Entertainment
BBC America	Animal Planet
BBC Canada	People+Arts
BBC Kids (Canada)	UKTV Australia
BBC Arabic	BBC Knowledge
BBC Lifestyle	

ON BBC WORLD SERVICE NOW

03:32 The Strand

The Strand reviews the new film No Strings Attached starring

(a)

http://student.mmu.edu.my/~108160

The world's messiest festivals Flash mob pillow fights and multicoloured flour

Top News Story

Obama condemns crackdown in Libya

Previous Next

(b)

Fig. 12. (a) BBC webpage desktop version and (b) adapted BBC webpage for the mobile

4.1 Framework Comparison

There are many significant differences between our proposed system and other existing content adaptation systems. The Xadaptor [2] framework builds on five components. It is a rule-based adaptive content adaptation system; fuzzy logic to model the adaptation quality and control the adaptation decision. The WebPage Tailoring System [6] is a complete framework to adapt contents for mobile devices. This system consists of three components. It uses mechanism that can determine which blocks in a webpage should be retained by user preferences and arrange the blocks. The CMo [5] framework builds on three components. It captures the context of the link, applies simple topic-boundary detection technique and uses the context to identify relevant information in the next page by using Vector machine. Our proposed Web-adaptor framework consists of five components. We compare the components on table 1 which they use to develop their systems.

Table 1. Comparison with Mobile Content Adapter Components

Components Used	(1) Xadaptor System	(2) CMo System	(3) Web Page Tailoring System	(4) Proposed System	Comments
User preferences	Yes	Yes	Yes	Yes	System 1 and 3 uses predefined user preferences from database but system 2 and 4 uses dynamic preferences from the users.
Transfer HTML page to Tree	Yes	Yes	Yes	Yes	System 1 use structure tree, system 2 use frame tree, system 3 use DOM tree and system 4 use simple HTML tree.
Block identification	Yes	Yes	Yes	Yes	System 1- 4 identify the blocks by HTML tags.
Object identification	Yes	No	No	Yes	System 1and 4 identify the objects by HTML tags.
Modification object elements	Yes	No	No	Yes	System 1 modifies the objects according to the user database but system 4 modifies the objects according to the dynamic data.
Re-arrange the sequence of the blocks for display	No	No	No	Yes	System 4 re-arranges the contents by the user preferences.
Use mechanism to select the target contents for display	No	Yes	Yes	Yes	System 2 delivers the related information, System 3 delivers information according to the tag pattern matching and system 4 delivers exact information.

Table.1. illustrates the comparison of the framework components. Other frameworks do not re-arrange the sequence of the blocks. In our proposed system, we re-arrange the block sequences to provide the target contents to the users.

4.2 Visualization Comparison

The Web Page Tailoring System, Xadaptor and our proposed system delivers the contents according to the user preferences. But there are differences in the visual outlook. The Web Page Tailoring System, users can zoom in which part of the page he wants to read. The multimedia content sometime gets oversized to the mobile screen because The Web Page Tailoring System doesn't adapt all kinds of contents. The Xadaptor and proposed system adapt the multimedia and text contents with rule base approach. The Xadaptor and Web Page Tailoring System display all the contents in a single column page but our proposed system deliver all the target contents in subpages.

5 Conclusion

In this paper, we proposed a new method for facilitating the browsing of a large webpage on a handheld device. The Web-adaptor converts HTML page into a tree and identifies the semantic blocks. Each object of the blocks gets profit and weight. The Pre-processor uses rules to modify the contents. The greedy algorithm selects the best contents. The system is able to generate subpages for the handheld devices. Our approach enables a new browsing experience to the users. The most significant information will appear at the first page of sub page. A new browsing method overcomes the limitation of a mobile device with a small screen and makes them truly useful for information access.

References

1. Xiao, X., luo, Q., Hong, D., Fu, H., xie, X., Ma, W.-Y.: Browsing on Small Displays by Transforming Web Pages into Hierarchically Structured Subpages. ACM Transactions on the Web 3(1), Article 4 (January 2009)
2. He, J., Gao, T., Hao, W., Yen, I.-L., Bastani, F.: A Flexible Content Adaptation System Using a Rule-Based Approach. IEEE Transactions on Knowledge and Data Engineering 19(1), 127–140 (2007)
3. Blekas, A., Garofalakis, J., Stefanis, V.: Use of RSS feeds for Content Adaptation in Mobile Web Browsing. In: International Cross-Disciplinary Workshop on Web Accessibility (W4A), pp. 79–85 (May 2006)
4. Lee, E., Kang, J., Choi, J., Yang, J.: Topic-SpecificWeb Content Adaptation to Mobile Devices. In: International Conference on Web Intelligence, pp. 845–848 (December 2006)
5. Borodin, Y., Mahmud, J., Ramakrishnan, I.V.: Context Browsing with Mobiles - When Less is More. In: International Conference on Mobile Systems, Applications and Services, pp. 3–15 (June 2007)
6. Kao, Y.-W., Kao, T.-H., Tsai, C.-Y., Yuan, S.-M.: A personal Web page tailoring toolkit for mobile devices. Computer Standards & Interfaces 31(2), 437–453 (2009)

7. Ahmadi, H., Kong, J.: Efficient Web Browsing on Small Screens. In: International Conference on Advanced Visual Interfaces, pp. 23–30 (May 2008)
8. Horia, M., Ono, K., Abe, M., Koyanagi, T.: Generating transformational annotation for web document adaptation: tool support and empirical evaluation. Journal of Web Semantics 2(1), 1–18 (2004)
9. Lee, E., Kang, J., Park, J., Choi, J., Yang, J.: ScalableWeb News Adaptation To Mobile Devices Using Visual Block Segmentation for Ubiquitous Media Services. In: International Conference on Multimedia and Ubiquitous Engineering (MUE 2007), pp. 620–625 (April 2007)
10. Pan, R., Wei, H., Wang, S., Luo, C.: Auto-adaptation of Web Content: Model and Algorithm. In: IET 2nd International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2008), pp. 507–511 (October 2008)
11. Chen, J., Zhou, B., Shi, J.: Function-Based Object Model Towards Website Adaptation. In: International World Wide Web Conference, pp. 587–596 (May 2001)
12. Yang, S.J.H., Zhang, J., Chen, R.C.S., Shao, N.W.Y.: A Unit of Information–Based Content Adaptation Method for Improving Web Content Accessibility in the Mobile Internet. ETRI Journal 29(6), 794–807 (2007)
13. Xu, K., Zhang, D., Zhu, M., Gu, T.: Context-Aware Content Filtering & Presentation for Pervasive & Mobile Information Systems. In: International Conference on Ambient MEdia and Systems and Workshops, Article 20 (February 2008)

QoS-Aware Web Services Selection with Interval-Valued Intuitionistic Fuzzy Soft Sets

Xiuqin Ma, Norrozila Sulaiman, and Mamta Rani

Faculty of Computer Systems and Software Engineering
Universiti Malaysia Pahang
Lebuh Raya Tun Razak, Gambang 26300, Kuantan, Malaysia
xueener@gmail.com,
{norrozila,mamta}@ump.edu.my

Abstract. With the increasing popularity of the development of service-oriented applications, it is imperative to measure the quality of services (QoS) for service consumers and providers. To find the most suitable service for different consumers, the QoS nonfunctional attribute will become an important factor in web service selection. However, non-functional QoS properties rely heavily on the subjective perceptions of service consumers that are not easy to assess due to their complexity and involvement of ill-structured information. The purpose of this paper is to introduce interval-valued intuitionistic fuzzy soft set theory for solving web service selection problems that take into account QoS requirement of consumers. Interval-valued intuitionistic fuzzy soft set theory, which is a new useful mathematical tool for dealing with uncertainties, is more effective to deal with uncertainties on non-functional QoS properties in web service selection. We present basic system architecture and the algorithm to solve fuzzy decision making problems for selecting web service based on interval-valued intuitionistic fuzzy soft sets. Finally, an illustrative example is employed to show our contribution.

Keywords: Web services; QoS; Soft sets; Interval-valued intuitionistic fuzzy soft sets; Service selection; Decision making.

1 Introduction

With the development of distributed computing technology, it is desirable to facilitate communication between applications and resource share in geographically distributed systems. As a result, the emergence of web services [1] brings changes to the traditional paradigm of distributed computing. A Web service [2] is a web accessible software that can be published, located and invoked by using the standard Web infrastructure. However, there exist a large number of web services which provide similarly functional characteristics. Multiple services with similarly functional characteristics give rise to the problem of service selection [3]. Consumers not only expect the service to meet functional aspects but they also demand good quality of services (QoS) such as service reliability, security, trust and execution cost. So QoS

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 259–268, 2011.

has been considered as a significant factor in the selection of web services. QoS of web services compose both functional and non-functional properties [4]; functional properties can be measured in terms of throughput, latency, response time; non-functional properties address various issues including integrity, reliability, availability and security of web services [5]. In recent years, a number of researchers have devised such techniques in order to help consumers in the service selection process. For example, Several QoS-based service matchmaking techniques [6] have been proposed to meet the needs of both consumers and providers. Huang et al. [7] employed fuzzy group decision-making methods and semantic web technologies to discover appropriate services by taking into account consumers' expectations and preferences. Lin et al. [8] proposed a QoS consensus moderation approach (QCMA) to analyze the group consensus based on their fuzzy opinion similarity and QoS preference with a number of QoS attributes for depicting how the group of consumers selecting a web service. However, there are still some problems that the single group based QCMA with opinion similarity and preference analysis for web service selection has not addressed. In order to overcome these problems, the FMG-QCMA (Fuzzy Multi-Groups based QoS Consensus Moderation Approach) [9] was proposed by the same author. At the same time, there are also some efforts which have been done to such issues concerning selecting services for building their composite service. By making use of matchmaking algorithms, Sirine et al. [10] depicted a goal-oriented and interactive composition approach to assist users filter and select services while building their composition service. Lin et al. [11] treated the selection of QoS-driven web service with dynamic composition as a fuzzy constraint satisfaction problem and applied an optimal search approach with adjustments to service composition. Wang et al. [12] designed a QoS-aware service selection model based on fuzzy linear programming (FLP) technologies to identify their dissimilarity on service alternatives and assist service consumers in selecting most suitable services with consideration of their expectations and preferences.

Soft set theory [13] initiated by Molodtsov is a new general mathematical tool for dealing with uncertainties which is free from the inadequacy of the parameterization tools. In recent years, work on the soft set theory has been active and great progress has been achieved [14, 15, 16, 17, 18, 19]. It is worthwhile to mention that Jiang et al. [20] presented the interval-valued intuitionistic fuzzy soft set theory by combining the interval-valued intuitionistic fuzzy sets and soft sets.

The purpose of this paper is to introduce interval-valued intuitionistic fuzzy soft set theory for solving web service selection problems that take into account QoS requirement of consumers. We present the algorithm to solve fuzzy decision making problems to select web service based on interval-valued intuitionistic fuzzy soft sets. To find the most suitable service for different users, the QoS nonfunctional attribute will become an important factor for users'decision in web service selection. Practically, non-functional QoS properties rely heavily on the subjective perceptions of service consumers that are not easy to assess due to their complexity and involvement of ill-structured information. However, the methods mentioned above are more suitable to quantify functional QoS properties than nonfunctional properties which are strongly affected by decision maker's subjective perception. Our approach employing Interval-valued intuitionistic fuzzy soft set theory, which is an useful mathematical tool for dealing with uncertainties, is more effective to deal with uncertainties on non-functional QoS properties in web service selection. An example is illustrated to demonstrate the proposed approach.

The rest of paper is organized as follows. Section 2 introduces the basic principles of soft sets and interval-valued intuitionistic fuzzy soft set theory. Section 3 gives basic system architecture. Section 4 presents a new approach of QoS-aware web services selection based on decision making algorithm of Interval-valued intuitionistic fuzzy soft sets. Section 5 illustrates an example to demonstrate the proposed approach. Finally, section 6 presents the conclusion.

2 Basic Notions

In this section, we review some definitions with regard to soft sets and interval-valued intuitionistic fuzzy soft sets.

Let U be a non-empty initial universe of objects, E be a set of parameters in relation to objects in U, $P(U)$ be the power set of U, and $A \subset E$. The definition of soft set is given as follows.

Definition 2.1. (See [13]). *A pair (F,A) is called a soft set over U, where F is a mapping given by*

$$F : A \to P(U) \ . \tag{1}$$

That is, a soft set over U is a parameterized family of subsets of the universe U.

Atanassov and Gargov [21] first initiated interval-valued intuitionistic fuzzy set (IVIFS), which is characterized by an interval-valued membership degree and an interval-valued non-membership degree.

Definition 2.2 (See [20, 21]). *An interval-valued intuitionistic fuzzy set on a universe X is an object of the form*

$$A = \{\langle x, \mu_A(x), \gamma_A(x)\rangle | x \in X\} \ . \tag{2}$$

where $\mu_A(x) : X \to Int([0,1])$ *and* $\gamma_A(x) : X \to Int([0,1])$ *(* $Int([0,1])$ *stands for the set of all closed subintervals of [0, 1]) satisfy the following condition:* $\forall x \in X, \sup \mu_A(x) + \sup \gamma_A(x) \le 1$.

Let U be an initial universe of objects, E be a set of parameters in relation to objects in U, $\zeta(U)$ be the set of all interval-valued intuitionistic fuzzy sets of U. The definition of interval-valued intuitionistic fuzzy soft set is given as follows.

Definition 2.3. (See [20]). *A pair* $(\tilde{\varphi}, E)$ *is called an interval-valued intuitionistic fuzzy soft set over* $\zeta(U)$, *where* $\tilde{\varphi}$ *is a mapping given by*

$$\tilde{\varphi} : E \rightarrow \zeta(U) . \tag{3}$$

In other words, an interval-valued intuitionistic fuzzy soft set is a parameterized family of interval-valued intuitionistic fuzzy subsets of *U*. Hence, its universe is the set of all interval-valued intuitionistic fuzzy sets of *U*, i.e. $\zeta(U)$.

3 Basic System Architecture

This proposed architecture integrates the ideas of [22, 23, 24, 25] and our own idea. It includes four roles [25]: service registry with QoS data type, service provider, service consumer, QoS agent, which are shown in Figure 1.

(1) Service Registry with QoS data type
Service providers publish their services and provide OoS information to Service Registry. This model extends Service Registry by adopting QoS data type. QoS agent evaluates services according to QoS information from providers and consumers. Service Consumers discover service based on QoS request and Service request from Service Registry.

(2) Service Provider
It can provide services which can be invoked by consumers and QoS information.

(3) QoS Agent

- QoS recorder: which is responsible for making confirmation whether any web service has been added or withdrawn from Service Registry. If related service and QoS information has been changed, it updates QoS database.
- QoS Evaluator: which evaluates services and give a ranking for services possessing the same or similar functions according to QoS information from providers and consumers.
- QoS aggregater: which collects the consumer's feedback and passes them to the QoS database and then aggregates QoS information from provider and consumers for the same web service into consistent QoS information.

(4) Service Consumer
It invokes all kinds of various services which are provided by service providers.

- Service request: which searches service in service registry according to functional service demand.
- QoS request: which is used to communicate with the QoS agent.
- QoS feedback: which aims to collect consumer's appraise for the used service.

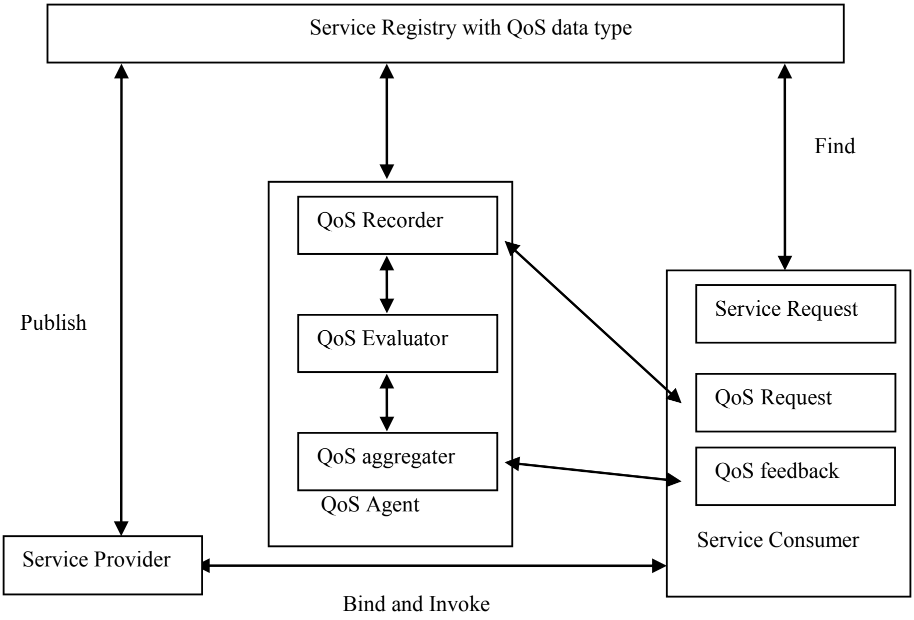


Fig. 1. Basic System Architecture

4 The Proposed Approach for Selecting Web Services

In this section, we address two problems (1) how to aggregate QoS information; (2) how to obtain the optimal web service according to service request and QoS request from consumers. It is essential to integrate QoS information as preprocess before evaluating web services.

4.1 Aggregating QoS Information

There are some different QoS information from the provider and many consumers who have used the service for the same service. Aiming to evaluate this service, it is necessary to aggregate different QoS information into consistent one, depending on the service requester's preference. If a user prefers the objective QoS information from the provider, the consumer can heighten the objective weight, or if the consumer believes QoS subjective information from the consumers' feedback, the consumer can heighten the subjective weight. The method of calculating the value for every QoS attribute is as follows:

$$f_i(S_j) = W_p V_p + W_c(\frac{V_{c1} + V_{c2} + \ldots + V_{cn}}{n}) \ . \tag{4}$$

Where W_p is the weight of QoS objective information from the provider; V_p is objective attribute value from provider; W_c is the weight of QoS subjective information from the consumers; $V_{cs}(s=1,...,n)$ is subjective attribute value from n consumers who have already used this service.

4.2 Evaluating and Selecting Web Services

In this research, we use the interval-valued intuitionistic fuzzy soft set to describe quality of web services for the decision maker. The selection of QoS-aware web services is modeled as a fuzzy decision-making problem that includes the following important aspects (1) imprecise preference. This method should be able to handle the problem that vague preferences for non-functional QoS attributes are employed by service consumers in the process of selecting web services. (2)QoS-aware service ranking. The approach should be capable of obtaining a QoS-based ranking on all the alternatives web services. We present the algorithm to solve fuzzy decision making problems to select web service based on interval-valued intuitionistic fuzzy soft sets in the following.

1) Input an interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi}, E)$ and the parameter set *E*. $U=\{S_1,S_2,\cdots,S_n\}$ be a set of web services, $E=\{e_1,e_2,\cdots,e_m\}$ be a set of QoS attribute, $\mu_{\tilde{\varphi}(e_j)}(S_i)=[\mu^-_{\tilde{\varphi}(e_j)}(S_i),\mu^+_{\tilde{\varphi}(e_j)}(S_i)]$ is the degree of membership an element S_i to $\tilde{\varphi}(e_j)$. $\gamma_{\tilde{\varphi}(e_j)}(S_i)=[\gamma^-_{\tilde{\varphi}(e_j)}(S_i),\gamma^+_{\tilde{\varphi}(e_j)}(S_i)]$ is the degree of non-membership an element S_i to $\tilde{\varphi}(e_j)$.

2) Compute score of membership degree $p_{\tilde{\varphi}(e_j)}(S_i)$ for e_j such that

$$p_{\tilde{\varphi}(e_j)}(S_i)=\sum_{k=1}^{n}[(\mu^-_{\tilde{\varphi}(e_j)}(S_i)+\mu^+_{\tilde{\varphi}(e_j)}(S_i))-(\mu^-_{\tilde{\varphi}(e_j)}(S_k)+\mu^+_{\tilde{\varphi}(e_j)}(S_k))] \cdot \tag{5}$$

3) Compute score of non-membership degree $q_{\tilde{\varphi}(e_j)}(S_i)$ for e_j such that

$$q_{\tilde{\varphi}(e_j)}(S_i)=-\sum_{k=1}^{n}[(\gamma^-_{\tilde{\varphi}(e_j)}(S_i)+\gamma^+_{\tilde{\varphi}(e_j)}(S_i))-(\gamma^-_{\tilde{\varphi}(e_j)}(S_k)+\gamma^+_{\tilde{\varphi}(e_j)}(S_k))] \cdot \tag{6}$$

4) Compute score $u_{\tilde{\varphi}(e_j)}(S_i)$ for e_j such that

$$u_{\tilde{\varphi}(e_j)}(S_i)=p_{\tilde{\varphi}(e_j)}(S_i)+q_{\tilde{\varphi}(e_j)}(S_i) \cdot \tag{7}$$

5) Compute the overall weighted score t_i for S_i such that

$$t_i=\omega_1 u_{\tilde{\varphi}(e_1)}(S_i)+\omega_2 u_{\tilde{\varphi}(e_2)}(S_i)+...+\omega_m u_{\tilde{\varphi}(e_m)}(S_i) \quad (0\le\omega_j\le 1, j=1,...,n) \tag{8}$$

Where ω_j represents the weight for various QoS attributes from the consumer's opinion.

6) Find k , for which $t_k=\max_{S_i\in U}\{t_i\}$. Then $S_k\in U$ is the optimal choice web service.

5 An Example

To illustrate the proposed approach, we give an example of selecting web services.

Example 1. Let $U=\{S_{p1},S_{c1},S_{p2},S_{c2},S_{p3},S_{c3}\}$ be a set of alternative web services, where it is assumed that for the same web service S_j, S_{pj} represents QoS appraises from the provider, S_{cj} expresses QoS appraises from the consumers. Let $E=\{e_1,e_2,e_3,e_4\}$ be a set of QoS attributes, where e_i stand for "high performance", "high reliability", "high security", "low cost" respectively. It is worthwhile to explain that there are many QoS attributes extracted from research report of W3C (2003) [24]. In order to briefly illustrate the proposed approach, we only choose four QoS attributes. A consumer is interested in obtain a web service on the basis of his choice parameters "high performance", "high reliability", "high security" and "low cost". The consumer give that W_p =0.4, W_c =0.6, which means this consumer prefers QoS subjective information from the consumers' feedback to objective QoS information from the provider. Furthermore we assume that each of QoS attribute is associated with a weight ω_j indicating its importance considered by this consumer. Here $\omega_1=0.8,\omega_2=0.4,\omega_3=0.5,\omega_4=0.9$ .The interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi},E)$ describes the "attractiveness of the web services" to the consumer. The consumer performs an evaluation of three web services and selects the best one. The proposed method is applied to solve this problem according to the following steps:

Step1: Integrating QoS information from the provider and the consumer into consistent one. It can be carried out by using Eqs.(4) shown in Table 2. The consumer give that W_p =0.4, W_c =0.6, which means this consumer prefers QoS subjective information from the consumers' feedback to objective QoS information from the provider.

Table 1. An interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi},E)$ for describing web services

U/E	e_1	e_2	e_3	e_4
S_{p1}	[0.60,0.80],[0.05,0.15]	[0.75,0.85],[0.05,0.15]	[0.70,0.85],[0.10,0.15]	[0.65,0.75],[0.10,0.20]
S_{c1}	[0.50,0.70],[0.10,0.20]	[0.60,0.80],[0.05,0.20]	[0.45,0.55],[0.20,0.40]	[0.60,0.70],[0.15,0.25]
S_{p2}	[0.70,0.80],[0.05,0.15]	[0.70,0.80],[0.10,0.20]	[0.75,0.85],[0.00,0.10]	[0.80,0.90],[0.00,0.10]
S_{c2}	[0.40,0.50],[0.40,0.50]	[0.65,0.80],[0.10,0.20]	[0.65,0.85],[0.05,0.10]	[0.80,0.90],[0.00,0.10]
S_{p3}	[0.60,0.70],[0.10,0.20]	[0.50,0.60],[0.20,0.30]	[0.60,0.70],[0.10,0.20]	[0.40,0.50],[0.20,0.30]
S_{c3}	[0.40,0.60],[0.20,0.30]	[0.30,0.40],[0.40,0.50]	[0.50,0.70],[0.10,0.20]	[0.60,0.80],[0.10,0.20]

Table 2. An new interval-valued intuitionistic fuzzy soft set $(\tilde{\varphi}', E)$ having consistent QoS informaton

U/E	e_1	e_2	e_3	e_4
S_1	[0.54,0.74],[0.08,0.18]	[0.66,0.82],[0.05,0.18]	[0.55,0.67],[0.16,0.30]	[0.62,0.72],[0.13,0.23]
S_2	[0.52,0.62],[0.28,0.36]	[0.67,0.80],[0.10,0.20]	[0.69,0.85],[0.03,0.10]	[0.80,0.90],[0.00,0.10]
S_3	[0.48,0.64],[0.16,0.26]	[0.38,0.48],[0.32,0.42]	[0.54,0.70],[0.10,0.20]	[0.52,0.68],[0.14,0.24]

Step2: Gaining the score of membership degree and score of non-membership degree by making use of Eqs.(5) and Eqs.(6) illustrated in Table 3 and Table 4, respectively.

Table 3. The score of membership degrees for $(\tilde{\varphi}', E)$

U/E	e_1	e_2	e_3	e_4
S_1	0.3	0.63	-0.34	-0.22
S_2	-0.12	0.6	0.62	0.86
S_3	-0.18	-1.23	-0.28	-0.64

Table 4. The score of non-membership degrees for $(\tilde{\varphi}', E)$

U/E	e_1	e_2	e_3	e_4
S_1	0.54	0.58	-0.49	-0.24
S_2	-0.6	0.37	0.5	0.54
S_3	0.06	-0.95	-0.01	-0.3

Step3: Obtaining the score of each alternative based on above results by Eqs.(7) given in Table 5.

Step4: Evaluating the overall weighted score. According to the given weight for every attribute, $\omega_1 = 0.8, \omega_2 = 0.4, \omega_3 = 0.5, \omega_4 = 0.9$, we gain the overall weighted score t_i employing Eqs.(8) and the results are shown in Table 5.

Table 5. The overall weighted score for each alternative

U/E	e_1	e_2	e_3	e_4	t_i
S_1	0.84	1.21	-0.83	-0.46	0.327
S_2	-0.72	0.97	1.12	1.4	1.632
S_3	-0.12	-2.18	-0.29	-0.94	-1.959

Step 5: Finding the optimal choice web service. From Table 5, we can draw conclusion that S_2 is the optimal choice web service, due to $t_2 = 1.632 = \max_{S_i \in U}\{t_i\}$, S_1 is the suboptimal web service, S_3 is the worst choice web service.

6 Conclusion

The interval-valued intuitionistic fuzzy soft set theory has been proposed by combining the interval-valued intuitionistic fuzzy sets and soft sets. However, up to the present, few documents have focused on its practical application. In this paper, we introduce interval-valued intuitionistic fuzzy soft set theory for solving web service selection problems that take into account QoS requirement of consumers. There are two main contributions of this work. First, this model applies interval-valued intuitionistic fuzzy soft set theory into web service selection. Interval-valued intuitionistic fuzzy soft set theory, which is a new useful mathematical tool for dealing with uncertainties, is more effective to deal with uncertainties on non-functional QoS properties in web service selection. Second, this approach considers not only objective QoS information from the provider but also subjective QoS information from the consumer, depending on the service requester's preference. In detail, we present basic system architecture and the algorithm to solve fuzzy decision making problems for selecting web service based on interval-valued intuitionistic fuzzy soft sets. Finally, an illustrative example is employed to show our contribution.

References

1. Kreger, H.: Web Services Conceptual Architecture, WSCA 1.0 (May 2001)
2. Vaughan-Nichols, S.J.: Web services: Beyond the hype. Computer 35(2), 18–21 (2002)

3. Lin, W.-L., Lo, C.-C., Chao, K.-M., Younas, M.: Consumer-centric QoS-aware selection of web services. Journal of Computer and System Sciences 74, 211–231 (2008)
4. Wang, P.: QoS-aware web services selection with intuitionistic fuzzy set under consumer's vague perception. Expert Systems with Applications 36, 4460–4466 (2009)
5. Zhou, C., Chia, L.-T., Lee, B.-S.: Semantics in service discovery and QoS measurement. IT Professional 7(2), 29–34 (2005)
6. Chao, K.-M., Younas, M., Lo, C.-C., Tan, T.-H.: Fuzzy match-making for Web Services. In: ANIA 2005, pp. 721–726 (2005)
7. Huang, C.-L., Lo, C.-C., Li, Y., Chao, K.-M., Chung, J.-Y., Huang, Y.: Service discovery through multi-agent consensus. In: Proceedings of IEEE International Workshop on Service-Oriented System Engineering, SOSE 2005, pp. 37–44 (2005)
8. Lin, W.-L., Lo, C.-C., Chao, K.-M., Younas, M.: Consumer-centric QoS-aware selection of web services. Journal of Computer and System Sciences 74, 211–231 (2008)
9. Lin, W.-L., Lo, C.-C., Chao, K.-M., Godwin, N.: Multi-group QoS consensus for web services, vol. 77, pp. 223–243 (2011)
10. Sirine, E., Parsia, B., Hendler, J.: Filtering and selecting semantic web services with interactive composition techniques. IEEE Intelligent Systems, 42–49 (2004)
11. Lin, M., Xie, J., Guo, H., Wang, H.: Solving QoS-driven Web service dynamic composition as fuzzy constraint satisfaction. In: Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service, IEEE 2005, pp. 9–14 (2005)
12. Wang, P., Chao, K.-M., Lo, C.-C.: On optimal decision for QoS-aware composite service selection. Expert Systems with Applications 37, 440–449 (2010)
13. Molodtsov, D.: Soft set theory_First results. Comput. Math. Appl. 37(4/5), 19–31 (1999)
14. Herawan, T., Mat Deris, M.: A Soft Set Approach for Association Rules Mining. Knowledge Based Systems (2010) doi: 10.1016/j.knosys.2010.08.005
15. Feng, F.: Generalized rough fuzzy sets based on soft sets. In: the Proceeding of 2009 International Workshop on Intelligent Systems and Applications, ISA 2009, pp. 1–4 (2009)
16. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. Journal of Fuzzy Mathematics 9(3), 589–602 (2001)
17. Maji, P.K., Roy, A.R.: An application of soft sets in a decision making problem. Computers and Mathematics with Applications 44, 1077–1083 (2002)
18. Maji, P.K., Biswas, R., Roy, A.R.: Intuitionistic fuzzy soft sets. Journal of Fuzzy Mathematics 9(3), 677–692 (2001)
19. Chen, D., Tsang, E.C.C., Yeung, D.S., Wang, X.: The parameterization reduction of soft sets and its applications. Computers and Mathematics with Applications 49(5–6), 757–763 (2005)
20. Jiang, Y., Tang, Y., Chen, Q., Liu, H., Tang, J.: Interval-valued intuitionistic fuzzy soft sets and their properties. Computers and Mathematics with Applications 60(3), 906–918 (2010)
21. Atanassov, K., Gargov, G.: Interval valued intuitionistic fuzzy sets. Fuzzy Sets And Systems 31(3), 343–349 (1989)
22. Chen, H., Yu, T., Lin, K.J.: QCWS: an implementation of QoS-capable multimedia web services. In: Chang, C.M. (ed.) IEEE Fifth International Symposium on Multimedia Software Engineering, Taichung, Taiwan (2003)
23. Ran, S.: A model for web services discovery with QoS. ACM SIGecom Exchanges 4(1), 1–10 (2003)
24. W3C, Web Services Architecture (2003), `http://www.w3.org/TR/2003/WD-ws-arch-20030808/`
25. Wang, H.-C., Lee, C.-S., Ho, T.-H.: Combining subjective and objective QoS factors for personalized web service selection. Expert Systems with Applications 32, 571–584 (2007)

DBrain Portal for Collaborative BioGrid Environment

Al Farisi and Mohd Fadzil Hassan

Computer and Information Sciences Department
Universiti Teknologi PETRONAS
Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia
el.farisi@ymail.com, mfadzil_hassan@petronas.com.my

Abstract. DBrain is a project to support dementia-affected people. One of the targeted deliverables from the DBrain project is a grid system that will be utilized for diagnosis, therapeutics, and treatment of dementia disease. This paper will present a study to provide a collaborative environment through a single portal interface to access all resources from multiple institutions in a grid. Users, including researchers and scientists, only need to log into the portal. They don't need to log into each of individual resources to use them. The portal will also support a variety of grid middleware platforms. It will serve as interface for a specific-purpose grid infrastructure, called BioGrid. This type of grid is used to support computational applications in bioinformatics and biomedical sciences.

Keywords: BioGrid, DBrain Project, Grid Computing, Grid Portal.

1 Introduction

The DBrain project is an ongoing project aimed to support dementia-affected people, especially in Malaysia. This project will develop medical support system, including hardware and software, for the benefit of dementia sufferers, caregivers, and healthcare workers. One of the targeted deliverables from this project is a grid system that involves resources from multiple institutions. The grid will deliver specific services in bio-computing, such as gene discovery, protein modeling, and drug discovery.

There are some related work that implement grid computing in the area of bioinformatics and biomedical sciences, such as CancerGrid (drug discovery) [1] and ProSim (protein molecule simulation) [2]. The implementation of grid computing will brings a number of new capabilities. Doing research faster, working in collaborative environment, and reducing costs are some of capabilities promised by grid computing. In DBrain project, the grid computing infrastructure will be utilized for diagnosis, therapeutics, and treatment of dementia disease. Some of these jobs, as mentioned previously, require a high performance computing facility to be completed.

However, it is not easy to build a complete grid system. To create a complete grid system requires a wide variety of protocols, services, and software development kits [3]. A grid system involves heterogeneous resources from different institutions. Even

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 269–277, 2011.

in the DBrain project, each partner institution already has their own grid computing facilities established. These grids use two major grid middleware, gLite and Globus Toolkit 4. This paper will present how grid interoperability can be solved at the grid portal layer. A grid portal can serve as an interface to the grid infrastructure. Users can be isolated from resource specific details, system changes/differences, etc. By creating a single portal interface, user can access all the resources easily to support their collaborative works.

2 Grid and Its Interoperability

2.1 Grid System

The definitions for a grid are numerous. It has been suggested that the definitions of a grid can be captured in a simple checklist as follows [4].

- Coordinates resources that are not subject to centralized controls …
  (There are many resources and users from different institutions involved in the DBrain project. All these resources and users, that live within different control domains, will be integrated and coordinated by the grid.)
- … using standard, open, general-purpose protocols and interfaces …
  (There are numerous and heterogeneous resources involved in a grid, and somehow all these resources need to be coordinated and integrated to solve a specific problem. The only way to do this is by using standard and open protocols and interfaces.)
- … to deliver nontrivial qualities of services.
  (A grid allows its constituent resources to be used in a coordinated fashion to meet complex user demands.)

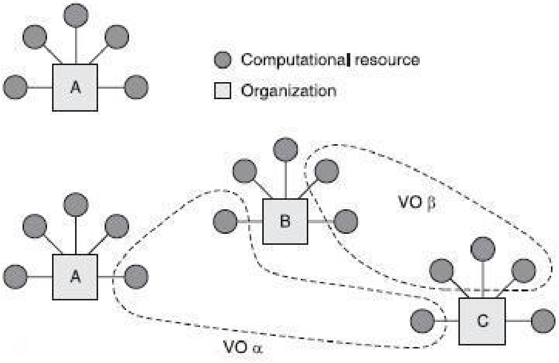


Fig. 1. Virtual Organizations [3]

Using a grid, resources from several institutions will be dynamically pulled into virtual organizations (VO) to solve a specific problem. Fig. 1 shows how the resources from different institutions can be combined into a virtual organization.

Each institution will be able to share its resources. Thus, VO members have direct access to each other's applications, files, data, hardware, and networks. Grid infrastructure that is presented in this paper, the BioGrid, will provide specific services to support computational applications in bioinformatics and biomedical sciences. The examples of application that supported by the biogrid are AutoDock [5] for molecular docking, and Gromacs [6] for molecular dynamics.

2.2 Challenges in Grid Computing

Heterogeneity. As mentioned earlier, there are heterogeneous resources from different institutions in a grid. Grid resources are owned by many different institutions that use different software and hardware. They also use different system for security and resources access. The major challenge is how the resources can communicate among themselves. That is why we need to use open, standard, general-purpose protocols and interfaces.

Resources Sharing. Grids give users access and control of resources from many institutions. Normally, each institution will put some conditions on the use of those resources, when and what resources are accessible. This sharing must be controlled, secure, flexible, and usually time-limited [7].

Security. To ensure secure access, grid developers and users need to manage three important things [7].

1. Access Policy: policies should be established on what resource is shared, who is allowed to share, and when sharing can occur.
2. Authentication: to identify a user or resource.
3. Authorization: to determine whether a certain operation is consistent with the rules.

Grid is usually secured using a combination of methods. The fundamental requirement is to enable VO to access resources that exist within multiple institutions as shown in Fig. 1. In order to coordinate resources, VOs need to establish trust, not only among users and resources in the VO, but also among the VO's resources [8].

3 Grid Portal

Because of its complexity, the grid needs to provide an easy-to-use interface for its users. Grid portal can be used for this purpose, to access grid infrastructures. Grid portal is a specialized web portal that provides an entry point to the grid infrastructure. For most users, grid portal is easier to use than command-line interface. Web technology has proven to be an effective way for delivering information to the user. By using grid portal, users can be isolated from grid complexity.

Grid portal can be accessed from anywhere and anytime. It uses HTTP/HTTPS, a general-purpose protocol, which is widely used for internet communication. This protocol is also used by the resources in a grid to communicate among themselves through a secure web-service mechanism.

3.1 Portal Requirements

There are several requirements that should be met by the DBrain portal as listed below which are based on [9] and [10].

1. Homogeneous: The portal should be capable to hide the heterogeneity of grid components by being homogenous.
2. Problem-oriented: The portal should allow its users (especially scientists, researchers, and physicians) to concentrate in their discipline without knowing the details about the grid back-end. It means that a portal should provide an abstraction layer on top of all the underlying diverse.
3. Easy-to-use: The portal should hide the complexity of grid infrastructure and provide a user-friendly interface. The use of graphical user interface will enhance the usability of grid portal. For example, by using graphs and diagrams to visualize data.
4. Persistent: Since operations in the grid may require significant time, it is needed to provide a persistent portal environment.
5. Collaborative: DBrain project is performed in a collaborative mode that involves participants from different institutions and locations. The portal must be able to support their collaborative works. For example, by providing collaboration tools for video conferencing, etc.
6. Widely Accessible: The portal needs to be accessible from any environment it is designed to support. On the users-side, there is no need for them to install software other than a browser. Some plug-in might be required to enhance graphical interface, such as Flash and Java plug-in.
7. Integrated: There are many resources and services in a grid. The portal must enable the integration and call of services provided by the grid.

3.2 Architecture

The proposed DBrain portal architecture can be divided into several layers as shown in Fig. 2. The highest layer is user interface. This layer interacts directly with users. The easy-to-use graphical interface is needed to help users easily access grid services.

The second layer is web-portal services. Web-portal services layer provides a collection of services to communicate with grid middleware. This layer should provide services that can be used to communicate to various grid middleware. The portal, with its services and interface, will be the liaison between users and grid services.

Resource access service is a middleware layer that provides tools to access grids. This layer will automate all the resources interactions to create a single and seamless grid. The lowest layer is distributed resources. This is the actual grid resources, such as computers, storage systems, data catalogues, hardware, and applications that are used in grids.

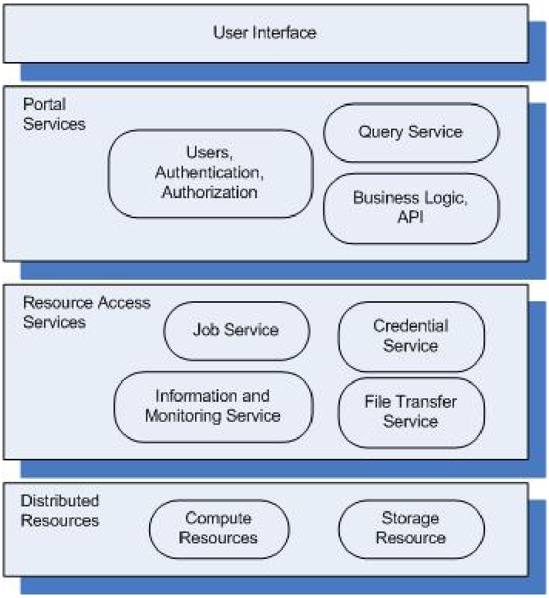


Fig. 2. Portal Architecture

3.3 Portal Components

The portal can be divided into five major components: credential management, job management, file/data management, information service, and collaboration tools.

Credential Management. This component is very important. It is used to manage users' proxy credentials. The portal must provides a mechanism from different level of security to make sure it is secured to be used. The credential management component will be combined with user account management component. Only authenticated user can log into the portal. Each component will be supplied with an access policy. Different users have different access on a particular component. Only permitted users can access specified components.

Once a user is authenticated, it is the job of the portal to act as a proxy in most grid interaction. Consequently, the portal must obtain a proxy credential that can be used on behalf of the user. Users must delegate to the portal the right to act on the user's behalf. Such grid resources are generally protected by the Grid Security Infrastructure (GSI) that has become the de-facto standard for grid security. While GSI supports such delegation, the standard web security protocols do not [11].

In order to bridge this incompatibility issue between web and grid security, the approach is by implementing MyProxy. Thus, the portal enables to use GSI-protected resources in a secure and scalable manner, as shown in Fig. 3. Users have to submit their proxy credential to MyProxy server. The portal then retrieves the proxy from MyProxy server and holds it for the duration of user's session.

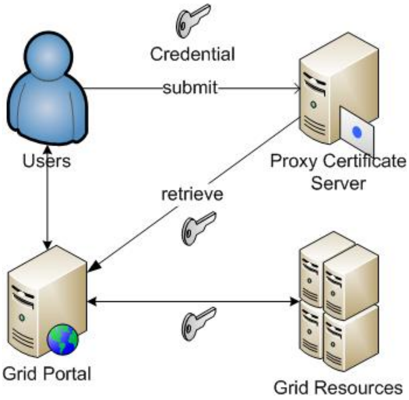


Fig. 3. Proxy Credential

Proxy credential is also used to decide whether users have access to VO's resources or not, including what resources and when it can be accessed.

Job Management. Users submit their job through this component. The portal will find resources available to be submitted to. Then it will schedule the job. The state of the job will be periodically stored in the job information center, which allows users to pull the latest state of the job or the result of the job [12]. Of course, to use this facility, users need to be authenticated and authorized first by logging into the portal and retrieving their proxy credential.

Data Management. In grids, data are stored on different file system using different access technologies. Data also stored in different locations. In most case, there is no shared file system. Data are also need to be described and located according to their content. To aid users in discovering relevant files within collections of thousands or millions of files, grids should provide metadata publication and search capabilities.

Data management component in biogrid portal will interact with data services in grids, such as metadata catalog and file transfer service. It allows users to manage their data.

Information Service. This component allows users to monitor and get information of grid resources, such as the usage of CPUs, memories, and jobs queue. This information is very useful to know what resource is available and ready to process a job.

Collaboration Tools. Collaboration tools include video conferencing tool, messaging, application and database repository. These tools are very useful to support collaboration among the users.

4 gUSE/WS-PGRADE Portal

There are several grid portal frameworks and standards which commonly used to build a grid portal, such as Vine Toolkit [13] and P-GRADE Portal [14]. This paper will explore the possibility of gUSE/WS-PGRADE Portal to be used in the DBrain project.

WS-PGRADE (Web Service Parallel Grid Runtime and Developer Environment) is the next generation of P-GRADE portal. It offers better functionality than P-GRADE portal. Just like P-GRADE, WS-PGRADE portal also tries to solve grid interoperability problem at the workflow level.

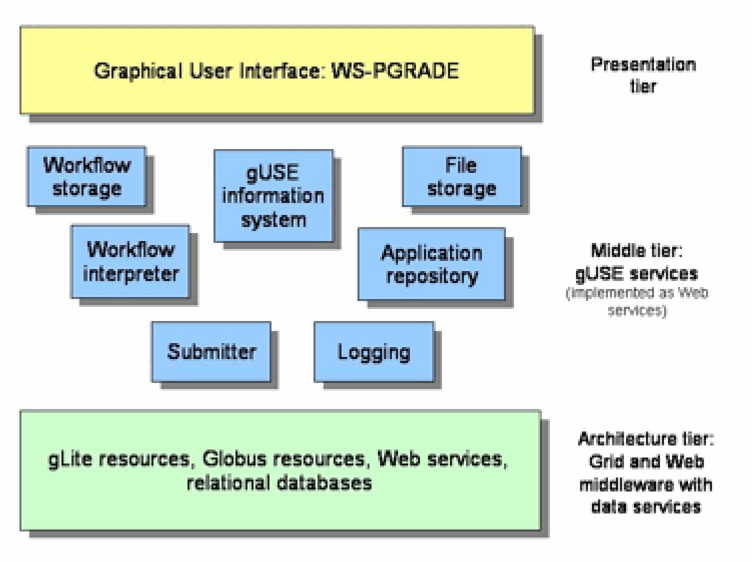

Graphical User Interface: WS-PGRADE
Presentation tier
Workflow storage
gUSE information system
File storage
Workflow interpreter
Application repository
Middle tier: gUSE services (implemented as Web services)
Submitter
Logging
gLite resources, Globus resources, Web services, relational databases
Architecture tier: Grid and Web middleware with data services

Fig. 4. gUSE Service Oriented Architecture [15]

WS-PGRADE uses the high-level grid service of gUSE (Grid User Support Environment), including workflow manager, storage, and job submitter. As shown in Fig. 3, gUSE is implemented as web services. gUSE supports various grid submission service, such as Globus Toolkit 2, LCG-2, Globus Toolkit 4, gLite, and BOINC [16]. Even, in the new release of gUSE (version 3.3), it already supports Cloud, such as GAE (Google App Engine) Cloud, and Amazon EC2 (Elastic Compute Cloud) [15][16].

gUSE/WS-PGRADE is a generic grid portal which then can be utilized to develop a specialized and customized portal. gUSE/WS-PGRADE is designed to support collaboration works by sharing applications and databases. A user can develop complex workflow-based applications. By publishing the applications in the repository, the application can be continued by other users to be developed and run. In addition, gUSE/WS-PGRADE comes with portlet-based technology based on Liferay [17]. Liferay has many plug-ins developed by the portal community, including for collaboration purpose. These plug-in can be easily integrated into the grid portal.

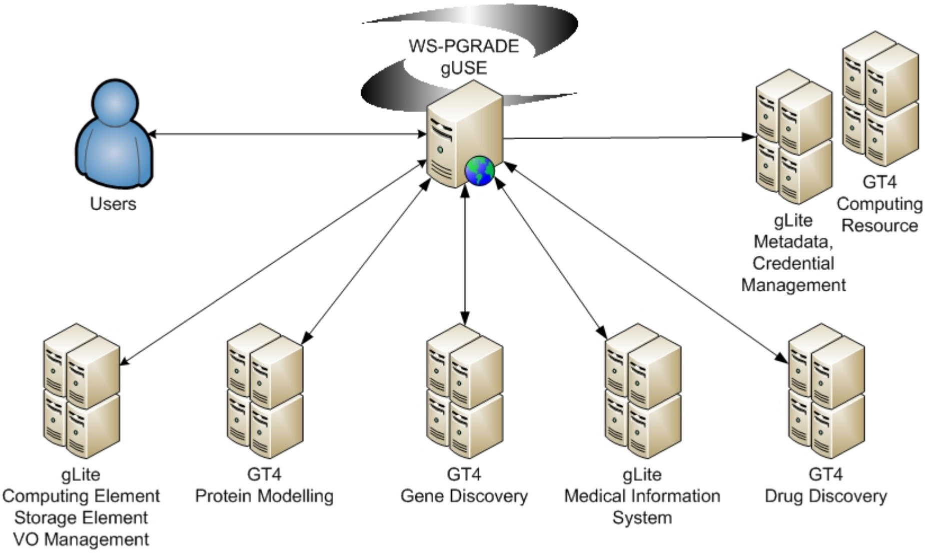


Fig. 5. DBrain Portal based on gUSE/WS-PGRADE

Considering all its features, gUSE/WS-PGRADE can be used as a basis to develop the DBrain portal. Fig. 5 shows how the gUSE/WS-PGRADE portal can be implemented to integrate and solve grid interoperability problem between partners' resources. It can be achieved by developing a specialized and customized portal, including specific workflow-based applications in the field of bioinformatics and biomedical sciences to support dementia-affected people.

5 Conclusion and Future Work

The DBrain Portal can serve as interface to the biogrid infrastructure. Users can be isolated from grid complexity, resources specific details, system differences, etc. By creating a single portal interface, users can access all resources in grids with ease. It will support their collaborative works.

Based on this study, the use of gUSE/WS-PGRADE can be considered to solve the grid interoperability problem at the portal and workflow level, especially in the DBrain project which involves heterogeneous resources and services. The next step is by developing a portal prototype and specific applications to meet the DBrain project requirements. In the future, there is a possibility for the portal to connect and access cloud infrastructure in Malaysia.

Acknowledgment. This paper was submitted under the TechnoFund grant (DBrain Project) Malaysia and support from the Universiti Teknologi PETRONAS.

References

1. CancerGrid, http://cancergrideu.w3h.hu/
2. ProSim Project,
   https://sites.google.com/a/staff.westminster.ac.uk/engage/Home/
3. Sotomayor, B., Childers, L.: Globus Toolkit 4 - Programming Java Services. Morgan Kaufmann Publishers, San Francisco (2006)
4. Foster, I.: What is the Grid? A Three Point Checklist. GRIDToday 1, 32–36 (2002)
5. AutoDock, http://autodock.scripps.edu/
6. Gromacs, http://www.gromacs.org/
7. GridCafe, http://www.gridcafe.org/
8. Welch, V., Siebenlist, F., Foster, I., Bresnahan, J., Czajkowski, K., Gawor, J., Kesselman, C., Meder, S., Pearlman, L., Tuecke, S.: Security for Grid Services. In: Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing 2003, pp. 48–57 (2003)
9. Németh, C., Dózsa, G., Lovas, R., Kacsuk, P.: The P-GRADE Grid Portal. In: Laganá, A., Gavrilova, M.L., Kumar, V., Mun, Y., Tan, C.J.K., Gervasi, O. (eds.) ICCSA 2004. LNCS, vol. 3044, pp. 10–19. Springer, Heidelberg (2004)
10. Von Laszewski, G., Foster, I., Gawor, J., Lane, P., Rehn, N., Russell, M.: Designing Grid-based Problem Solving Environments and Portals. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences 2001, p. 10 (2002)
11. Novotny, J., Tuecke, S., Welch, V.: An Online Credential Repository for the Grid: MyProxy. In: Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing 2001, pp. 104–111 (2002)
12. Cai, Y., Cao, J., Li, M., Chen, L.: Portlet-based Portal Design for Grid Systems. In: Fifth International Conference on Grid and Cooperative Computing Workshops, 2006. GCCW 2006, Hunan, pp. 571–575 (2006)
13. Vine Toolkit, http://vinetoolkit.org/
14. P-GRADE Grid Portal, http://portal.p-grade.hu/
15. gUSE - grid User Support Enviroment, http://www.guse.hu/
16. Kacsuk, P.: WS-PGRADE Portal and Its Potential Use in Grid Malaysia and in Cloud. Knowledge Sharing Session - MIMOS. Kuala Lumpur, Malaysia (2010)
17. Liferay, http://www.liferay.com/

Assessing the Use of Mash-Ups in Higher Education

Rabiu Ibrahim and Alan Oxley

Department of Computer and Information Sciences
Universiti Teknologi PETRONAS
31750 Tronoh, Perak, Malaysia
brancyringim@yahoo.com, alanoxley@petronas.com.my

Abstract. Today mash-ups have been accepted by many users and by many organizations. The development of mash-ups is everywhere, yet until now there has been insufficient research and development of mash-up applications for use in Higher Education. In this work we focus on mash-ups in Higher Education: their potential, how to provide training for users, and how to assess their effectiveness in the dissemination of information. A mash-up architecture consists of three main components: first is the content/service provider, second is the provider of the editor used to create the mash-up, and lastly is the server to host it and the client to whom the mash-up results are presented. In this work we investigate the available mash-up editors from which we chose one; we used the selected mash-up editor for the development of a mash-up application at University Teknologi PETRONAS.

Keywords: Data mash-up, Higher Education, Library, Mash-up editor, Development Methodology, Users, Web 2.0, Social, Media.

1 Introduction

Mash-ups are Web 2.0 applications which use and reuse data and services, which are accessible on the Web, in combinations. Moreover, mash-up applications are developed rapidly and in an ad-hoc fashion [1][2]. Mash-ups can be regarded as something similar to a simple Service Oriented Architecture (SOA) application. Both integrate information and this can be either locally produced information or information available on the Web.

A study by [3] gives an overview of the mash-up phenomenon; we now describe a few of the issues mentioned there. The motivation for mash-up development occurs due to the fact that development can be made quickly, easily, and affordably. On this last point, application development can be done by making a new application by reusing an existing application, which has already been developed by another user and which has been tested and paid for. Mash-ups use a large amount of data feeds, API's and other platform services that exist over the Web. For mash-up development, the level of Web platform does not really matter since mash-up applications use a basic level of Web platform for the universal exchange of data and knowledge. A user can make use of an existing Application Programming Interface (API), the availability of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 278–291, 2011.

which has facilitated the growth of mash-ups; also, users can simply design and implement a new API if a suitable one is not available. Mash-ups are a powerful technique and means of accessing the nearly unlimited amount of data and services that are available on the Web. In addition, the development and the enhancement of mash-up tools used for building mash-up applications is impressive. They have reached a level of usability whereby normal Internet users, i.e. non technical users, can use the tools to create their own mash-up solutions.

Let us consider the mash-up architecture which is presented in Figure 1; this shows that mash-ups are capable of combining different data and services from different sources, irrespective of whether the data or service is local or remote, as long as there is the right to access the data or services. Web services, local or remote database files, Really Simple Syndication (RSS) feeds, platform services like Google Maps, Yahoo! Maps, flat files, and other services are the sources of data for mash-ups. Existing Web widgets and badges are also utilized by most of the mash-up developers.

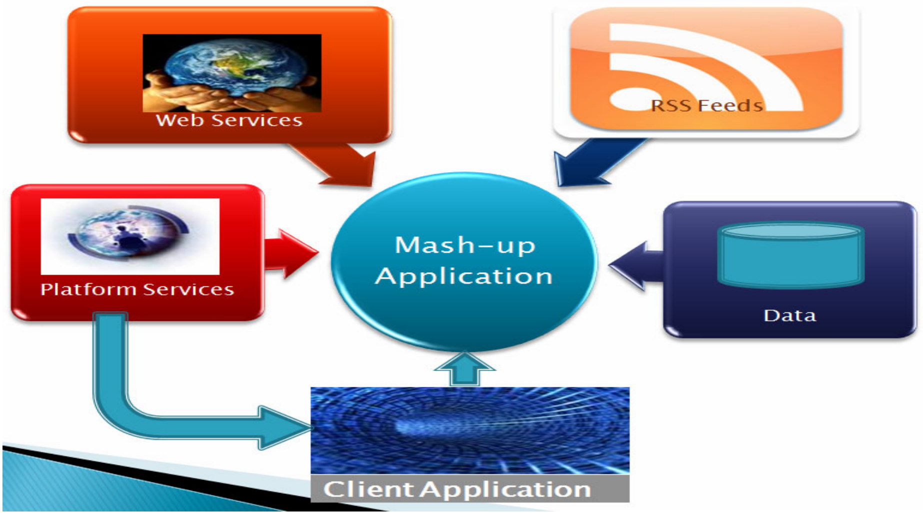


Fig. 1. Mash-ups Architecture: Base on [4]

Mash-up data sources are classified into three different categories [6]. The categories are as follows:

- Online data resources whose underlying data cannot be controlled by users and where access to the data requires special software that is available from the resource owner. For example, the data held by eBay, Flicker, Google can be accessed using the appropriate API.
- Local and other files that belong to a user and, therefore, can be controlled and accessed by the user.
- Online data resources whose underlying data cannot be controlled by users and where access to the data required is readily available. Websites and RSS feeds are examples of this data category. The data is publicly available.

Some general terms relevant to the discussion of mash-ups are ‘combination’ (also called ‘aggregation’) and ‘visualization’. We can also model a mash-up as a set of layers: the data layer, the application layer, and the presentation layer [6]. A mash-up combines multiple data source services and applications into a single application. The data layer is being referred to here. Visualizing the data with the aid of a User Interface (UI) can be thought of as the presentation layer.

Grouping data and manipulating it, with varying functions, usually happens in the application layer and can be done using the powerful aggregation features of mash-ups. Mash-ups are applications that are situational. Mash-ups are used by the ‘long tail’, a place in which there is a huge amount of unmet demand. In the ‘long tail’ the traditional IT cost structure for an application is no longer applicable [7].

2 Review

Data mash-ups in education and research are part of an emerging, richer information environment with greater integration of mobile applications, sensor platforms, e-science, mixed reality, and semantic, machine-computable data [11].

“Mash-up development is a promising End User Development (EUD) application area for several reasons”[5]. Web technology is considered as an essential platform which provides access to the rising number of publicly available services and allows developers to find the sophisticated services, then to develop a UI to which these services are joined together and presented to the user of the mash-up. These processes require a basic level of technical know-how in programming and are supported by tools. Demand guided by rapid, opportunistic development of situated applications, with short lifetimes, intended for small audiences are the basis for the development model. Furthermore, the Web is a platform that facilitates community building and mash-up sharing.

Too many different explanations of mash-ups have resulted in some confusion regarding the term ‘mash-up’ and its use [3]. Nevertheless, the commonly used definition of mash-ups classifies them into two divisions: Web mash-ups and enterprise mash-ups. The Web mash-up comprises of only data and people, while the enterprise mash-up is designed to integrate data and people for use by a business process.

Currently many available mash-up editors are only for business purposes. There are others that are used mostly for fun. It is possible to develop an application for Higher Education (HE) and libraries by using some of these non-business editors. There is a clear cut reason for classifying mash-up editors according to their purpose and context. [2] is an investigation of mash-up editors, and we now discuss some of the issues it raises. The study classifies editors into three groups. First, an editor that retrieves data from one or more data sources, processes the data and, lastly, publishes the result into a Web feed, or widget, is classified as an Information mash-up editor. Second, an editor that automates processes by orchestrating services, forms and other resources in a workflow, and often includes data entry, is classified as a Process mash-up editor. Third, a Web Site Customization editor is one that allows a user to mash-up by customizing a Web page; this process includes removing elements, adding additional widgets and changing the UIs of websites.

In this study the methodology for using information mash-up editors to develop mash-ups for HE and libraries (educational mash-ups) is used. There are a few things that we consider while working with mash-ups for HE - things like the commonly available mash-up editors, the community of practices, and other related issues. Below are shown the topics considered during this study, detailing some of the many aspects under consideration.

2.1 Yahoo Pipes

According to the Wikipedia definition:

> *Yahoo! Pipes is a web application from Yahoo! that provides a graphical user interface for building data mash-ups that aggregate web feeds, web pages, and other services, creating Web-based apps from various sources, and publishing those apps. The application works by enabling users "pipe" information from different sources and then set up rules for how that content should be modified (for example, filtering)* [8].

2.2 An Open University Blog

Tony Hirst writes: "OUseful.info is a blog in part about… things that I think may be useful in an higher education context, one day…" [9]. In this blog there are several examples of how to make mash-ups in Google or Yahoo! using pipes; one mash-up allows a user to retrieve information concerning his/her studies, another mash-up allows a user to map the school location.

2.3 Library LookUp

The Library Lookup project started in December 2002 as five lists of bookmarklets for libraries using these catalog systems: Innovative, Voyager, iPac, DRA, and Talis. The system allows a user to create a link using a bookmarklets generator. (There are some links already available in the system.) The user can drag and drop links [10].

2.4 A Study to Find the Best Social Media HE Mash-Up

One study which was undertaken to identify the uses of mash-ups in HE, came up with what it regarded as the best social media mash-up for HE [10]. This study considered a number of mash-up applications. The study's author, Power, commented that although there are plenty of mash-up applications for HE, he favours few. This is what he has to say:

> *"The hardest part of any cohesive social media campaign is pulling it all together.*
>
> *It's why I'm so impressed when colleges or universities embrace the social stream and preset it on their own terms in a creative and meaningful way. I'm not talking about social media directories where a school lists all its accounts. I'm talking about a high-quality mash-up where colleges and universities wrangle feeds from blogs, Facebook, Twitter, YouTube, Flickr, and more to create a compelling page that gives a real-time snapshot of all an institution has to offer. (Hat tip to the folks over at mStonerblog; they have been talking about mash-ups for months.)".*

Some example of mash-up applications for HE considered by the study are now described.

2.4.1 University of Maryland Baltimore County

The social stream was broke down by the UMBC mash-up in more different ways than anticipated. Surely, this is a sort of a profound display of all that is taking place within the campus and likewise a profound resource in connecting both the students and organizations based on activity. Photos, tabs, tweets, videos, blogs, music, and organizations are there for students. The page defaults to everything in case someone needs it all as shown in figure 2.

2.4.2 Missouri State University

Figure 3 present the Missouri State University mash-up. The State University of Missouri compiles its social stream in a position of excellent prominence-openly on top of the University's home page. The stream was drawn from university-based Facebook accounts, Twitter feeds, and news sends to fill up the center (News and Events) column. This is quite an enormous example the way the school is embarking on social media whilst giving new substance to the home page all over the course of the day.

2.4.3 Tufts University

A comprehensive effort of drawing together various feeds into a single interface is the social media hub at Tufts University shown in figure 4. Here are tabs such as YouTube, Twitter, LinkedIn, Facebook, Flickr, and select university blogs. The most impressive part of it is, Tufts didn't simply draw the available widgets from every site, but got the code customized to match the appearance and feel on the site already.

2.4.4 Savannah College of Art and Design

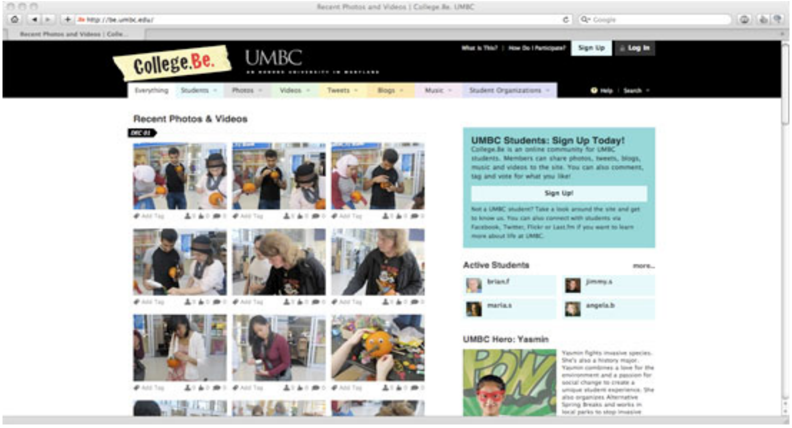
The University for Creative Careers, that's the very motto The Savannah College of Art and Design bills itself. No wonder the school's social media mash-up is one of the most prevailing and creative out there see figure 5. The University sets its blog, Twitter, Flickr, and YouTube feeds into a 12-box slider that draws pictures, text, and video into a fascinating display.

2.4.5 College of William and Mary

The College of William and Mary's social stream is so user-friendly and visually captivating, actually due to its simplicity as in figure 6. There are six boxes that feed the latest from the stream, being it a blog post, uploaded photo, or tweet. Visualizing preceded posts over the stream is as simple as choosing the numbers down the left side.

2.4.6 Vanderbilt University

The social media mash-up of Vanderbilt University makes it simple. Its Twitter-esque stream that's easy to follow and scan was produced by the blog, YouTube, and Twitter feeds. Moreover, in the tabs across the top of the page are the social media options. It would be better if users could access this stream with just a single click, see figure 7.



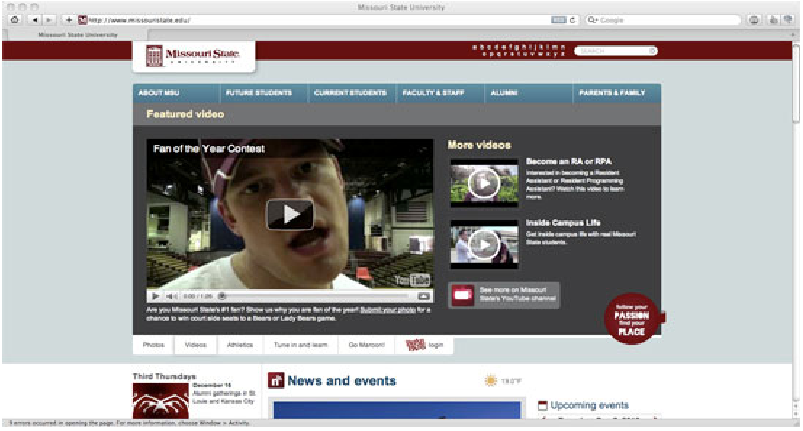
UMBC Students: Sign Up Today! College.Be is an online community for UMBC students. Members can share photos, tweets, blogs, music and videos to the site. You can also comment, tag and vote for what you like!

Sign Up!

Not a UMBC student? Take a look around the site and get to know us. You can also connect with students via Facebook, Twitter, Flickr or Last.fm if you want to learn more about life at UMBC.

Yasmin fights invasive species. She's also a history major. Yasmin combines a love for the environment and a passion for social change to create a unique student experience. She also organizes Alternative Spring Breaks and works in local parks to stop invasive

Fig. 2. University of Maryland Baltimore County mash-up interface



Interested in becoming a Resident Assistant or Resident Programming Assistant? Watch this video to learn more.

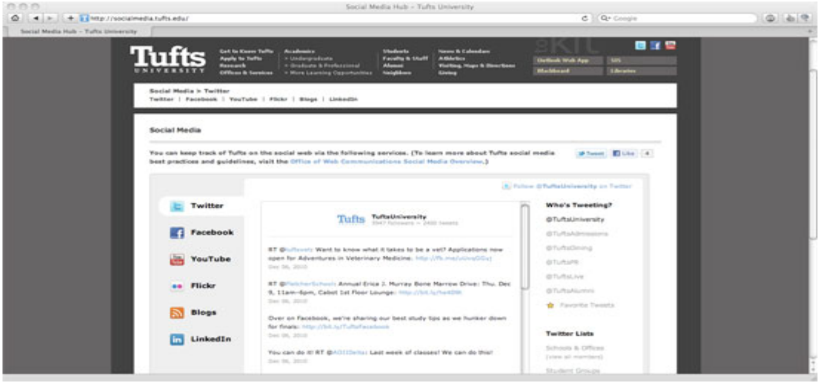
Get inside campus life with real Missouri State students.

Are you Missouri State's #1 fan? Show us why you are fan of the year! Submit your photo for a chance to win court side seats to a Bears or Lady Bears game.

December 15 Alumni gatherings in St. Louis and Kansas City

For more information, choose Window > Activity.

Fig. 3. Missouri State University mash-up application interface



You can keep track of Tufts on the social web via the following services. (To learn more about Tufts social media best practices and guidelines, visit the Office of Web Communications Social Media Overview.)

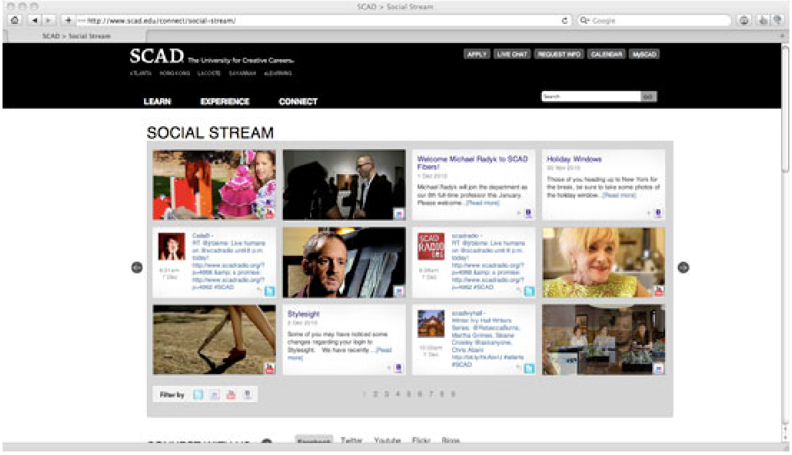
RT @tuftsvets: Want to know what it takes to be a vet? Applications now open for Adventures in Veterinary Medicine. http://fb.me/[illegible]
Dec 06, 2010

RT @FletcherSchool: Annual Erica J. Murray Bone Marrow Drive: Thu. Dec 9, 11am–6pm, Cabot 1st Floor Lounge: http://bit.ly/[illegible]
Dec 06, 2010

Over on Facebook, we're sharing our best study tips as we hunker down for finals: http://bit.ly/TuftsFacebook
Dec 06, 2010

You can do it! RT @ADIIDelta: Last week of classes! We can do this!
Dec 06, 2010

Fig. 4. Tufts University mash-up interface



Welcome Michael Radyk to SCAD Fibers!

Michael Radyk will join the department as our 8th full-time professor this January. Please welcome... [Read more]

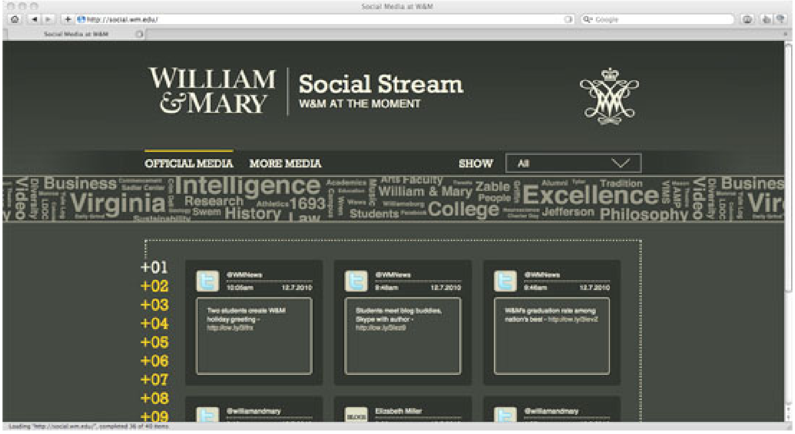
Holiday Windows

Those of you heading up to New York for the break, be sure to take some photos of the holiday window... [Read more]

Stylesight

Some of you may have noticed some changes regarding your login to Stylesight. We have recently ... [Read more]

Fig. 5. Savannah Collage of Art and Design



@WMNews 10:09am 12.7.2010
Two students create W&M holiday greeting - http://ow.ly/3ifrx

@WMNews 9:48am 12.7.2010
Students meet blog buddies, Skype with author - http://ow.ly/3iez0

@WMNews 9:46am 12.7.2010
W&M's graduation rate among nation's best - http://ow.ly/3iew2

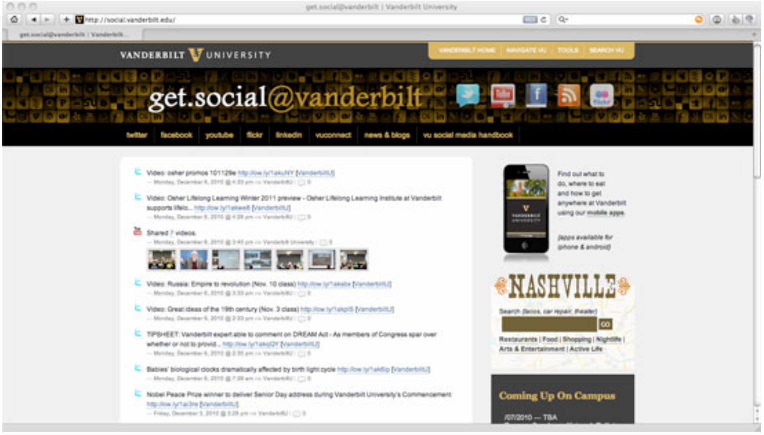
@williamandmary

Elizabeth Miller

@williamandmary

Fig. 6. College of William & Mary mash-up interface

Power identifies the above six mash-up applications as his favourites [10]. The author looked into the use of social media in HE. The examples of mash-up applications developed by the students mostly involve social media sites such as YouTube and Twitter. However, our research is focusing on general applications in HE.



Video: osher promos 101129e http://ow.ly/1akuNY [VanderbiltU]
— Monday, December 6, 2010 @ 4:33 pm via VanderbiltU | 0

Video: Osher Lifelong Learning Winter 2011 preview - Osher Lifelong Learning Institute at Vanderbilt supports lifelo... http://ow.ly/1akwe8 [VanderbiltU]
— Monday, December 6, 2010 @ 4:28 pm via VanderbiltU | 0

Shared 7 videos.
— Monday, December 6, 2010 @ 3:40 pm via Vanderbilt University | 0

Video: Russia: Empire to revolution (Nov. 10 class) http://ow.ly/1akabx [VanderbiltU]
— Monday, December 6, 2010 @ 3:33 pm via VanderbiltU | 0

Video: Great ideas of the 19th century (Nov. 3 class) http://ow.ly/1akpl5 [VanderbiltU]
— Monday, December 6, 2010 @ 2:33 pm via VanderbiltU | 0

TIPSHEET: Vanderbilt expert able to comment on DREAM Act - As members of Congress spar over whether or not to provid... http://ow.ly/1akqQY [VanderbiltU]
— Monday, December 6, 2010 @ 2:35 pm via VanderbiltU | 0

Babies' biological clocks dramatically affected by birth light cycle http://ow.ly/1ak6ip [VanderbiltU]
— Monday, December 6, 2010 @ 7:38 am via VanderbiltU | 0

Nobel Peace Prize winner to deliver Senior Day address during Vanderbilt University's Commencement http://ow.ly/1ai3re [VanderbiltU]
— Friday, December 3, 2010 @ 3:28 pm via VanderbiltU | 0

Find out what to do, where to eat and how to get anywhere at Vanderbilt using our mobile apps.

(apps available for iphone & android)

NASHVILLE

Search (tacos, car repair, theater)

GO

Restaurants | Food | Shopping | Nightlife | Arts & Entertainment | Active Life

Coming Up On Campus

/07/2010 — TBA

Fig. 7. Vanderbilt University mash-up interface

3 Methodology of This Research

We are looking at mash-up applications for use at Universiti Teknologi PETRONAS (UTP). We have come up with five areas of HE that have the potential to benefit from the application of mash-ups. The five areas are teaching and learning, the library, research, administration, and security. We are focusing on only three of these. These are teaching and learning, the library and research. We have identified categories of application which are suited to mash-up applications. We also wished develop a framework for mash-up applications.

Let us describe our work on the teaching and learning area. In the Information Technology programme at our university, one of the courses involves students studying Web 2.0 architecture. Mash-ups are a Web 2.0 architectural pattern. Teaching mash-ups to the students enrolled on the Web2.0 architecture course was one of our plans, as an initial step in investigating the use of mash-ups in teaching and learning. Due to the fact that the students were majoring in IT and were studying a course on Web 2.0 architecture we expected that they would be able to easily master the production of mash-ups. That is, the students are a special case as they are IT savvy and Web savvy.

We spent time thinking of a way to apply mash-ups in this HE teaching and learning setting and we formulated an appropriate strategy to carry it out. Part of our strategy was to design and conduct a sequence of mash-up tutorials to this class of students. The content of the tutorials had to be decided upon. We intended to teach the students the basic and most common ways of creating mash-up applications. We planned to teach the students how to construct mash-ups by using and reusing data that is stored either locally or remotely. The sessions were to take place in the computer lab and to made use of a well-known mash-up editor, Yahoo! Pipes. Initially, the students were to be given a lecture on Yahoo! Pipes in order to convey an understanding of the Yahoo! Pipes editor environment and the modules in it; the intention was to make sure that the students became familiar with the editor environment. We planned to demonstrate to the students how to create a simple mash-up application, using the Yahoo! Pipes editor, one that produced results. In subsequent sessions they were to work their way through a list of exercises at their own pace. These involved the students being asked to make a Web search for five educational mash-ups that might be useful in an educational setting and later, to develop five mash-ups of their own. Each of the mash-ups to be developed was to have specific features. We planned to have the students work individually, this was to ensure that each student participated and was creative.

The second application area is the library. A library is a huge repository of physical and electronic resources. The integration of data and the organization of data are the key activities of a library. Metadata, such as the location of an item and details about the item are key requirements for library users. A library consists of both physical and digital items, yet the accessibility of these different types of item remains an issue in our library today. Library users need to be able to speedily locate a resource; the information describing the resource on a computer screen needs to be appropriate, and the users need to be able to speedily access the resource. These aspects are concerned with resource discovery. However, there is another aspect, that of disseminating information to users, such as notifying users of the arrival of a new book. Mash-ups have a big role to play in the library. There is considerable scope for using mash-ups to organize, distribute, and integrate information on its resources. There is the possibility for mash-ups to become widespread throughout the library. We designed a simple model of how information can be discovered by the users, and how information can be disseminated to users.

We planned to make use of mash-ups to resolve some of the issues faced by the information resource centre (IRC) at UTP. We focussed on assisting users searching for online material, such as journal papers. Prior to our work, users had to search each subscribed-to database one at a time. We planned to alleviate this problem with the aid of a mash-up that automatically searches multiple databases. Work on this mash-up application started following a request by the one of the librarians.

The third application area is research. Postgraduate students can make use of mash-ups as a way of exploring, researching, discovering, analyzing and so on. Mash-ups are formed from collections of data. This can be remixed in different ways. Other possibilities include getting real time updates on his/her area of expertise, rapid data sample collection, and rapid access to research output/developments taking place throughout the world or easy access to other research materials. Mash-ups can be a good tool for postgraduates to utilize, since he/she can create the desired mash-up application. Postgraduates can tailor-make mash-ups to their own field of research. There are a few issues that we are currently looking into. One of these is the amount of basic training, if any, that needs to be given to the researcher or the postgraduate. In respect to this, we have been focusing on one individual at a time case. The motivation of the student to participate in the experiment is also an issue. In addition there are several other issues related to the effective deployment of mash-ups.

More generally this work we have identified four other categories of educational mash-ups, as follow:

- Frequent Web usage. There is potential for providing a more coordinated system for all of the services that a university student needs access to: email; course-related material; university-related material; library.
- Portal, wikis, online resources, and GPS. There is potential for providing the following: a Web portal containing event information, news, etc.; wikis that staff and students can contribute to; an on-line repository of research material created by students; GPS-based information such as the location of experts in the various research areas.
- Podcasts and radio. There is potential for providing the following: podcasts of lectures, warnings, guides, announcements, news, music, poetry; a radio system accessible over the Web by students which could also be a vehicle for promoting the university externally.
- Maps, Images, and Videos. There is potential for including all of these media in mash-ups.

4 Results

Following our review of mash-up applications, a small number of mash-ups applications were developed by us. Eleven of these new mash-up applications were tested and in all cases the Yahoo! Pipes mash-up application editor was found to be the editor with the greatest potential for our work. We found it to be well-suited, user-friendly and a tool that we consider one would wish to adopt.

Part of this work is also the construction of a mash-up development framework for HE see in figure 8 below. We recommend that development is done using the framework.

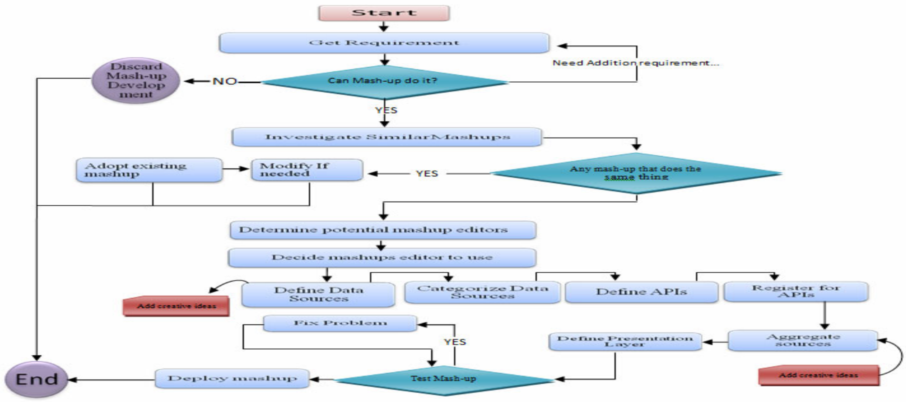
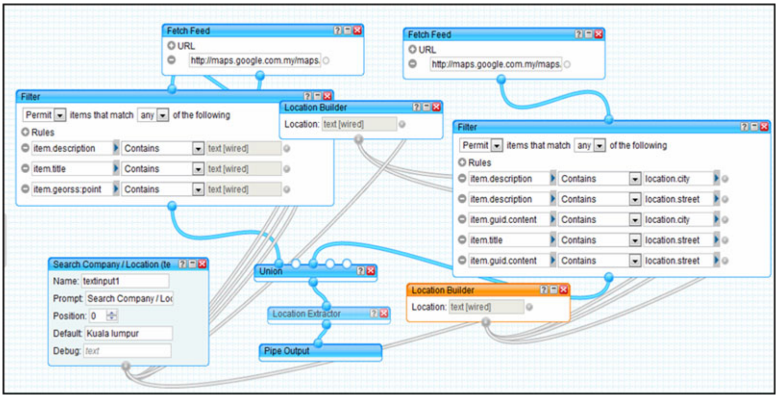


Fig. 8. Mash-up Development framework

As the research progresses, the potential for the use of mash-ups in HE is becoming clearer. We are, however, mainly focusing on the core activities of HE: teaching, research, and the library. Presently we have drilled down into each of these areas. As regards teaching and learning, we conducted tutorials with 16 students. We found that the result were impressive, with very encouraging performances by the students during the sessions. Among this student we thought few of them accept and adopt the concept of mash-up for their final year project, three of were selected for the Engineering Design Exhibition (EDX), one of the student present the Mash-up for Internship Placement, the creation of this mash-up is with our help and is totally using our lead model which is recommended in our research. The student takes the chance to demonstrate own experience and understanding of the concept of mash-up which we thought during our class. In figure 9 below shows the Mash-up for Internship Placement source / engine design while the figure 10a and 10b present the Mash-up for Internship Placement result to the user view in map this show the complete result that from between the range of Selangor and Kuala-Lumpor, while figure 10c present the search result base on the user input, this input can be the name of the company, location, or the type of placement in the company. Lastly is the figure 10d which present the result of this mash-up in a list view, user can select to view the result in map, list, or image in some mash-up applications.



Fetch Feed
URL: http://maps.google.com.my/maps.

Fetch Feed
URL: http://maps.google.com.my/maps.

Filter
Permit items that match any of the following
Rules
item.description Contains text [wired]
item.title Contains text [wired]
item.georss:point Contains text [wired]

Location Builder
Location: text [wired]

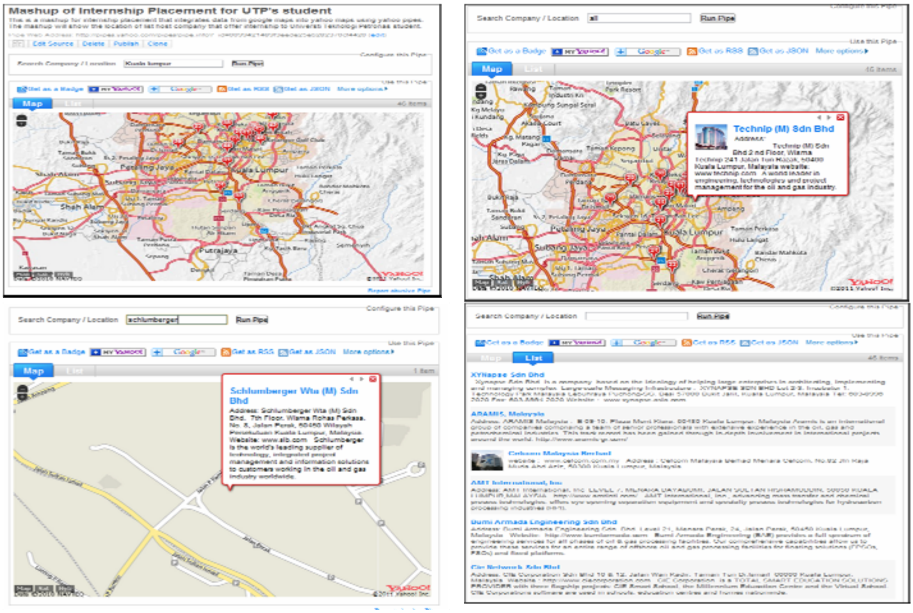
Filter
Permit items that match any of the following
Rules
item.description Contains location.city
item.description Contains location.street
item.guid.content Contains location.city
item.title Contains location.street
item.guid.content Contains location.street

Search Company / Location (te
Name: textinput1
Prompt: Search Company / Loc
Position: 0
Default: Kuala lumpur
Debug: text

Location Builder
Location: text [wired]

Fig. 9. Mashup engine using Yahoo Pipes widget

We developed a mash-up application using Yahoo! Pipes to ease accessibility of online material, such as journal papers. Now the UTP IRC users do not have to go through each of the subscribed-to databases one at a time. A screenshot from the library mash-up which we developed is shown in figure 11. The system has currently been tested by many users in our academic department and so far there are positive results, which help with the progress of our research.



Mashup of Internship Placement for UTP's student

This is a mashup for internship placement that integrates data from google maps into yahoo maps using yahoo pipes. The mashup will show the location of list host company that offer internship to Universiti Teknologi Petronas student.

Search Company / Location: Kuala lumpur [Run Pipe]

Map | List — 40 items

Search Company / Location: all [Run Pipe]

Map | List — 46 items

Technip (M) Sdn Bhd
Address: Technip (M) Sdn Bhd 2 nd Floor, Wisma Technip 241 Jalan Tun Razak, 50400 Kuala Lumpur, Malaysia website: www.technip.com A world leader in engineering, technologies and project management for the oil and gas industry.

Search Company / Location: schlumberger [Run Pipe]

Map | List — 1 item

Schlumberger Wta (M) Sdn Bhd
Address: Schlumberger Wta (M) Sdn Bhd. 7th Floor, Wisma Rohas Perkasa, No. 8, Jalan Perak, 50450 Wilayah Persekutuan Kuala Lumpur, Malaysia. Website: www.slb.com Schlumberger is the world's leading supplier of technology, integrated project management and information solutions to customers working in the oil and gas industry worldwide.

Search Company / Location: [Run Pipe]

Map | List — 46 items

XYNapse Sdn Bhd
Xynapse Sdn Bhd is a company based on the ideology of helping large enterprises in architecting, implementing and managing complex Large-scale Messaging Infrastructure. XYNAPSE SDN BHD [illegible] Technology Park Malaysia [illegible] Kuala Lumpur, Malaysia Tel: [illegible] Website: www.xynapse-asia.com

ARAMIS, Malaysia
Address: ARAMIS Malaysia [illegible] Kuala Lumpur, Malaysia Aramis is an international group of companies combining a team of senior professionals with extensive experience in the oil, gas and petrochemical industries. [illegible]

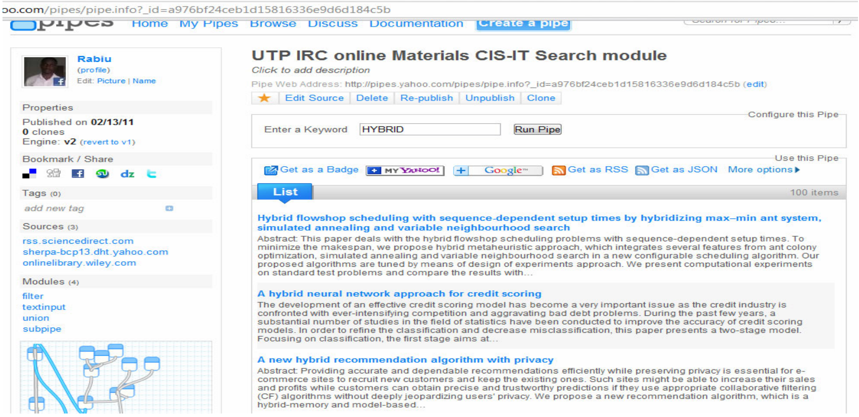
Celcom Malaysia Berhad
website: www.celcom.com.my Address: Celcom Malaysia Berhad Menara Celcom, No.82 Jln Raja Muda Abd Aziz, 50300 Kuala Lumpur, Malaysia

AMT International, Inc
Address: AMT International, Inc LEVEL 7, MENARA DAYABUMI, JALAN SULTAN HISHAMUDDIN, 50050 KUALA LUMPUR [illegible]

Bumi Armada Engineering Sdn Bhd
Address: Bumi Armada Engineering Sdn. Bhd. Level 21, Menara Perak, 24, Jalan Perak, 50450 Kuala Lumpur, Malaysia Website: http://www.bumiarmada.com Bumi Armada Engineering (BAE) provides a full spectrum of engineering services for all phases of oil & gas processing facilities. [illegible]

Cle Network Sdn Bhd
Address: [illegible] Kuala Lumpur, Malaysia Website: [illegible] is a TOTAL SMART EDUCATION SOLUTIONS PROVIDER [illegible]

Fig. 10. Mash-up for Internship Placement map and list view



oo.com/pipes/pipe.info?_id=a976bf24ceb1d15816336e9d6d184c5b

Rabiu
(profile)
Edit: Picture | Name

Properties
Published on 02/13/11
0 clones
Engine: v2 (revert to v1)

Tags (0)

Sources (3)
rss.sciencedirect.com
sherpa-bcp13.dht.yahoo.com
onlinelibrary.wiley.com

Modules (4)
filter
textinput
union
subpipe

UTP IRC online Materials CIS-IT Search module

Click to add description

Pipe Web Address: http://pipes.yahoo.com/pipes/pipe.info?_id=a976bf24ceb1d15816336e9d6d184c5b (edit)

Enter a Keyword: HYBRID [Run Pipe]

List — 100 items

Hybrid flowshop scheduling with sequence-dependent setup times by hybridizing max–min ant system, simulated annealing and variable neighbourhood search
Abstract: This paper deals with the hybrid flowshop scheduling problems with sequence-dependent setup times. To minimize the makespan, we propose hybrid metaheuristic approach, which integrates several features from ant colony optimization, simulated annealing and variable neighbourhood search in a new configurable scheduling algorithm. Our proposed algorithms are tuned by means of design of experiments approach. We present computational experiments on standard test problems and compare the results with…

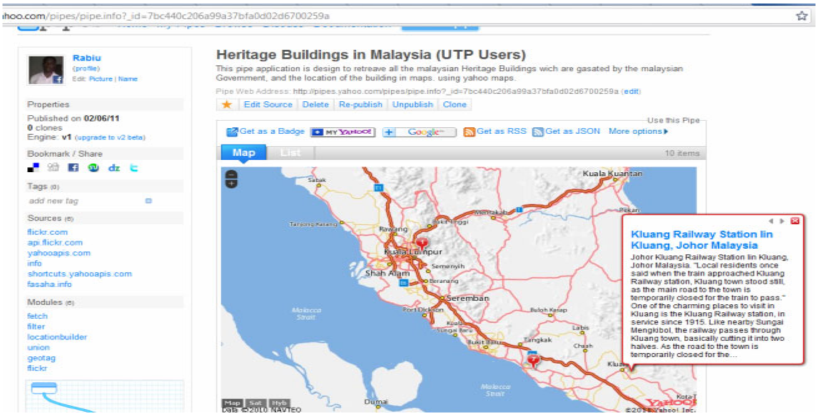
A hybrid neural network approach for credit scoring
The development of an effective credit scoring model has become a very important issue as the credit industry is confronted with ever-intensifying competition and aggravating bad debt problems. During the past few years, a substantial number of studies in the field of statistics have been conducted to improve the accuracy of credit scoring models. In order to refine the classification and decrease misclassification, this paper presents a two-stage model. Focusing on classification, the first stage aims at…

A new hybrid recommendation algorithm with privacy
Abstract: Providing accurate and dependable recommendations efficiently while preserving privacy is essential for e-commerce sites to recruit new customers and keep the existing ones. Such sites might be able to increase their sales and profits while customers can obtain precise and trustworthy predictions if they use appropriate collaborative filtering (CF) algorithms without deeply jeopardizing users' privacy. We propose a new recommendation algorithm, which is a hybrid-memory and model-based…

Fig. 11. UTP IRC mash-up interface

As regards the research area of application, we are experimenting with a few postgraduates at the moment; this has involved providing training to those participants. Our aim here is to try to assess their acceptance of mash-up applications, the areas of use, and how to teach them to develop their own useful mash-ups. This is a win-win situation as the work assists in our research yet, at the same time, the mash-ups help the researchers in their work. Currently we picked a few students from different departments to assist in carrying our work. These were all students studying different forms of engineering.

Hamid (not his real name) is a postgraduate from the Civil Engineering department currently working in heritage building facility management. He is the first postgraduate to be involved in our work. We took about 25 minutes to explain to Hamid what a mash-up is and how to develop a mash-up application. He has been able to develop a useful mash-up application. (See figure 12 for a screenshot of the mash-up application developed by this postgraduate.) Some more mash-ups could be explored with other postgraduates in the future.


hoo.com/pipes/pipe.info?_id=7bc440c206a99a37bfa0d02d6700259a

Heritage Buildings in Malaysia (UTP Users)

This pipe application is design to retreave all the malaysian Heritage Buildings wich are gasated by the malaysian Government, and the location of the building in maps. using yahoo maps.

Pipe Web Address: http://pipes.yahoo.com/pipes/pipe.info?_id=7bc440c206a99a37bfa0d02d6700259a (edit)

Rabiu (profile)

Properties
Published on 02/06/11
0 clones
Engine: v1 (upgrade to v2 beta)

Tags (0)

Sources (6)
flickr.com
api.flickr.com
yahooapis.com
info
shortcuts.yahooapis.com
fasaha.info

Modules (6)
fetch
filter
locationbuilder
union
geotag
flickr

10 items

Kluang Railway Station lin Kluang, Johor Malaysia

Johor Kluang Railway Station lin Kluang, Johor Malaysia. "Local residents once said when the train approached Kluang Railway station, Kluang town stood still, as the main road to the town is temporarily closed for the train to pass." One of the charming places to visit in Kluang is the Kluang Railway station, in service since 1915. Like nearby Sungai Mengkibol, the railway passes through Kluang town, basically cutting it into two halves. As the road to the town is temporarily closed for the...

Fig. 12. Heritage Buildings in Malaysia mash-up interface

5 Conclusions

Note that teaching and learning was the first activity that we engaged in; our motivation started from there after obtaining effective and impressive results from the class studying the Web2 .0 architecture course. We feel that this training of students on mash-ups could be extended to students of any discipline. An understanding of mash-ups is likely to be of benefit to other students on campus, particularly to those where maps form a part of the curriculum. In addition have few student to adopt our ideology of mash-up for their final year project (FYP) help us to see how effective is mash-up in the HE teaching and learning, at the same time we get impress that our effort to train the student is achieved in this section and it is fantastic.

For the library everything went perfectly. Here the users just benefit and do not contribute to mash-up development. We believe that this is only the beginning for the use of mash-ups in UTP IRC and we think that this is the start of a new era for UTP IRC.

Throughout all of our work we came across so many issues, especially with the research area of application. Nevertheless, we tried our best by making use of the available participants in the work, which has helped in making progress to date.

References

[1] Peenikal Mashups, S.: The Enterprise, White paper Mphasis (September 2009)

[2] Grammel, L., Storey, M.A.: An End User Perspective on Mashup Makers., University of Victoria Technical Report (September 2008)

[3] Ogrinz, M.: Mashup Patterns: Designs and Examples for the Modern Enterprise. Addison-Wesley Professional, Reading (2009)

[4] Clarkin, L., Holmes, J.: Enterprise Mashups. The Architecture Journal 13(5) (October 2007)

[5] Bolin, M.: End-User Programming for the Web. In: Department of Electrical Engineering and Computer Science, May 5, MIT, Cambridge (2005)

[6] Craig, L.: Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design. Prentice-Hall Inc., Englewood Cliffs (2002)

[7] Krill, P.: Serena to ship enterprise mashup platform. Infoworld, November 30 (2007), http://www.infoworld.com/d/developer-world/serenaship-enterprise-mashup-platform-632

[8] Wikipedia.com Definition of Yahoo Pipes, http://en.wikipedia.org/wiki/Yahoo!_Pipes (accessed on February 21, 2011)

[9] Tony Hirst, Open University, http://ouseful.wordpress.com/(accessed on 21/02/2011)

[10] Udell, J.: LibraryLookUp,
http://jonudell.net/udell/2006-01-30-further-adventures-in-lightweight-service-composition.html (accessed 10, 09, 2009)

[11] Patrick Powers marketing and web communications professional, higher education and lives in St. Louis, Mo.,
http://patrickpowers.net/2010/12/best-social-media-mash-ups-in-higher-education/
(accessed 26, 02, 2011)

[12] BattyAndrew, M., Hudson-Smith, C., et al.: JISC report, In: Data Mashups and the Future of Mapping, University College London, Centre for Advanced Spatial Analysis (CASA) and University of Nottingham, Centre for Geospatial Science (CGS) (September 2010)

Influence of Informal Visual Environments and Informal Motivational Factors on Learners' Aesthetic Expectations from Formal Learning Visual Environments

Sadia Riaz[1,*], Dayang Rohaya Awang Rambli[1], Rohani Salleh[2], and Arif Mushtaq[1]

[1] Computer and Information Sciences Department
Universiti Teknologi PETRONAS
Tronoh, Perak, Malaysia
[2] Management and Humanities Department
Universiti Teknologi PETRONAS
Tronoh, Perak, Malaysia

Abstract. Media aesthetic researchers believe that Informal Visual Environments (IVEs) have changed our aesthetic perceptions and made us perceptually selective by establishing a new schema (set of aesthetic expectations) based on information visualization. This may have caused learners' to have aesthetic expectations from Formal Learning Visual Environments (FLVEs). Likewise, learners' get cognitively fatigued when experience difference between what they aesthetically expect and what they see in a FLVE, this lowers their learning motivation. The purpose of this empirical study is to investigate how learners' aesthetic expectations from FLVEs are influenced by IVEs (Motion-Pictures, Video-Games, and Social Networking Websites) and Informal Motivational Factors (Challenge, Curiosity, Control and Fantasy). Keller's and Malone & Leppers' motivational models are used as research baseline to ascertain aesthetic expectations in formal and informal visual environments and how they jointly determine learners' aesthetic expectations from FLVEs. The study investigates four research questions by computing Two-Way Analysis of Variance, Pearson Correlation Coefficient and Multiple Regression Analysis. Results support the argument discussed in the paper and show a strong influence of both IVEs and Informal Motivational Factors on learners' aesthetic expectations from FLVEs.

Keywords: Aesthetic Expectations, Visual Aesthetics, Learning Motivation, Formal Learning Visual Environments, Informal Visual Environments.

[*] Sadia Riaz has received her MBA (IT) degree from Muhammad Ali Jinnah University, Islamabad (Pakistan) in 2002. She is associated with Board of Investment (BOI), Government of Pakistan as Director (Investment). At present she is pursuing Doctorate studies in Computer & Information Sciences from Universiti Teknologi PETRONAS, Malaysia. Her area of research is integration of media psychology to propose a model based on learners' aesthetic perceptions and motivation in formal and informal visual environments (email: missriaz@yahoo.com).

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 292–305, 2011.

1 Introduction

The digital age of today is rich in information and media aesthetics; it also requires us to be continual learners [1]. It is also largely believed that best form of learning is natural learning, which takes place as a process and not as a series of events [2]. This form of learning is best described as informal in nature. Informal learning occurs throughout our lives in a highly personalized manner and is based upon our particular interests, needs, motivation and past experiences [3]. According to researchers in the field of media aesthetics [4] viewers personalized interaction and viewing of visual aesthetics of informal digital environments has created a new schema. This is because television, motion-pictures and visual computer or screen displays are no longer considered as means of simple message distribution today by researchers but essential elements for communicating media aesthetics [4].

Edward Branigan [5] defines schema *as an arrangement of knowledge* already possessed by the perceiver to predict and classify new sensory data. When it is said that viewers personalized interaction with informal visual environments has formed a new schema, it means it has established a new set of aesthetic expectations. Ursyn [13] in article titled “Aesthetic Expectations for Information Visualization” has stressed upon raising the level of aesthetic designing by combining computer based information visualization techniques with principles of aesthetic designing and emphasized upon knowledge visualization by fulfilling aesthetic expectations of users.

Learners’ aesthetic expectations are formed through a cognitive process in which brain organizes, filters information and compares information with their past experiences (schema) in order to derive a meaning [6]. The filtration carried out by brain can also make learners’ perceptually selective in judging aesthetics of formal learning visual environment. This leads to formation of prejudiced aesthetic perceptions due to the contextual frame of reference [4] in which learners’ view and compare visual aesthetics of formal learning visual environments with that of informal visual environments. When there is a difference between *what learners aesthetically expect* and *what is there* in the visual design of formal learning visual environment, or if interaction involves a lot of cognitive work, they get “fatigued” which is known in some literature as “ego depletion”. This influences upon their learning motivation in formal learning visual environment.

Robins & Holmes [7] believe that design has an impact beyond decoration and research has already shown that user-interface of a learning environment has a very strong impact on the learning experience of the learners’ and the amount of knowledge to be retained [8]. This means learners’ aesthetic expectations in visual environments are related to the way they visualize information e.g., visual clutter, visual noise, colors, visual context [1].

There is lack of research on learners’ aesthetic perceptions that are formed due to their interaction with informal visual environments and its influence on their learning motivation in formal learning visual environments. For that matter, this empirical study is undertaken to investigate (a) learners’ set of aesthetic expectations from formal learning visual environments, (b) how that is rated against their favorite informal visual environment and informal motivational factor and (c) how it is reflected in their integration assessment of informal motivational factors into designing of formal learning visual environments to enhance their motivational and aesthetic appeal.

Moreover, this empirical study uses two motivational models as a research baseline, i.e. (1) Keller's Motivation Model and (2) Malone & Lepper's Motivation Model. These two motivational models share a certain degree of overlap as well in terms of their motivational variables. Motivational critiques Hardré [9] suggests that an integration of the two may provide an optimal instructional design model. Since no study was found in the literature that has integrated the two models, it is therefore also important to investigate how the two models jointly predict learners' aesthetic expectations in formal learning visual environments.

2 Research Objectives

- To investigate how learners' aesthetic expectations from formal learning visual environments are influenced by informal visual environments and informal motivational factors.
- To investigate how learners' aesthetic expectations from formal learning visual environments are jointly determined by informal and formal motivational factors.

3 Literature Review

Motivation is an emotion or a sense of feeling that captivates positive senses in our brain by enforcing a desired behavior [2]. Among educational researchers, motivation has become a buzzword and is considered to be critical in sustaining learners' involvement in the learning environment. Particularly, intrinsic motivation is considered to be more influential in directing self-regulated learning behavior, primarily because intrinsically motivated learners' are those who choose to participate in learning environments for no obvious compulsions, such as external stress, pressure or reward.

Moreover, in web-based learning environments, formal and informal learning may not be considered as completely distinct entities, but rather a part of single continuum. Therefore, in this empirical study we have identified two motivational models that essentially address motivation in view of characteristics of formal and informal visual environments.

3.1 Formal Learning Visual Environments (FLVEs)

As the name signifies, formal learning visual environment is formal in characteristics. It follows a prescribed schedule and defined learning objectives. It occurs formally of which learners' are aware of. Learners' are *'pushed'* towards learning in formal learning visual environment.

Keller's Model (Table 1) is appropriate to address the formality of the formal learning visual environment because this model adopts a ceremonial approach by viewing learning motivation from behaviorist perspective and has been validated by numerous studies, at different educational levels across different cultures [10]. Its motivational factors include: (1) Attention, (2) Relevance, (3) Confidence, and (4) Satisfaction.

3.2 Informal Visual Environments (IVEs)

The word informal means anything but prescribed, defined or formal. Informal visual environments are digital environments or visual media technologies that we use and interact with in our everyday life, anywhere, anytime and without any compulsions - thus, intrinsic motivation is higher. Learning also occurs in informal visual environments, but it is more or less discovery learning or incidental learning [11]. Some researchers even refer to it as learning while being not aware of it. Learners' are *'pulled'* towards informal visual environment that results in incidental learning.

As for the selection of motivational model, Malone & Lepper's Model [12] is based upon casual dimensions of learning motivation. It is best suited to meet the requirements of informal visual environments as they are casual in characteristics, e.g., multimedia based environments or other interactive visual environments. Its motivational factors include: (1) Fantasy, (2) Challenge, (3) Curiosity, and (4) Control.

Table 1. John Keller's Motivational Model

Attention	Attention is grabbed in a learning environment by using colors, creating novelty, providing interaction, generating participation, wittiness and sound effects.
Relevance	By providing realistic scenario, a meaningful contextual interpretation is created between the learner and the learning environment.
Confidence	Engagement provided by the learning environment tends to enhance learners' confidence level and proves to be a confidence-building experience for them.
Satisfaction	By accepting the benefits of learning environment and expressing aspiration to continue pursuing similar goals through it, indicates satisfaction on part of learners.

Table 2. Malone & Lepper's Motivational Model

Fantasy	Cognitive engagement to be provided by learners by making them experience situations in fantasy contexts that are not actually present, but intrinsically motivating.
Challenge	The difficulty of the activities to be performed by learners should be kept at an optimal level, otherwise they will get bored or frustrated.
Curiosity	To enhance sensory and cognitive curiosity in activities to be performed by learners, the environment may be designed as such to make learners believe that their current knowledge structure is incomplete, incompatible, or vague.
Control	The learning environment should promote a positive sense of control in learners, so that they are aware of the fact that their learning outcomes are dependent upon their own actions.

4 Research Questions

In order to achieve study objectives, the following research questions are designed.

RQ1: How do learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal visual environment and is reflected in their assessment of informal motivational factors for integration to make formal learning visual environments motivationally engaging?

RQ2: How do learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and is reflected in their assessment of the same for integration to make formal learning visual environments motivationally engaging?

RQ3: How do learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and reflected in their choice of informal visual environment?

RQ4: How the two motivational models, (1) Keller's ARCS model which has behaviorist formal learning motivational variables of Attention, Relevance, Confidence & Satisfaction and (2) Malone & Lepper's motivational model, which is based on casual or informal dimensions of Fantasy, Challenge, Curiosity & Control, jointly determine learners' aesthetic expectations from formal learning visual environments?

5 Methodology

The four research questions were investigated through a quantitative research method based on survey questionnaire.

400 copies of the questionnaire were hand distributed and emailed to undergraduate and postgraduate IT students at Universiti Teknologi PETRONAS (UTP) and Universiti Sains Malaysia (USM). The overall response rate of the questionnaire was 289 (72.25%) of which, 249 (86.1%) were usable as most items were adequately answered.

130 (52.46%) responses were from UTP, while 119 (47.79%) responses were from USM. The analysis was done using SPSS v. 11.

6 Results and Analysis

RQ1: How do learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal visual environment and is reflected in their assessment of informal motivational factors for integration to make formal learning visual environments motivationally engaging?

A Two-Way Analysis of Variance tested aesthetic expectations of the respondents who reported integration of informal motivational factors will make formal learning visual environments motivationally engaging or disengaging, and also indicated their choice of informal visual environment from the given three options of, (1) Social Networking Websites (SNWs), (2) Motion-Pictures, (3) Video-Games (Fig. 1).

Respondents who indicated that integration of informal motivational factors will make formal learning visual environment motivationally engaging, showed significantly higher aesthetic expectations from formal learning visual environments (F = 3.681 , p = .010, ŋ^2 = .029) than those who reported otherwise.

Aesthetic expectations from formal learning visual environments also differed significantly (F = 4.083, p = .002, ŋ^2= .038) across respondents who indicated their choice of informal visual environment (Social Networking Websites (SNWs), Motion-Pictures, Video-Games).

As Fig. 1 shows, respondents who opted for ‘Social Networking Websites’ as their favorite choice of informal visual environment reported highest level of aesthetic expectation from formal learning visual environments, followed by ‘Video-Games’ and ‘Motion-Picture’ adopters.

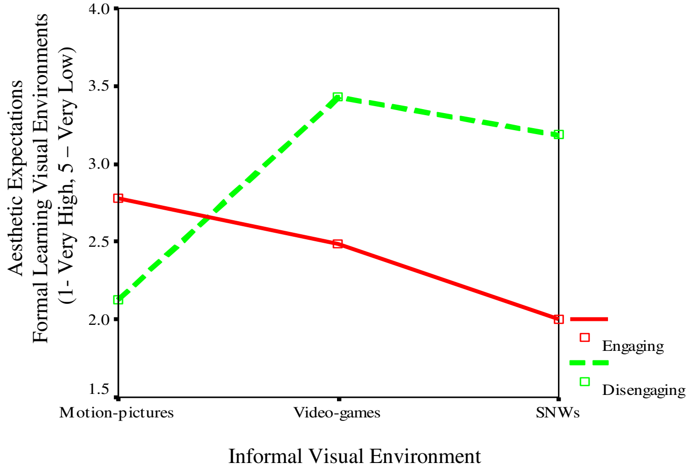


Fig. 1. Means Plot of Learners’ Aesthetic Expectations from FLVEs by their choice of Informal Visual Environment (IVE)

The aesthetic expectations pattern emerged similar across ‘Video-Games’ and ‘Social Networking Websites’ adopters. However, for ‘Motion-Picture’ adopters the pattern emerged in reverse form. This indicates respondents who reported integration of informal motivational factors into formal learning visual environments will make the later motivationally disengaging, reported higher level of aesthetic expectations than those who reported otherwise.

This reverse interaction of informal visual environment (Motion Pictures) with engagement also shared an interaction effect, which was significant (F = 6.880, p = .044, ŋ^2 = .094).

RQ2: How do learners’ set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and is reflected in their assessment of the same for integration to make formal learning visual environments motivationally engaging?

A Two-Way Analysis of Variance tested aesthetic expectations of the respondents who indicated integration of informal motivational factors will make formal learning visual environments motivationally engaging or disengaging, and also rated their favorite informal motivational factor from the given four options of, (1) Fantasy, (2) Challenge, (3) Curiosity (4) Control (Fig. 2).

Respondents who reported that integration of informal motivational factors will make formal learning visual environments motivationally engaging, depicted significantly higher aesthetic expectations from formal learning visual environments (F = 6.681 , p = .054, ŋ²= .017) than those who reported otherwise.

The aesthetic expectations pattern emerged similar across all four informal motivational factors, i.e., Fantasy, Challenge, Curiosity, and Control.

Aesthetic expectations for formal learning visual environments also differed significantly (F= 3.553, p=.000, ŋ²= .049) across respondents who indicated their choice of informal motivational factor.

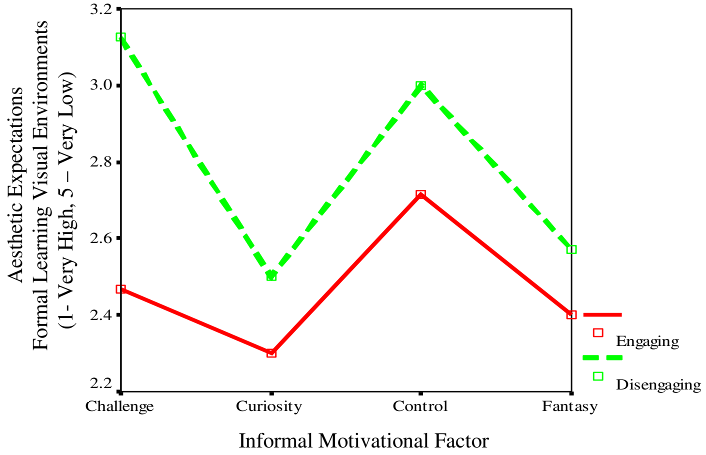


Fig. 2. Means Plot of Learners' Aesthetic Expectations from Formal Learning Visual Environments by their choice of Informal Motivational Factor

Informal motivational factor 'Curiosity' is the leading factor that sets high learners' aesthetic expectations from formal learning visual environments. This is closely followed by factor 'Fantasy', while factors 'Challenge' and 'Control' have smaller influences upon learners' aesthetic expectations in formal learning visual environments.

RQ3: How do learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and reflected in their choice of informal visual environment?

A Two-Way Analysis of Variance tested aesthetic expectations of the respondents who indicated their choice of informal visual environment from the given three options of (1) Motion-Pictures, (2) Video-Games and (3) Social Networking Websites and at the same time picked their choice of informal motivational factor from the given four options of (1) Fantasy, (2) Challenge, (3) Curiosity and (4) Control. See Table 3 for descriptive statistics and Fig. 3 for Means Plot.

Table 3. Mean and Standard Deviation of Informal Visual Environments and Informal Motivational Factors

Favorite Informal Visual Environment		**Favorite Informal Motivational Factor**			
		Challenge	***Curiosity***	***Control***	***Fantasy***
Motion-Pictures	Mean	2.60	2.17	3.67	2.43
	SD.	1.265	1.169	1.155	0.976
Video-Games	Mean	1.40	2.64	2.80	2.94
	SD.	0.548	0.924	1.229	1.289
Social Networking	Mean	2.75	2.00	2.33	3.00
Websites	SD	0.886	1.414	0.816	1.342

Levels of aesthetic expectations from formal learning visual environments differed significantly (F= 4.350, p=.038, ŋ^2= .138) across respondents who indicated their choice of informal visual environment (Social Networking Websites (SNWs), Motion-Pictures, Video-Games) and also picked their favorite informal motivational factor (Fantasy, Challenge, Curiosity, Control).

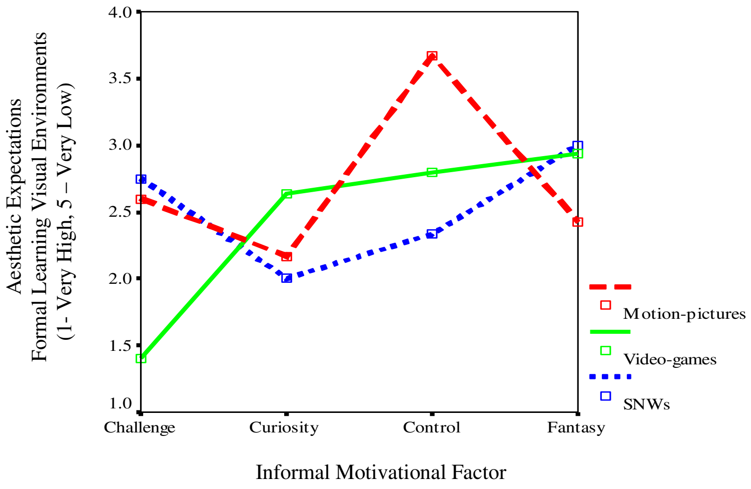


Fig. 3. Means plot of Learners' Aesthetic Expectations from Formal Learning Visual Environments by their choice of Informal Visual Environment and Informal Motivational Factor

Aesthetic expectations from formal learning visual environments were highest among respondents who picked 'Video-Games' as their favorite informal visual environment, while informal motivational factor ‘Challenge’ led this derive. Likewise, factor ‘Fantasy’ was found to be least tempting in the same visual environment.

This was followed by respondents who picked ‘Social Networking Websites’ (SNWs) as their favorite informal visual environment. This choice was led by informal motivational factor ‘Curiosity’, while factor ‘Fantasy’ was again found to be least motivating in the same visual environment.

Lastly, respondents who picked ‘Motion Pictures’ over ‘Video-Games’ and ‘SNWs’ as their favorite informal visual environment reported that informal motivational factor ‘Curiosity’ was behind their choice. Likewise, ‘Motion Picture’ adopters reported to be least motivated by informal motivational factor ‘Control’ in the same visual environment.

RQ4: How the two motivational models, (1) Keller’s ARCS model which has behaviorist formal learning motivational variables of Attention, Relevance, Confidence & Satisfaction and (2) Malone & Lepper’s motivational model, which is based on casual or informal dimensions of Fantasy, Challenge, Curiosity & Control, jointly determine learners’ aesthetic expectations from formal learning visual environments?

This research question was investigated by performing statistical procedure in two steps.

Step 1 – Pearson Correlation Coefficients of the eight motivational factors was computed to determine their association with aesthetic expectations from formal learning visual environments and to also ascertain their individual range and strength of association (Table 4).

Table 4. Pearson Correlation Coefficients of the Eight Predicting Motivational Factors with Aesthetic Expectations from Formal Learning Visual Environments

Motivational Factors	**Learning Motivation, *r***	**Sig. (2 tailed), *p***
Fantasy	.352	.002 **
Control	-.077	.042 *
Curiosity	.452	.004 **
Challenge	-.275	.001 **
Attention	.413	.009 **
Relevance	.383	.000 **
Confidence	.458	.000 **
Satisfaction	.211	.011 *
**denotes significance at the p < 0.01 *denotes significance at the p < 0.05		

Correlations go from zero (0), which indicates a non-linear relationship, to one (1) which indicates a perfect linear relationship and means everything falls exactly on the regression line. While positive and negative relationships are simply an indication whether it is an uphill or a downhill relationship or a direct or an inverse association. The Table 4 shows that all of these correlations are statistically significant at p < 0.01 or 0.05 levels, indicating they are reliably different from zero.

Motivational factor Confidence (ARCS) has a Pearson Correlation r = .458 which is a high positive value, depicts a strong correlation and indicates this motivational factor positively determines learners' aesthetic expectations from formal learning visual environments. This is followed by motivational factor Curiosity (Malone & Lepper) r = .452, Attention (ARCS) r = .413, Relevance (ARCS) r = .383, Fantasy (Malone & Lepper) r = .352. Motivational factor Satisfaction (ARCS) has the smallest but positive correlation, r = .211, which is again significant at $p < 0.05$.

Motivational factor, Control (Malone & Lepper) and Challenge (Malone & Lepper) have a negative correlation coefficient, with r = -.077 and r = -.275, respectively. Although, both of these correlations are negatively associated, they are still statistically significant at $p < 0.05$ and $p < 0.01$ respectively. This also indicates that higher motivational influence of factor Control and Challenge can even lower aesthetic expectations from formal learning visual environments.

Step 2 – The analysis technique known as Multiple Regression was used in Step 2 to determine how the two motivational models, (1) Keller's ARCS model, which is based upon behaviorist formal motivational variables of *Attention, Relevance, Confidence & Satisfaction* and (2) Malone & Lepper's motivational model, which is based upon casual or informal dimensions of *Fantasy, Challenge, Curiosity & Control*, jointly determine learners' aesthetic expectations from formal learning visual environments?

This analysis predicts values on a quantitative outcome variable, using several other predicting variables (Table 5)

Table 5. Multiple Regression Analysis of the Eight Motivational Factors Predicting Learners' Aesthetic Expectations from Formal Learning Visual Environments

R	R Square	Adjusted R Square	Std. Error of the Estimate
.805 (a)	.648	.634	.714

a Predictors: (Constant), Attention, Relevance, Confidence, Satisfaction, Challenge, Control, Curiosity, Fantasy

The value of Multiple Correlation Coefficient (R) between all eight predicting motivational variables and learning motivation of formal learning visual environments is 0.805. The maximum value of multiple correlation coefficients is 1, positive or negative and indicates correlation of all variables for predicting one single outcome, which in this case is 0.805, suggesting a strong relationship of all predicting motivational variables with aesthetic expectations from formal learning visual environments.

The value of the R^2 is a measure of how much of the variability in the outcome is accounted for by the predictors, which in this case are a combination of motivational factors given by Malone & Lepper and John Keller. Its value is 0.648, which means all predicting motivational variables approximately account for 65% of the variation in predicting learners' aesthetic expectations from formal learning visual environments.

The adjusted R² gives some idea of how well our model generalizes and the closer its value is to R², the better it is for our model. In this case, difference for the model is reasonable (0.648 - 0.634 = 0.014 or 1.4%). This shrinkage means that if the model was derived from the population rather than sample, it would account for approximately 1.4% less variance in the outcome.

Analysis of Variance tests whether the model is significantly better at predicting the outcome, than using the mean as a best guess. The F-result, labeled as regression in the below Table 6, is the ratio of improvement in prediction relative to the inaccuracy that still exists in the model, labeled as residual in the table. This model has an F-ratio = 68.350 which is highly significant at p <.001. Therefore, it can be said that model significantly improves our ability to determine learners' aesthetic expectations from formal learning visual environments.

Table 6. Analysis of Variance of the Eight Motivational Factors Predicting Learners' Aesthetic Expectations from Formal Learning Visual Environments

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	63.422	8	7.92	68.350	.000(a)
Residual	27.960	241	.116		
Total	91.382	249			

Finally, for this model, significant motivational factors that determine learners' aesthetic expectations from formal learning visual environments include, Fantasy t (241) = 5.477 at p<.001, Curiosity t (241) =3.497 at p<.01, Attention t (241) = 7.260 at p < .05, and Confidence t (241) = 2.667 at p <0.01. While motivational factors Relevance, Control and Satisfaction do not essentially contribute towards predicting the same and are found to be statistically insignificant (Table 7).

Table 7. Model Parameters for Predicting Learners' Aesthetic Expectations from Formal Learning Visual Environments

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	1.866	.334		5.654	.000
Attention	.835	.115	.502	7.260	.015
Relevance	.027	.068	.030	.393	.695
Confidence	.318	.119	.217	2.667	.006
Satisfaction	.322	.132	.157	2.439	.075
Challenge	-.714	.119	-.488	-6.022	.000
Control	-.136	.118	-.079	-1.153	.251
Curiosity	.550	.157	.303	3.497	.007
Fantasy	.905	.165	.516	5.477	.000

7 Discussion

This study investigated how learners' aesthetic expectations and learning motivation in formal learning visual environment are influenced by informal visual environments and informal motivational factors. In this regard, four (04) research questions were investigated.

RQ1 investigated how learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal visual environment and is reflected in their assessment of informal motivational factors for integration to make formal learning visual environments motivationally engaging. Results showed that learners' who indicated integration of informal motivational factors will make formal learning visual environments motivationally engaging, reported higher level of aesthetic expectations from formal learning visual environments than those who reported otherwise. This means informal visual environments have altered learners' acceptance threshold of visual aesthetics by making them perceptually selective. Moreover, influence of three informal visual environments, (1) Social Networking Websites, (2) Motion Pictures and (3) Video Games was investigated in this study. Results showed that Social Networking Websites have the highest motivational influence on learners' aesthetic expectations. This is because aesthetic designing of Social Networking Websites such as facebook, myspace, twitter etc., is based upon causal dimensions to encourage social informal learning. Social learning is learning with and through others. It also promotes learning as a pastiche of small chunks of observing how others do things, asking questions, trial and error, sharing stories with others and casual conversation. This makes aesthetic appeal of such websites high for learning purposes and sets aesthetic expectations higher for formal learning visual environments.

RQ2 investigated how learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and is reflected in their assessment of the same for integration to make formal learning visual environments motivationally engaging. Four informal motivational factors, (1) Fantasy (2) Challenge (3) Curiosity and (4) Control were investigated in the study. Results showed that Curiosity was the highest rated informal motivational factor in terms of setting learners' aesthetic expectations higher in formal learning visual environment. This factor remains at the heart of informal visual mediums which are meant for entertainment purposes. It stimulates cognitive learning process by evoking arousal and seeking attention through unusual and puzzling stimulus.

RQ3 examined how learners' set of aesthetic expectations from formal learning visual environments are rated against their choice of informal motivational factor and reflected in their choice of informal visual environment. Results showed that aesthetic expectations from formal learning visual environments were highest among respondents who picked Video Games as their favorite informal visual environment. Video Games remained favorite among respondents due to informal motivational factor Challenge. It is suggested that instructional designing of formal learning visual environments should apply and use psychology of entertainment used in Video Games. Research shows that students, who are expert computer game players, learn to process information more quickly than non-players. Moreover, it would be misleading to say that general characteristics of a video game do not fit into designing of formal learning visual environments. When playing a video game, the learner gets immediate feedback, experiences personalized situations to adjust to their unique learning styles, is allowed to adjust game settings to accommodate his/her learning needs and pace. Video Games, deploy informal motivational factors challenge, curiosity, fantasy, and control, which are all elements commonly found in a motivating learning environment.

RQ4 examined how the two motivational models, (1) Keller's ARCS model which has behaviorist formal learning motivational variables of Attention, Relevance, Confidence & Satisfaction and (2) Malone & Lepper's motivational model, which is based on casual or informal dimensions of Fantasy, Challenge, Curiosity & Control, jointly determine learners' aesthetic expectations from formal learning visual environments. It is believed that for aesthetic designing and optimizing learning motivation in formal learning visual environments, integration of informal and formal learning motivational factors is the best solution. The study results also supported and showed that a predicting model based on informal and formal motivational factors, significantly improved our ability to determine learners' aesthetic expectations from formal learning visual environments.

8 Conclusion

The study concludes that today's formal learning visual environments are designed without accounting for changes that have occurred in learners' aesthetic perceptions. This is presumably due to massive proliferation of digital technologies which are rich in media aesthetics. It has formed a new schema or a set of aesthetic expectations based on information visualization of digital environments. Due to the lack of experimental IS research on learners' aesthetic perceptions that are formed due to their interaction with informal visual environments and its influence upon their learning motivation in formal learning visual environments, the argument requires theoretical and empirical justification.

The study also concludes that learners' aesthetic expectations from formal learning visual environment are influenced by informal visual environments and informal motivational factors. Moreover, Keller's Motivational Model meant for formal learning visual environments and Malone & Lepper's Motivational Model intended for informal visual environments, when tested jointly, improved our ability to determine learners' aesthetic expectations for formal learning visual environments. This finding supports the argument discussed in this paper that informal visual environments do tend to influence upon learners' aesthetic perceptions and have created a new schema (set of aesthetic expectations) for formal learning visual environments. Formal learning visual environments will become cognitively engaging if the difference between what learners' 'aesthetically expect' is incorporated in aesthetic designing parameters of the environment, as this will have a positive influence on their learning motivation.

9 Future Works

The findings of this study will be used to propose an aesthetic perception and motivation model for visual learning environments. The proposed model will facilitate in aesthetic designing of formal learning visual environments by accounting for changes in learners' aesthetic expectations.

References

1. Riaz, S., Rambli, D.R.A., Salleh, R., Mushtaq, A.: Media Psychology & Ergonomics: Perceptual Visual Limitations to Human Cognition Factors in Aesthetic and Minimalist Designing of Web-based Learning Environments. In: Proceedings of the International Conference on Intelligence and Information Technology (ICIIT), Lahore, Pakistan, October 28-30, vol. 1, pp. 493–498 (2010)
2. Riaz, S., Rambli, D.R.A., Salleh, R., Mushtaq, A.: Study to Investigate Learning Motivation Factors Within Formal and Informal Learning Environments and their Influence upon Web-based Learning, vol. 5(4), pp. 1863–1383 (2010), doi:10.3991/ijet.v5i4.1338 ISSN: 1863-0383
3. Conlon, T.: A Review of Informal Learning Literature, Theory and Implications for Practice in Developing Global Professional Competence. Journal of European Industrial Training 28(2/3/4/), 283–295 (2003)
4. Zettl, H. (ed.): Applied Media Aesthetics, "Sight, Sound, Motion", 3rd edn., p. 4. Wadsworth Publishing (2008)
5. Branigan, E.: Narrative Comprehension and Film, Routledge, London (1992) as cited in Adrian Miles Hypertext Structure as the Event of Connection HT'01 8/01 Aarhus, Denmark. pp. 63, 2001. ACM ISBN 1-59113-420-7/01/0008
6. Riaz, S., Rambli, D.R.A., Salleh, R., Mushtaq, A.: Integrating Media Psychology Within a Theoretical Framework of Instructional Design for Web-Based Learning Environments (WBLEs). In: Proceedings of 5th International Conference on E-Learning, Penang, Malaysia, July 12–13, pp. 463–471 (2010)
7. Robins, D., Holmes, J.: Aesthetics and Credibility in Website Design. Information Processing and Management 44, 386–399 (2008)
8. Deubel, P.: An Investigation of Behaviorist and Cognitive Approaches to Instructional Multimedia Design. Journal of Educational Multimedia and Hypermedia 12(1), 63–90 (2003)
9. Hardré, P.: Designing Effective Learning Environments for Continuing Education. Performance Improvement Quarterly 14(3), 43–74 (2001), doi:10.1111/j.1937-8327.2001.tb00218.x
10. Keller, J.M.: What is a Motivational Design, Florida State University (2006), `http://www.pdf-finder.com/What-Is-Motivational-Design?1.html`
11. Hanley, M.: Introduction to Non-formal Learning. E-Learning Curve Blog (2008), `http://michaelhanley.ie/elearningcurve/introduction-to-non-formal-learning-2/2008/01/28/`
12. Malone, T., Lepper, M.: Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. In: Snow, R., Farr, M. (eds.) Aptitude, Learning, and Instruction: III. Conative and Affective Process Analysis, Lawrence Erlbaum, Hillsdale (1987)
13. Ursyn, A.: Aesthetic Expectations for Information Visualization. International Journal of Creative Interfaces and Computer Graphics (IJCICG) 1(1), 19–39 (2010)

Zero Watermarking for Text on WWW Using Semantic Approach

Nighat Mir

Computer Science Department
Effat University
nighat_mir@hotmail.com

Abstract. Information security can be achieved by different standard methodologies like Steganography, Cryptography and Digital Watermarking. In this research semantic watermarking approach has been used to offer security for online content. The proposed approach uses a secret key based on the idea of public key cryptography. The proposed method is developed by embedding semantic watermarking with a standard cryptographic method.

In the proposed model electronic web content is secured by a secret key. This secret key is composed of a semantic key and a unique author ID which is further encrypted by a cryptographic algorithm. Semantic key is generated on the basis of most frequently repeating letters in English language which are “th” and “wh” and a unique author ID is issued by the certification authority. Resultant of these two is further subjected to another layer of security and the secret key is encrypted using a cryptographic algorithm- Advanced Encryption Standard (AES). The system has been implemented using C# language and is further tested with different popular websites.

Keywords: Semantic, security, watermarking, cryptography, certification authority, web.

1 Introduction

In this paper I have discussed robust ways to protect and discourage the redistribution of online text. With a massive growth of Internet and its easy and low cost access to an author; Internet has attracted billions of writers. The growing factor of electronic publishing has somehow an effect on the print media. But in the meantime copyright protection of electronic text is becoming more and more elusive. Other then preventing unauthorized access to copy the content, a discouraging factor can also be focused and added to the web based data.

A unique ownership can be given to a web publisher so that it allows an author to determine the identity and rights to his/her content.

Objective of Steganography, Cryptography and Watermarking is to provide a security mechanism to the information, but behavior of each one is implemented in a different way. Steganograhy is an art which hides the message in such a way that only intended user or receiver can reveal the existence of hidden message. Where,

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 306–316, 2011.

cryptography seems more suspicious to an intruder with its understandable format. Watermarking is a mechanism of hiding information into a digital indicator to protect the copyright of an author.

2 Related Work

Different techniques have been studied for securing the text information based on different characteristics like image based, semantic based and syntactic based.

Some Image based algorithms for the text security have been studied in this section. Low N. F. and Brassil in [1] [2] have proposed Line shifting and Word shifting image based techniques which moves words vertically and horizontally. Line shifting changes the document by moving lines up or down based on the binary signal which is to be embedded. Where, Word shifting method moves words slightly to the left or right of their normal position. Young, Moon and Oh in [3] have proposed an algorithm which is based on the word classification and inter-word space that adds space in each line of text. The spaces are adjusted with respect to the sine wave of a particular wave and frequency. H. and Kot in [4] have proposed a watermarking method for owner authentication based on inter characters and words spaces. Algorithm utilizes the right and left spaces of a document.

In earlier periods, Syntactic approach has been used for the text watermarking based on the structure of a text. Atllah [5] has used Text Meaning Representation (TMR) methods for embedding a watermark in a document. Hasan in [11] has proposed morpho-syntactic tools for text watermarking by performing these alteration to the Turkish language text. The unmarked text is first transformed into a syntactic tree diagram where the syntactic hierarchies and the functional dependencies are coded. The watermarking software then operates on the sentences in syntax tree format and executes binary changes under control of Word-net to avoid semantic drops. Hassan and Mohammad in [7] have proposed a feature coding algorithm in which embedding can be achieved by performing a vertical displacement of points in Persian and Arabic languages.

Semantic methods are considered more secure against retyping of text but sometimes they may change the meaning of a text. Nighat and Afaq in [8] have proposed a Synonyms and Acronyms based technique for XML files to prove the intellectual ownership and to achieve the security. XML is taken as an input object to be watermarked on sender side and Tags are constructed according to the manipulated Synonyms List (SL) or Acronyms List (AL), each one at a time. XML file is validated and translated by HTML and the relative information is displayed on the browser. Linguistic based algorithms [9] have been discussed in terms of pre-supposition and transformation such as passivization, topicalization and preposing are used to embed the watermarks in a text document.

Web pages have a large capacity due to a great amount of available bandwidth and this can be utilized in an optimal way for hiding extra information like embedding a secret key. Many tools and applications are available for the web development but all browsers translate the codes into HTML format. As HTML is a basic component for the web development. There are different versions available for the HTML and is actually based on list of elements. Web pages must conform to the rules of HTML in order to be displayed correctly in a web browser [10].

Researchers have made attempts for hiding information using HTML and XML. Lachen and Sun in [10] have proposed different techniques for web page watermarking. They have proposed techniques based on the features of markup tags. Most of the techniques which have been proposed are by using white spaces, line breaks, attributes ordering, string delimiter and color values. Some of these have been tested to see the effect where some of them are not experimented.

Alaa, S. and A. in [11] have combined the HTML with cryptography and have used white spaces and have applied the cryptographic algorithm DES only for using color tag (value used for the colors).

Shingo and Kyoko in [12] have proposed some methodologies for hiding information using XML files. Techniques proposed are using empty elements, white spaces in tags, attribute and element ordering but these have not been implemented but recommended as a future work.

Different techniques for text steganography were proposed in [13] which exploit random character, reverse character, tags shuffling and attribute shuffling techniques for XML files. Mentioned techniques were implemented to show the results with respect to the security and the bandwidth. Nighat in [14] has used natural language digital watermarks for securing web based information. Several robust techniques of web page imperceptible digital watermarking using Verbs, Articles and Prepositions are studied and implemented for the protection of online content.

3 Proposed Methodology

In this research the copyrights for a web author has been proposed to be integrated based on some English language constructs and are usually called semantic based watermarks. The frequently occurring words are taken into consideration to generate a semantic key to protect the web authorization. The words which are used in this proposed methodology are from the list of most frequently occurring letters in English Language which is “th” and “wh”. These two letter occur as initials for many words e.g. for “th” there are (the, though, that, this, then, than, their, there, these, those, thank, through) and for “wh” there are (what, which, where, whether, who, when, while, whole, why, whom).

In this system only first two letter are considered so that every word starting with these two initials will contribute towards generating a secret key. The secret key used in this system is based on the mentioned words and all of these are taken into account while implementation.

Figure 1 shows the embedding phase of the proposed model where HTML files is taken as a carrier. HTML file is read and parsed to extract the natural watermarks, words starting with letters “th” and “wh” are used to generate the semantic key. During embedding a unique author ID is added to the semantic key to generate the secret key (Skey). This Skey goes through the encryption process where an encrypted SKey is finally taken and added to the HTML page to make it secure and now this page is ready to be communicated on Internet with an author secret key and author can prove the authentication to his/her text by this key. It is preferred that author uses the same key for different content similar to anyone’s name as an owner.

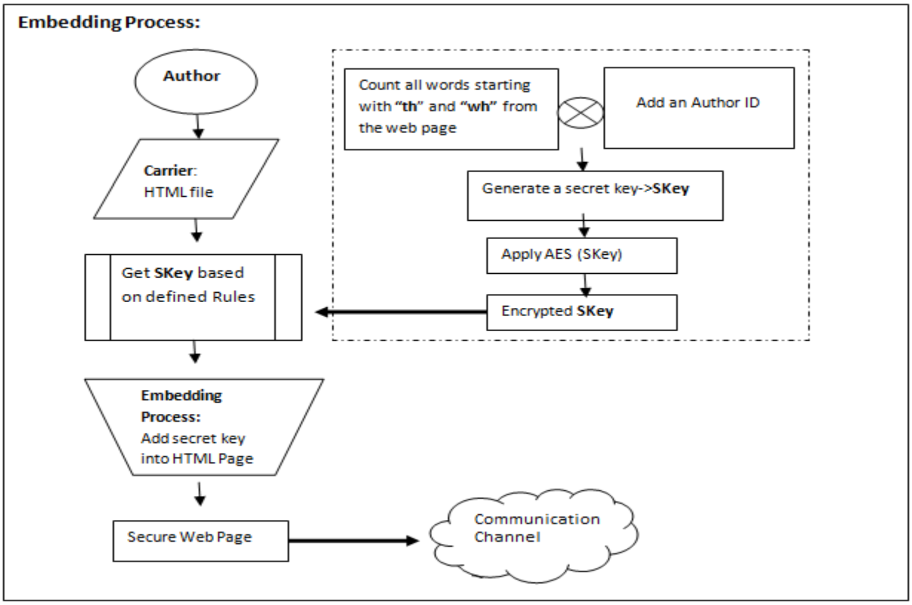

Embedding Process:
Author
Carrier:
HTML file
Get SKey based on defined Rules
Embedding Process:
Add secret key into HTML Page
Secure Web Page
Communication Channel
Count all words starting with "th" and "wh" from the web page
Add an Author ID
Generate a secret key->SKey
Apply AES (SKey)
Encrypted SKey

Fig. 1. System diagram of the proposed mechanism, where HTML files works as a carrier of the SKey which is encrypted using AES algorithm in an embedding process

Figure 2 illustrates the extraction phase of the proposed system, where a secured web page combined with the secret key is subjected to the system. Initially the secret key SKey will be decrypted and then will be verified against the natural semantic watermarks.

The extraction phase takes more time then the embedding as it needs to first check all possible words made up of these natural watermarks, then counting them to generate a number for each word is quite time consuming and prune to errors which may eventually lead to an incorrect key and conform to the robustness of the embedding phase.

The whole system has been implemented using C# language in Visual Studio.net framework. The system works accurately for the dynamic published website, static offline pages and a user can also create a test web page at the run time, following the rules of creating a web page. The program is tested on dynamic published websites for the robustness purposes and results are shown in this research. The working scenario starts with reading and parsing the information starting from the <body> tag till it finds the closing tag of </body>. So, every text in between the two tags is read and parsed based on the defined rules. The mentioned words act as natural watermarks which are combined with the unique author ID. The author ID can be a registered ID with a certified agency and this implemented system can also create one at the runtime. First a website address (URL) is provided to the program, it counts how many and what the watermarks (as defined) and then combines the unique author ID with those watermarks and then further it is encrypted using the AES algorithm to generate a cipher text for the corresponding secret key. Further, this ciphered secret

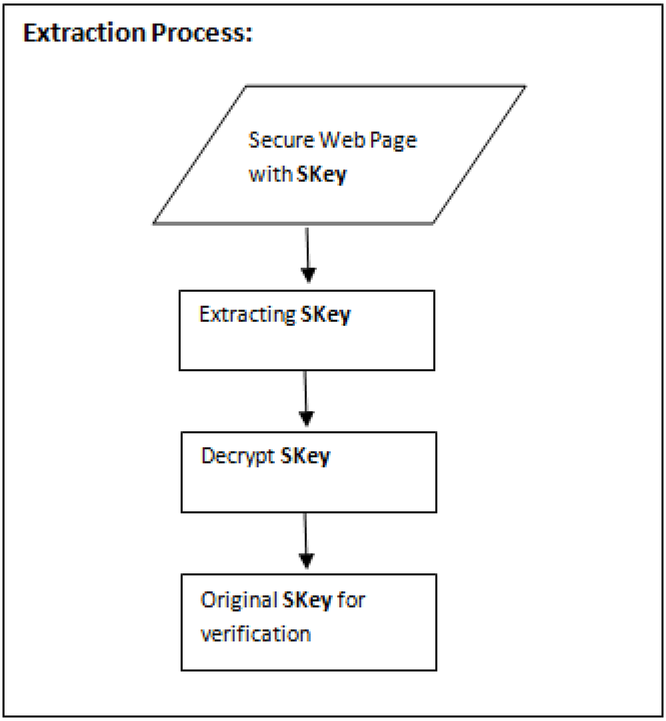


Fig. 2. This shows the extraction process where the SKey is recovered by the reverse of AES algorithm

key is added to the HTML page. The tag used to add this secret key is <meta> tag which is used to provide the meta data about the HTML document. The reason for choosing this element is based on the fact that information in this tag does not get displayed on the page by the browser and it is supported by all major browsers. Meta element is usually used to specify the page description, keywords, author details or modification dates. Equation 1 shows what elements are combined to get a key.

$$SKey = AES(\sum_{i=1}^{i=n}(th *. NoT + wh *. NoT + aID) \tag{1}$$

Where:

SKey = secret key
i = counter
th* = the, though, that, this, then, than, their, there, these, those, thank, through
wh* = what, which, where, whether, who, when, while, whole, why, whom
NoT = number of times
aID = author identification number

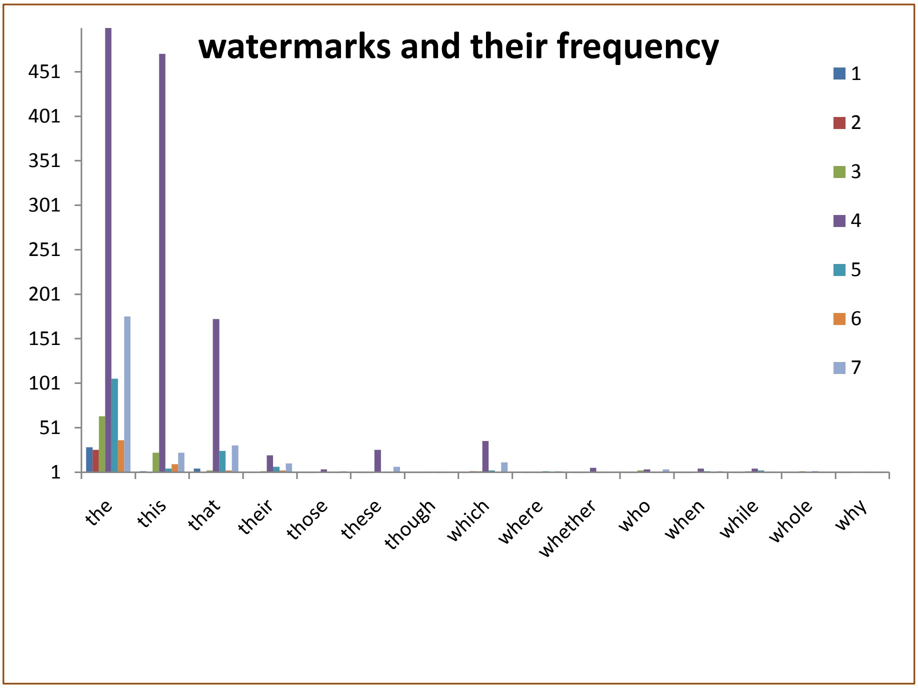
Equation (1) explains the working mechanism of generating the secret key and what components it is composed of.

The words used as watermarks are not considered or converted directly into cipher text as letter during the embedding phase but each letter is given a code during the implementation. The code is a decimal number e.g. if there are 7 “the” found in a web page and let assume that the code defined for “the” is 14, it will take a product of watermark into number of times it appears (14.7+key) and the resultant secret key is encrypted by AES into a cipher text before adding to the meta tag of a web page.

4 Experimental Results

During the testing phase different popular websites have been trained with the system. Table 2 shows the websites used in experiments. URL’s were subjected to the system and based on the defined rules, textual information was parsed and the corresponding watermarks were extracted. After the watermarks were extracted an author ID was given at the same time. For the uniformity purpose only one key has been used for all experiments. Key defined for these experiments has been set to a word “author”. The key has further been added to the extracted watermarks before encryption and the product of author ID and watermarks is encrypted using AES to generate a final secret key.

Graph 1: This shows a graphical view of watermarks and the pictorial frequency of each word



Graph1 shows the pictorial form of each watermark found in every website experimented and it also shows the frequency of occurrence for each word. Table 1 shows the actual numbers of occurrence for each watermark for different websites. Numbers are used to represent the website in Table 1 for the simplicity purposes and later the numbers are explained by the actual URL's in Table 2.

Table 1. This shows the frequency of every natural watermark extracted from different websites (from 1 to 7)

watermarks	***Websites(each number represents a website mentioned in Table 2)***						
	1	**2**	**3**	**4**	**5**	**6**	**7**
the	29	26	64	1051	106	37	176
this	2	0	23	471	5	10	23
that	5	0	3	173	25	3	31
their	0	1	2	20	7	3	11
those	0	0	1	4	1	0	2
these	0	0	0	26	1	0	7
though	0	0	0	1	0	0	1
which	0	2	2	36	3	1	12
where	1	0	1	0	2	1	2
whether	0	0	1	6	0	0	0
who	0	0	3	4	1	0	4
when	0	0	1	5	2	0	2
while	0	0	2	5	3	0	0
whole	0	0	2	1	2	0	0
why	0	0	1	1	0	0	0

Table 1 demonstrates the results achieved for all natural watermarks parsed from the studied websites. It clearly describes the words analyzed and their frequency of occurrence in each website. It has also been observed that for few words there is a zero value which proves their less frequency occurrence in English content.

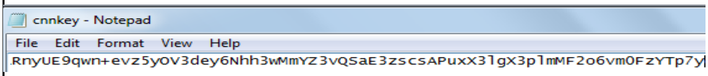
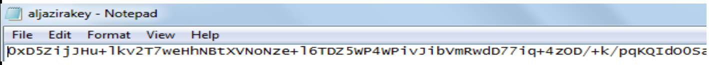
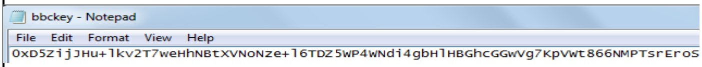
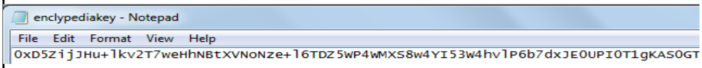
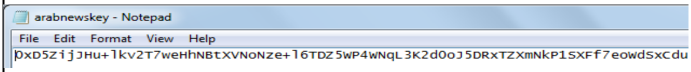
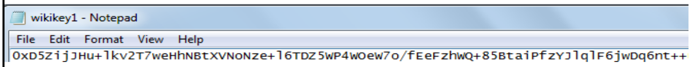
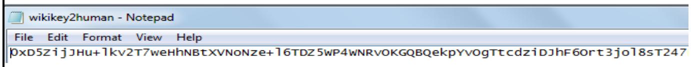
Table 2. This shows the websites URL's used for the testing of system; some famous websites are taken for experiments

Numbers used in Table 1	***Corresponding Web sites***
1	http://www.cnn.com/
2	http://english.aljazeera.net/
3	http://www.bbc.co.uk/news/magazine-12493236
4	http://www.encyclopedia.com/topic/global_warming.aspx
5	http://arabnews.com/lifestyle/science_technology/
6	http://en.wikipedia.org/wiki/Educational_accreditation
7	http://en.wikipedia.org/wiki/Human_computer_interaction

Table 2 shows the names of different websites used in the testing process of the proposed system. Famous websites are considered for experiments for the authentication purposes. Most of these are news websites due to a large number of viewers or readerships. Two pages from Wikipedia are taken due to its popularity and authors contributions as well threats for the authorships. Websites are numbered for the simplicity and readability.

Table 3 illustrates the secret keys which are in encrypted form; these are captured as it is from the system to show their actual format as, the system at the end produced the key in a text file format. Keys produced are numbered in the given table for the simplicity, readability and uniformity with other tables.

Table 3. This table shows the secret keys generated for all 7 websites used during the experiments. Numbers for websites are used for the simplicity purposes and corresponding names are detailed in Table 2

Numbers used in Table 1	*Secret Keys generated for each website with author ID "author"*
1	 RnyUE9qwn+evz5yOV3dey6Nhh3wMmYZ3vQSaE3zscsAPuxX3lgX3plmMF2o6vm0FzYTp7y
2	 0xD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WPivJibVmRwdD77iq+4zOD/+k/pqKQIdOOSa
3	 OxD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WNdi4gbHlHBGhcGGwVg7KpVWt866NMPTsrEroS
4	 OxD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WMXS8w4YI53W4hvlP6b7dxJEOUPIOT1gKASOGT
5	 0xD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WNqL3K2dOoJ5DRxTZXmNkP1SXFf7eoWdSxCdu
6	 OxD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WOeW7o/fEeFzhWQ+85BtaiPfzYJlqlF6jwDq6nt++
7	 0xD5ZijJHu+lkv2T7weHhNBtXVNoNze+l6TDZ5WP4WNRVOKGQBQekpYVOgTtcdziDJhF6Ort3jol8sT247

5 Conclusion

English based natural watermarks have been used in this research for the copyright protection of web based content. Two groups of words starting with letters "th" and "wh" further composed of different sets of frequently used words in English language have been used to propose this methodology. <meta> tag which is a part of <head> section of a HTML document is used to embed the secret key which is a cipher text based of watermarks and a unique author ID. The proposed model have been implemented, verified and tested for different popular and authentic websites. Some of the results have been shown in the research to verify the tests performed. The algorithm used for encryption is AES which is considered as most sure and the advanced algorithm and hence a blended security mechanism of digital watermarking with semantic based watermarks along with the cryptography has been proposed in this research.

6 Recommendations and Future Work

During the experiments it has been analyzed that some watermarks have very high frequency thus only those high frequency water marks can be considered for generating the secret key. But on the other hand it may be more vulnerable to be decoded. And some other methods can be considered to overcome this issue. The reason for choosing many watermarks in generating a secret key is to hide the perceptibility, readability and to make it more difficult to decode the secret key, which can further be analyzed and studied against different parameters. Further to this some other letters or words can also be considered for using as a watermark for the copyright protection purposes. And this idea can also be extended towards other languages. Another issue which has been noticed in during the experiments is the length of the secret key. In case of finding many watermarks in a page there is a probability of having a long key and this issue can further be resolved shortening the length of the key which can add another layer of security. One possible and easy way is to use Huffman encoding for the key only. Another case which was considered in this research at later stages was to use a low number of key for AES which can help in addressing the issue. Other strong mechanism can be studied further as well. In this research all possible words starting with the initials “th” and “wh” are taken into consideration for generating a semantic key. Taking all possible words can be studied and analyzed into two different angles. First taking all possible words makes the secret key strong and more difficult and time consuming to get extracted but in the meantime it may threat the robustness when there is a huge content. For a very large content the system will take more time in generating the key and sometimes more complex as well. To overcome this we can choose few specific words starting with these initials with respect to their occurrence frequency to generate a semantic key. Further some other ways can be studied and proposed to overcome this issue as well.

References

1. Low, H., Maxemchuk, N.F., Brassil, J.T., O’Gorman, L.: Document Marking and Identification using Both Line and Word Shifting, AT&T Bell Laboratories, pp. 0743-166W95. IEEE, Los Alamitos (1995)
2. Low, H., Maxemchuk, N.F., Brassil, J.T., O’Gorman, L.: Copyright Protection for Electronic Distribution of Text Documents, vol. 87(9). IEEE, Los Alamitos (1999)
3. Kim, Y.-W., Moon, K.-A., Oh, I.-S.: A Text Watermarking Algorithm based on Word Classification and Onter-word Space Statistics. ICDAR. IEEE, Los Alamitos (2003)

4. Yang, H., Kot, A.C.: Text Document Authentication by Integrating Inter Character and Word Spaces Watermarking, Vol. ICME, vol. 2. IEEE, Los Alamitos (2004)
5. Zhou, X., Zhou, W., Wang, Z., Pan, L.: Security Theory and Attack Analysis for Text Watermarking. In: EBISS (2009)
6. Meral, H.M., Sevinc, E., Unkar, E., Sankur, B., Ozsoy, A.S., Gungot, T.: Syntatic tools for Text Watermarking, Bogazici Univ. and TUBITAK project (2009)
7. Shirali-Shahreza, M.H., Shirali-Shahreza, M.: A New Approach to Persian/Arabic Text Steganography. In: ACIS, IEEE, Los Alamitos (2006)
8. Mir, N., Hussain, S.A.: Web Page Watermarking: XML files using Synonyms and Acronyms, Under Publication, Procedia Computer Science (2010)
9. Macq, B., Vybornova, O.: A Method of Text Watermarking using Presuppositions. In: SPIE, vol. 6505 (2007)
10. Mohammad Laheen, Sun XingMing: Techniques with Statistics for Web page Watermarking, NSFC No.60373062 (2005)
11. Ala'a, H., Mazin, S., Al Hamami, M.A.: A Proposed Method to Hide inside HTML Web Page File
12. Shingo, K., Ichiro, O.: A Proposal on Information Hiding Methods using XML, Mitsubishi Research Institute, Communication Research Laboratory, Yokohama National University and The University of Tokyo (2002)
13. Aasma, Sumbul, Asadullah: Steganography: A New Horizon for Safe Communication through XML. JATIT (2008)
14. Mir, N.: Robust Techniques of Web Watermarking using verbs, articles and prepositions. IJCSIS International Journal of Computer Science and Information Security (2011)

Sentiment Classification from Online Customer Reviews Using Lexical Contextual Sentence Structure

Aurangzeb Khan, Baharum Baharudin, and Khairullah Khan

Universiti Teknologi PETRONAS Perak, Malaysia
aurangzebb_khan@yahoo.com, bharb@petronas.com.my,
khairullah_k@yahoo.com

Abstract. Sentiment analysis is the procedure by which information is extracted from the opinions, appraisals and emotions of people in regards to entities, events and their attributes. In decision making, the opinions of others have a significant effect on customers, ease in making choices regards to online shopping, choosing events, products, entities, etc. When an important decision needs to be made, consumers usually want to know the opinion, sentiment and emotion of others. With rapidly growing online resources such as online discussion groups, forums and blogs, people are commentating via the Internet. As a result, a vast amount of new data in the form of customer reviews, comments and opinions about products, events and entities are being generated more and more. So it is desired to develop an efficient and effective sentiment analysis system for online customer reviews and comments. In this paper, the rule based domain independent sentiment analysis method is proposed. The proposed method classifies subjective and objective sentences from reviews and blog comments. The semantic score of subjective sentences is extracted from SentiWordNet to calculate their polarity as positive, negative or neutral based on the contextual sentence structure. The results show the effectiveness of the proposed method and it outperforms the word level and machine learning methods. The proposed method achieves an accuracy of 97.8% at the feedback level and 86.6% at the sentence level.

Keywords: Sentiment Classification, Feature Extraction, Reviews Mining.

1 Introduction

With the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic classification of text documents and online customer reviews becomes the key method for information organization and knowledge discovery. Proper classification of e-documents, online news, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to get meaningful knowledge [1]. Rapidly growing online resources, online discussion groups, and forums and blogs has led to people commentating via the internet and a vast amount of new data in the form of customer reviews, comments and opinions about a product, events and entities being generated more and more. The reviews about any entity, e.g. banks, hotels, airlines and online shopping items

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 317–331, 2011.

including books, digital cameras, mobile phones, notebooks, etc. are useful in decision making for both the customer and manufacturer. The sentiments from online reviews have a great influence on others in decision making [2].

With the rapid growth of social media content on the internet in the last few years, the world has been altered, and the web is the best way for people to express their views regarding anything on the various social network sites, discussion forums and blogs. If we want to buy a product, travel abroad or stay at a hotel, we are no longer limited to asking our friends and families because there are many user reviews available on the Web. For a company, there may no longer be a need to conduct surveys from focus groups in order to gather consumer opinions about its products and those of its competitors because there is plenty of such information publicly available on the internet [3].

So it is desirable to develop an efficient and effective sentiment analysis technique that is able to analyze the customer review and classify it into positive, negative or neutral opinions about any entity. Several researchers have been working on the sentiment analysis using a domain dependent framework for feature and feedback level opinion classification. A few are using machine learning techniques for classification at the document level. In this work, we proposed a domain independent rule based method for semantically classifying sentiment from online customer reviews and comments. The method is effective as it takes a review, checks individual sentences and decides its semantic orientation considering its structure and the contextual dependency of each word.

The rest of the paper is organized as in section-2 present the related research and motivation for the proposed work. Section-3 descried the proposed method with preprocessing steps .in section-4 we explain to extraction and classification of semantic score for each review and sentence. Section-5 elevates the results and finally in section-6 we conclude our proposed work.

2 Background and Related Research

Researchers have taken a keen interest in sentiment analysis for the last few years. It has attracted a great deal of attention because of its challenging research problems and the wide range of applications for both academia and industry. It needs a computational study for extracting knowledge from the people's opinions, appraisals and emotions toward entities, events and their attributes. In today's international global world market and highly growing internet usage, people prefer online shopping, banking, ticket reservation, hotel booking, etc. So sentiment analysis from online customer reviews is becoming a requirement of an organization, customer and also manufacturer. Different researchers have been working on different aspects of this area. The existing work on sentiment analysis can be categorized into document, sentence and word/feature level classification.

Word or feature level sentiment analysis gets much importance by applying the natural level processing and statistical methods. Several researchers have worked on extraction of features and opinion-oriented words [4], [5], [6] using a predefined seed word list for extracting semantic orientation and opinion classification. In [4] and [6], the authors used product features for extraction of customer opinions. [7] present

Natural Language Processor Linguistic Parser to parse each review, to split text into sentences and to produce part of speech tags for each word like noun, verb, adjective, etc. A few authors have taken term senses into account and assume that a single term can be used in different senses and can present different opinions. They use Synset from WordNet for different senses of the same term. The [8] used opinionated words for opinion mining from blogs.

The machine learning techniques performed better then lexicon and rule based approaches [9]. They use bag-of-words (BoW), Part-Of-Speech (POS) information and sentence position as features for analyzing reviews and representing reviews as feature vectors to a learning device usually Naïve Bayes and SVM. But these feature extraction methods are also dependent on tools like POS Tagger and no contextual information is considered. In [10], the authors proposed a method for sentiment classification based on conditional random fields (CRFs) in response to the two special characteristics of "contextual dependency" and "label redundancy" in sentence sentiment classification. CRFs capture the contextual constraints on the sentence sentiment. A Hierarchical framework is used for introducing redundant labels and capturing the label redundancy among sentiment classes. The Hierarchical structure is very costly and ineffective in a large scale data set.

Most of the existing work focused on document level sentiment classification [11]. In [12], the authors use a machine learning technique with a minimum cuts algorithm for sentiment classification. Topic oriented classification models normally represent a document as a set of terms in which topic sensitive words are important. In contrast, polar terms such as "excellent" and "worst" are considered essential to sentiment-oriented classification. The sentiment structures in sentence context are more expressive than individual polar term based features [13]. 'The full story of how lexical items reflect attitudes is more complex than simply counting the valences of terms' [14].

In [15], the problem of attributing a numerical score (one to five stars) to a review is presented. They use the feature representations of reviews and describe it as a multi-label classification (supervised learning) problem, and present two approaches using Naïve Bayes (NB) and Support Vector Machines (SVM's). In [16], a system is presented which classifies documents and then checks subjectivity of sentences in it. The machine learning approach with the integration of compositional semantics of sentiment classification is presented in [17]. The support vector machine (SVM) algorithm with 'bag of words' (BoW) to classify movie reviews is presented in [18], in which a few types of special features are selected. However, the limitation of this approach is that, it only focuses on adjectives and their modifiers that express appraisal. The method in [19] extracts the polarity of phrases using the point-wise mutual information (PMI) between the phrases and seed words. Most of the above mentioned techniques use flat feature vector (a bag-of-words) BoW methods used to represent the documents. However, statistical based techniques rely on subject, domain and language style to gather large amounts of significant data with statistics while neglecting contextual information and syntactical structure, which in turn affects the accuracy of the sentiment classification at small textual composition levels. So the techniques may not accurately represent the information that can be extracted at sentence level. To measure sentiment on the phrase or sentence level, opinion oriented words were proposed by simple methods for combination of individual sentiments [20] and supervised [21] statistical techniques. [22] proposed a machine learning method using both lexical and syntactic features for sentiment analysis. These methods, however, missed vital contextual information. So, the individual sentence is important for extracting semantic orientation.

Rule based techniques approaching the analysis of word dependency and structure of contextual information for sentiment orientations were proposed in [23], [24]. In [23], the authors proposed opinion extraction from noisy text data at multiple levels of granularity using domain knowledge for contextual structure and WordNet for semantic orientation.

The limitations of these techniques are manually developed domain-dependent lexicons and inability to deal with long complex sentences. A lexical system for sentiment analysis at various grammatical levels is presented in [24]. This approach used a wide-coverage lexicon, accurate parsing and sentiment sense disambiguation semantic orientation So, the contextual information of all the parts of speech is essential for the semantic orientation, as was shown in [25]. All the content, parts of speech and the structure of the sense in the sentence play a vital role in sentiment analysis The main limitations of the existing approaches are the concentration on sentence structure and the contextual valance shifter is low; lexicon based systems suffer from limitations in lexical coverage, Word since disambiguation which is ignored, rule of term weighting, and the polarity score is too generalized; moreover, less attention is given to attenuation or the imperial expression or the confidence level of the sentiment orientation in the expression is ignored, and there is no proper rule for handling the noisy text with photonic symbols and special characters.

In this paper we proposed a method of sentiment classification at the sentence level applying rules for all parts of speech to score their semantic strength, contextual valence shifter, expression or sentence structure based on dynamic pattern matching, and word sense disambiguation is addressed. The system identified opinion type, strength, confidence level and reasons. It deals with the SentiWordNet [26], WordNet[1], as the knowledge base with the additional capability of strengthening the knowledge base with modifiers and contextual valence shifter information, and is used for all parts of speech.

3 Proposed Framework

For classifying and analyzing sentiments from online reviews and blog comments, we use lexical contextual information at the sentence level to check whether sentences are objective or subjective, and to classify subjective sentences into positive, negative or neutral opinions. In our previous work [27], we used a machine learning algorithm for classifying sentences into objective and subjective and for finding their polarity. In this work, we proposed a rule based lexicon method to determine subjectivity from objectivity sentences. From subjective sentences, we extract the opinion expression and check their semantic scores using the SentiWordNet directory. The final weight of each individual sentence is calculated after considering the whole sentence structure, contextual information and word sense disambiguation. Fig-1 shows the overall

[1] `http://wordnet.princeton.edu`

process of the sentiment analysis of the proposed system. The steps are described below.

- Split reviews into sentences and make a Bag of Sentences (BOS).
- Remove noise form sentences using spelling correction, convert special characters and symbols (photonics) to their text expression, use POS for tagging each word of the sentence and store the position of each word in the sentence.
- Make a comprehensive dictionary (feature vector) of the important feature with its position in the sentence.
- Classify the sentences into objective and subjective sentences using lexical approach.
- Using a lexical dictionary as a knowledge base, check the polarity of the subjective sentence as positive, negative or neutral.
- Check and update polarity using the sentence structure and contextual feature of each term in the sentence.

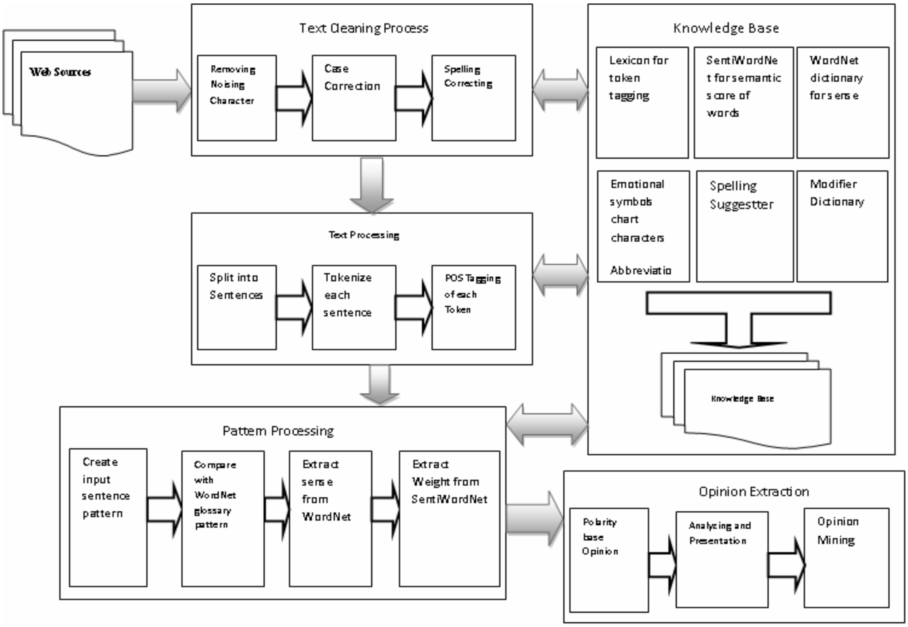

Web Sources
Text Cleaning Process
Removing Noising Character
Case Correction
Spelling Correcting
Text Processing
Split into Sentences
Tokenize each sentence
POS Tagging of each Token
Pattern Processing
Create input sentence pattern
Compare with WordNet glossary pattern
Extract sense from WordNet
Extract Weight from SentiWordNet
Knowledge Base
Lexicon for token tagging
SentiWordNet for semantic score of words
WordNet dictionary for sense
Emotional symbols chart characters Abbreviatio
Spelling Suggestter
Modifier Dictionary
Knowledge Base
Opinion Extraction
Polarity base Opinion
Analyzing and Presentation
Opinion Mining

Fig. 1. Proposed Architecture for Sentiment Analysis

3.1 Sentence Splitter and Processing Noisy Text

In this section the reviews are spitted into sentences to extract feature level sentiment score by SentiWordNet. Making BOS form spitted sentences and stored as with review Id and sentence id. After applying the POS, the position of each word in the sentence is also stored for further processing.

The sentence boundaries identification is important to split the reviews into correct sentences. For this purpose we have implemented a rule based module is which "." is consider as sentence boundary, when it is not preceded by predefined word i.e. Pvt., Ltd., etc. and also ignore the "." after abbreviation list (defined in dictionary) and immediate after digits which not follow space character. To remove noise from text we applied an algorithm to remove symbols, check spelling and corrected those words which are wrong written.

3.2 Part of Speech (POS) Tagger

For assigning tag to each word in sentence, we used POS tagger by adopting the Stanford trigger lexical database as knowledgebase and connect it with our system with some changes for efficient and effective tagging. The system extract the reviews and comments from web using crawler and then clean it and apply the pos for tagging. All the words are tag. JJ, JJS, VB, VBS, RB, NN, NNS, DT, etc. ad described in Table-1.

Table 1. POS types

POS_Name	POS_ Abbreviation	SentiWordNet_Abrv
Noun	NN	n
Adjective	JJ	a
Verb	VB	v
Adverb	RB	r
Nouns	NNS	n
Adjectives	JJS	a

3.3 Feature and Opinion Word Position Extraction

The algorithm for sentiment classification uses opinion word to determine the polarity of sentence based on the contextual information and sentence structure. The position of each word in a sentence is important for the semantic orientation and correct pattern extraction for word science disambiguation. Also the product feature and opinion word are extracted from the tagged sentences using word position. We select the feature from the list at run time after suggesting the most frequent feature extracted from the opinionated sentences. To extract opinion words from the sentences first we focus to find the features that emerge explicitly as nouns or noun phrases in the reviews. The following steps are used.

- Use Part of speech (POS), to tag every word of the sentence and store each word position with assigned tag.
- Collect noun, noun phrases and adjectives with their positions
- Noun phrase are observed as product features.
- For each sentence in the review if contain any feature word, extract the nearby adjective, and consider such an adjectives as opinion words.
- Adjectives and or adjective processed by adverbs are observed by opinion words.
- Frequent product feature are selected from the key noun phrases

3.4 Sentiment Sentences Extraction

In this section we apply subjective sentences extraction method to classify the sentences into objective and subjective one. In our previous work [27] we used the supervised learning approach to extract the subjective sentences. In this work we use rule based module to extract those sentences which contains opinion or subjective words referring from SentiWordNet, WordNet or subjectivity lexicons knowledge base. One unique aspect of this work is to check the word sense disambiguation by using our new proposed method in which we extract the pattern of that sentence using POS and the word position in the sentences, then extract all possible patterns of each sense from the WordNet glossaries, the system locate for the exact match with pattern of the sentence with id found that exact sense pattern score is the extract from the SentiWordNet, that gives very efficient results score. If the patterns are not exactly matched, then it checks for the nearest pattern and the score of that nearest is extract from the SentiWordNet.

3.5 Knowledge Base for the Sentence Structure and Contextual Information

Knowledge base conations SentiWordNet, WordNet and predefined intensifiers dictionary for domain independent polarity classification for positive, negative and neutral opinion. Sentiment words are usually classified into positive and negative categories. For this purpose we extract the semantic score of each opinion word using the SentiWordNet dictionary containing the semantic score of more then117662 words. Then considering the sentence structure and the associated words which affect the weight of the opinion word and update the polarity accordingly. The main aspect of this work is a knowledge base for the contextual information of each word in a sentence which really modifies the strengths of opinion. The knowledge base (semantic strength calculation for each sentence) contains negation words, enhancers, reducers, model nouns, context shifters and other intensifiers with their semantic scores.

- Negation

Negation word are (Not, Never, N't, Does'nt, Cannt, Nor, Don't, Would'nt, No, etc) it will reverse its polarity.

- Contact shifter

There is few type of context shifter to populate our knowledge base with semantic score followed by some specific rules for semantic weight extraction from sentences.

- The contact shifter (But, expect, however, only, although, though, while, whereas, etc)
- Contradictory nature contact shifter (Although, Despite, While)
- Mobilizing or modal contact shifter (Would, Should)
- Pre-Supposition contact shifter (Miss, forget, refused, assumed, hard, harder, less, etc)

If sentence have any such type of word, then the polarity will be recalculated because these words affect the polarity of the opinion word. The negation words reduce its effect to nothing.

- Modifiers (Enhancer and Reducer)

  Modifier word in the sentence e.g. (Slightly, Somewhat, Pretty, Really, Very, Extremely, (the) most), if find the word then recalculate the polarity referring the weightage dictionary the same process will be repeated that score of which opinion word will be effected. e.g, in the sentence "the staff were very nice and cooperative", in this sentence the very is enhance the weight of the nearest opinion word i.e. nice.

- Modifiers of certain Nouns

  Curtains nouns like (a(little)bit of (a), a few, Minor, Some, a lot, Deep, Great, a ton of) effect the sentence polarity, so recalculate the polarity if such type of word occur. From the dictionary of the weights of words/terms, assign weights to each sentence accordingly.

4 Contextual Semantic Orientation of Sentences

In this section we describe the process of assigning weight to each sentence and to decide about the review to be positive, negative or neutral. We used rule base method to check the polarity of the sentences and contextual information at the sentences level. The process is used to extract the contextual information from sentence and calculates their semantic orientation using SentiWordNet, WordNet and predefined intensifier semantics score dictionaries. The following process shows the overall polarity calculation of the proposed method to take the sentence structure:

- Split the reviews into sentences, and a Bag of sentences is created (BoS).

```
REVIEWS: = split corpus
SENT: = Split Reviews
REW_ID:= Assign ID to each Review
SENT_ID:= Assign ID to each sentence
WORD_LIST:= list of words in sentence
WORD_POSITION: = position of each word in a sentence
```

- Classify the sentences into subjective and objective.

- Applying POS and clean the sentences and take subjective sentences for further processing.

```
SENT – sentences to be tagged
WORD_LIST:= list of words in sentence
For each WORD in SENT compare with LEXICON and tag it.
RETURN TAG_SENT
```

- Check each sentence, and find the required word *(WRD),* if exist in the sentence, the extract its position in the sentence. *X= POS_WRD.*

- Check the opinion word *(OW)* in the sentence by calculating its position as *(X-5)* and *(X+5)* in the sentence. If found then mark is as opinion sentence and assign The word to *N ie (N=OW)*

- For the correct sense, extract the sense-id from WordNet using semantic pattern of the desired sentence, refer to SentiWordNet the semantic score of *WRD* is extract on the basis of that sentence structure.

```
SELECT only NN JJ RB VB from TAG_SENT
Place WRD at NN JJ RB VB place
CONCATINATE tags of k+3 and k-3 with WRD
RETURN DES_PATTERN
SLT_PATTERN – extracted pattern from wordnet glossary
SELECT SENSE_NO of SLT_PATTERN from WORDNET
```

- Now calculate its word semantic orientation and assign a weight to this word from the SentiWordNet dictionary.(OW←SEM_SCOR)

```
SENTIM_WORD_SCORE:= extract positive negative score from the SentiWordNet according to SENSE_NO
IF the POSITIVE_SCORE is greater than NEGATIVE_SCORE THEN
SENTIM_WORD_SCORE:= POSITIVE_SCORE
ELSE the POSITIVE_SCORE is less than NEGATIVE_SCORE THEN
SENTIM_WORD_SCORE:= NEGATIVE_SCORE
```

- Sentence level polarity is calculated as consider the sentences to calculate the average score in the sentences the following rules are take into consideration.

```
MODIFIER_WEIGHT:= weight of SENT_SENTIM_WORD in MODIFIER_DICT
MODIFIER_DICT: = list of Modifier which affects the score of positive and negative polarity
IF SENT_SENTIM_WORD is similar JJ OR SENT_SENTIM_WORD is similar RB THEN
CHECK (SENT_SENTIM_WORD + 3) and (SENT_SENTIM_WORD - 3) for Modifier from MODIFIER_DICT
IF WORD found as MODIFIER THEN calculate overall weight.
```

- If there is negation word (Not, Never, N't, Does'nt, Cannt, Nor, Don't, Would'nt, No) near the *N*, *Check (N+3) and (N-3)* then reverse its polarity. e.g. *(OW=+0.8 →OM= -0.8)*

- If there is any type of context shifter in the sentence then the polarity will be recalculated because these words affect the polarity. The position of the contact shifter were checked in sentences , then check the nearest opinion word may *be JJ, JJS, noun NN, NNS or VB, VBS* , if its score is negative then it will be change it after recalculating its weights and vice versa. The negation words reduce its effect to nothing.

- Check the modifier word in the sentence, if exists then recalculate the polarity referring the weightage dictionary the same process will be repeated that score of which opinion word will be effected. e.g, in the sentence “the staff were very nice and cooperative”, in this sentence the very is enhance the weight of the nearest opinion word i.e. nice.
- Curtains nouns affect the sentence polarity, so recalculate the polarity if such types of word occur. From the dictionary of the weights of words/terms, assign weights to each sentence accordingly. The steps of rule base system for contextual valance shifter is describes as below.

```
IF the MODIFIER is a negation modifier THEN
SENTIM_WORD_SCORE:= Reverse the polarity of SENT_SENTIM_WORD
IF the MODIFIER is a intensifier THEN
SENTIM_WORD_SCORE:= intensifying MODIFIER_WEIGHT obtained from MODIFIER_DICT
SENTIM_WORD_SCORE:= SENTIM_WORD_SCORE + MODIFIER_WEIGHT
IF the MODIFIER is a decelerator OR IF the MODIFIER is enhancer OR IF the MODIFIER is context shifter THEN
SENTIM_WORD_SCORE:= intensifying MODIFIER_WEIGHT obtained from MODIFIER_DICT
SENTIM_WORD_SCORE:= SENTIM_WORD_SCORE + MODIFIER_WEIGHT
```

- Calculate the final weights of each sentence and review to decide about positive, negative or natural. The below equation-3and 4.

$$SentenceSc\ ore(Sen) = \frac{\sum_{i=1}^{n} Score(i)}{n} \tag{3}$$

Where,
Score (Sen), are the positive or negative score of sentence Sen, Score(i) is the positive, negative score of ith word in sentence S. n is the total no. of words in Sen.

$$\mathrm{Re}\ viewScore\ (\mathrm{Re}\ w) = \frac{\sum_{i=1}^{n} Score\ (Sen\ )}{n} \tag{4}$$

Where,
Rew(Score), are the positive or negative score of Review Rew, Score(Sen) are the positive, negative score of ith sentences in review. n is the total no. of sentences in the review.

5 Experiments and Results

For evacuation of our proposed method, we collected three types of online customer reviews datasets to check the system performance (1) popular publicly available corpus from movie-review polarity dataset v2.0 IMDB movie reviews[2] . The data set consists of 1000 positive and 1000 negative reviews in individual text files, and also

[2] http://www.cs.cornell.edu/people/pabo/movie-review-data/

the sentences polarity dataset (includes 5331 positive and 5331 negative processed sentences / snippets [28]. Table-2 shows datasets information.

We take the positive and negative sentences to check the performance of our proposed system. (2)We extracted 1000 reviews from the Skytrax[3], where more than 2.5 million independent reviews for over 670 airlines and 700 airports. After spilt the reviews into sentences there 8 average sentences per review is found. We extracted the subjective lexicons and semantic orientation from all the positive and negative sentences. (3) We performed our experiments on the dataset of about 2600 hotel reviews downloaded, which is collected from TripAdvisor[4,] that is one of the popular review sites about hotels and traveling. We extracted only text of reviews using text file. Table -2 and Table -3 one shows the customer reviews, no of sentences per review and the objective and subjective sentences in the reviews.

Table 2. Processed Datasets

Datasets	comments	sentences	Sentences/comments (average)
Movie Reviews	Already processed	10662	10
Airlines Reviews	1000	7730	8
Hotel Reviews	2600	25663	10

We processed all the datasets to remove the noise and clean from the special characters, symbols and check for the spelling mistakes, applied POS tagger and classify into subjective and objective as shown in Table-3. The movie reviews data has already been processed for positive and negative sentences. We consider only the subjective sentences for further processing to find semantic orientation at individual sentence level. Fig-1 shows the classification of subjective and objective sentences (taken from our proposed system) for the airline and movie review datasets.

We processed the subjected sentences for semantic orientation by taking the contextual features and SentiWordNet for semantic score. The weight is calculated using the Equation-3 and 4 for the final opinion orientation. We evaluate our results using precession and recall. Table-4 shows the overall accuracy of our proposed method.

Table 3. Sum of opinion sentences

Dataset	Reviews	Sents	Subjective	Objective	Percentage(Sub /Obj)
Movie Reviews	Already processed	10662	8530	2132	80/20
Airlines Reviews	1000	7730	5405	2325	70/30
Hotel Reviews	2600	25663	17704	7969	68/32

[3] http://www.airlinequality.com/

[4] http://www.tripadvisor.com/

Fig-3 shows the graphical representation of positive, negative or neutral opinion orientation for comments and review data. The system achieves accuracy of about 91% for feedback level and about 86% at sentence level. So the rule base system with lexical system performs better than machine learning and word level sentiment analysis. Our system is also able to extract the opinion strength of different attributes.

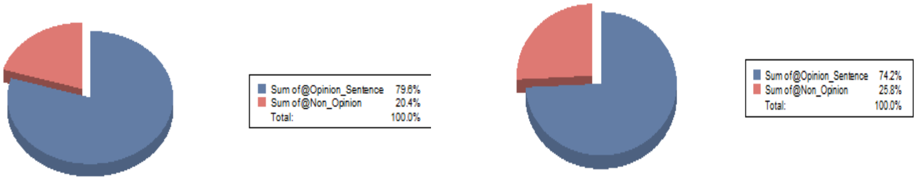


Fig. 2. Percentage of subjective and objective sentences for customer reviews

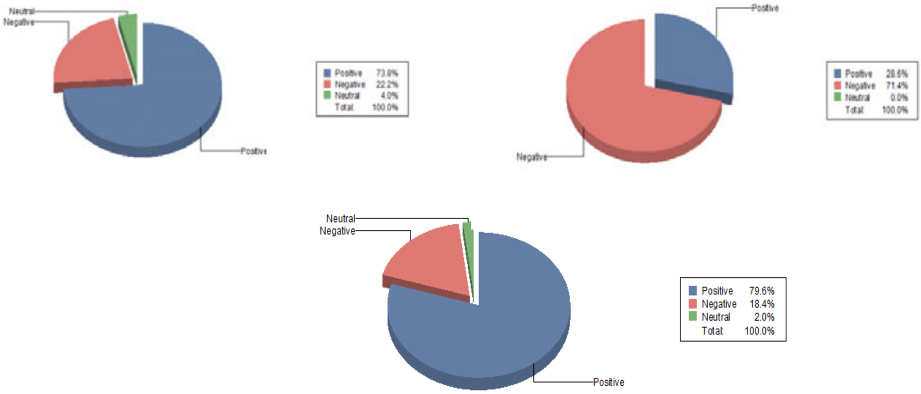


Fig. 3. Accuracy of the proposed system at different data sets

Table 4. Accuracy of opinion orientation for positive and negative

Datasets	Sentiment Orientation	Sentence level	Feedback level
		Accuracy	Accuracy
Movie Reviews	Positive	86.5%	96.8%
	Negative	84.8%	95.7%
	Weighted average	86%	97%
Hotel Reviews	Positive	81.8%	83.7%
	Negative	76.2%	79.5%
	Weighted average	80%	82%
Airlines Reviews	Positive	87.8%	94.8%
	Negative	83.7%	89.9%
	Weighted average	86%	92%

6 Conclusion and Future Work

In this paper, we proposed a rule based sentiment analysis approach for opinion classification. The contextual information and sense of each individual sentence is extracted according to the pattern structure of the sentence. The semantic score for the extracted sense is assigned to the sentence using SentiWordNet. The final semantic weight is calculated after checking each semantic orientation of each term in the sentence; the decision is then made as to polarity of positive, negative or neutral. The results show that sentence structure and contextual information in the sentence are important for the sentiment orientation and classification. The sentence level sentiment classification performs better than the word level semantic orientation. The limitations include dependency on lexicons and lack of term sense disambiguation.

We evaluate our work few types of customer review datasets. From the results, it is clear that the proposed system achieved an average accuracy of 86.6% at the sentence level and 97.8 % at the feedback level for movie, airline review and hotel review datasets. In future, we will improve extraction of the acute sense of the sentence for an efficient semantic orientation; we will also improve the knowledge base for semantic scores of all parts of speech.

References

1. Baharudin, B., Lee, L.H., Khan, K.: A review of machine learning algorithms for text-documents classification. Journal of Advances in Information. Techchnology 1, 4–20 (2010), `http://ojs.academypublisher.com/index.php/jait/article/view/01010420`, doi:10.4304/jait.1.1.4-20.
2. Liu, B.: Sentiment Analysis and Subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) To Appear in Handbook of Natural Language Processing, 2nd edn., pp. 1–38. University of Illinois at Chicago, USA (2010a), `http://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf`
3. Liu, B.: Sentiment analysis: A multi-faceted problem. IEEE Intelligent Syst. 1, 1–5 (2010b), `http://www.cs.uic.edu/~liub/FBS/IEEE-Intell-Sentiment-Analysis.pdf`
4. Popescu, A.M. and O. Etzioni (2004), Extracting product features and opinions from reviews, `http://turing.cs.washington.edu/papers/emnlp05_opine.pdf`
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDDM 2004), pp. 168–177. ACM, New York (2004b), doi:10.1145/1014052.1014073
6. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: Proceedings of the 19th National Conference on Artifical Intelligence, (AI 2004), pp. 755–760. ACM, New York (2004a)
7. Andreevskaia, A., Bergler, S.: Mining wordnet for fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In: Proceedings of the 11th Conference of European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, pp. 209–216 (2006)
8. Attardi, G., Simi, M.: Blog mining through opinionated words. In: Proceedings of the 15th Text Retrieval Conference, November 14-17, pp. 2–7. National Institute of Standards and Technology, Maryland (2006)

9. Baccianella, S., Esuli, A., Sebastiani, F.: Multi-facet Rating of Product Reviews. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 461–472. Springer, Heidelberg (2009), doi:10.1007/978-3-642-00958-7_41
10. Zhao, J., Liu, K., Wang, G.: Adding redundant features for CRFs-based sentence sentiment classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, USA, pp. 117–126 (October 2008)
11. Pang, B., Lee, A.L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd ACL, November 16-24, pp. 271–278. Kimberly Patch, Technology Research (2004)
12. Pang, B., Lee, L., Vaithyanathan, S., Jose, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of the Conference on EMNLP (EMNLP 2002), USA, pp. 79–86 (2002)
13. Hu, Y., Li, W.: Document sentiment classification by exploring description model of topical Terms. Comput. Speech Language 25, 386–403, doi:10.1016/j.csl.2010.07.004
14. Polanyi, L., Zaenen, A. (2004), Contextual valence shifters., `http://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-020.pdf`
15. Sarvabhotla, K., Pingali, P., Varma, V.: Supervised learning approaches for rating customer reviews. J. Intelli. Syst. 19, 79–94 (2010), `http://www.reference-global.com/doi/abs/10.1515/JISYS.2010.19.1.79`, doi:10.1515/JISYS.2010.19.1.79
16. Yu, H., Hatzivassiloglou, V.: Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), USA, pp. 129–136 (2003), doi:10.3115/1119355.1119372
17. Choi, Y., Cardie, C.: Learning with compositional semantics as structural inference for subsentential sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), USA, pp. 793–801 (2008), `http://portal.acm.org/citation.cfm?id=1613715.1613816`
18. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management (IKM 2005), pp. 625–631. ACM, USA (2005), `http://dx.doi.org/10.1145/1099554.1099714`
19. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 417–424 (July 2002), `http://acl.ldc.upenn.edu/P/P02/P02-1053.pdf`
20. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (CL 2003), USA, pp. 1367–1373 (2003), doi:10.3115/1220355.1220555
21. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Proceedings of the Human Language Technology Conference on Empirical Methods in Natural Language, USA, pp. 579–586 (October 1990)
22. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT 2005), USA, pp. 347–354 (2005), doi:10.3115/1220575.1220619

23. Moilanen, K., Pulman, S.: Sentiment composition. In: Proceedings of the Recent Advances In Natural Language Processing International Conference, Borovets, Bulgaria, September 27-29, pp. 378–382 (2007)
24. Dey, L., Haque, S.M.: Opinion mining from noisy text data. Int. J. Document Anal.,Recognition 12, 205–226 (2009), `http://www.springerlink.com/content/1265305p65512357/` 10.1007/s10032-009-0090-z
25. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Semantically distinct verb classes involved in sentiment analysis. In: Proceedings of the International Conference on Applied Computing (AC 2009), Japan, pp. 27–34 (2009)
26. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Italy, pp. 417–422 (2006), `http://nmis.isti.cnr.it/sebastiani/Publications/LREC06.pdf`
27. Khan, A., Baharudin, B., Khan, K.: Sentence based sentiment classification from online customer reviews. In: Proceedings of the Conference on Frontiers of Information Technology (FIT 2010), pp. 1–6. ACM, New York (2010), doi:10.1145/1943628.1943653
28. Pang, B., Lee, L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 2005), USA, pp. 115–124 (2005), doi:10.3115/1219840.1219855

Experimental Approximation of Breast Tissue Permittivity and Conductivity Using NN-Based UWB Imaging

Saleh Alshehri[1], Sabira Khatun[3], Adznan Jantan[1], R.S.A. Raja Abdullah[1],
Rozi Mahmud[2], and Zaiki Awang[4]

[1] Department of Computer and Communication systems Engineering,
Faculty of Engineering
saaas101@gmail.com, adznan@eng.upm.edu.my,
RSA@eng.upm.edu.my
[2] Department of Imaging, Faculty of Medicine and Health Sciences,
Universiti Putra Malaysia, Serdang, Malaysia
rozi@medic.upm.edu.my
[3] Department of Computer Systems & Networks, Faculty of Computer systems & Software
Engineering, Universiti Malaysia Pahang, Pahang, Malaysia
sabira@ump.edu.my
[4] Microwave Technology Center, Faculty of Electrical Engineering,
Universiti Teknologi Mara, Shah Alam, Malaysia
zaiki437@salam.uitm.edu.my

Abstract. This paper presents experimental study to distinguish between malignant and benign tumors in early breast cancer detection using Ultra Wide Band (UWB) imaging. The contrast between dielectric properties of these two tumor types is the main key. Mainly water contents control the dielectric properties. Breast phantom and tumor are fabricated using pure petroleum jelly and a mixture of wheat flour and water respectively. A complete system including Neural Network (NN) model is developed for experimental investigation. Received UWB signals through the tumor embedded breast phantom are fed into the NN model to train, test and determine the tumor type. The accuracy of the experimental data is about 98.6% and 99.5% for permittivity and conductivity respectively. This leads to determine tumor dielectric properties accurately followed by distinguish between malignant and benign tumors. As malignant tumors need immediate further medical action and removal, this findings could contribute to save precious file in near future.

Keywords: breast cancer, Neural Network, breast tissues dielectric properties.

1 Introduction

Malignant tumor is cancerous while benign tumor is not usually harmful. Most of current breast cancer detection methods do not differentiate between these two types. It is important to make immediate action if the detected tumor is malignant.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 332–341, 2011.

There are several indicators to distinguish them, such as the shape, margin, size and stiffness [1,2]. Tumor viscosity is another important indicator [3]. Malignant tumor is more viscous than benign. Malignant tumor dielectric properties are higher because it usually contains more water and blood contents [4,5,6,7,8,9].

It is presented in Reference [6] that at frequency 4.7 GHz, about 67% of the benign tumors have dielectric constant less than or equal to 50, 10% is greater than 55 and the rest are in between. Also, 70% of the malignant tissues have dielectric constant greater than 70 and 25% is less than 50. This demonstrates a soft margin for the dielectric constant values between 50 and 55. Tissues with high dielectric properties produce more scattered signal than tissues with less dielectric properties for incident UWB signals [6,10].

There have been some works done to distinguishing between the two tumor types [1,2]. These studies conducted the classification based on either the shape or by use of Ultrasonic waves [1,2] and not in very early stage. To the best of our knowledge, no study been done using Neural Network (NN) based Ultra Wide Band (UWB) system to measure the dielectric values and then to distinguish between malignant and benign tumors. By correctly detecting the tumor dielectric properties, it is possible to have stronger decision on the discrimination between malignant and benign tumors.

The paper is organized as follows. The experimental method is presented in next section, followed by the results and finally a conclusion.

2 Method

The high contrast in dielectric properties between normal breast tissues and tumor tissues is the principle behind breast cancer detection using UWB imaging [6]. It is not important to know the exact values while the ratio between them is important [11,12]. The tumor to adipose breast tissue dielectric properties can be 10:1 while it could be low as 1.1:1 for glandular breast tissue [6]. The experiment setup and breast phantom used in Ref. [13] is also considered here. Pure petroleum jelly is used to mimic the breast fatty tissue because of ease manipulation. The materials dielectric property measurement is done using Agilent N5230A VNA and HP 85070B dielectric coaxial probe as show in Fig. 1. Fig. 2 shows the experimental system set-up. The UWB transmitter (Tx) and receiver (Rx) are connected to a PC using Ethernet hub and can be controlled through the PC. The breast phantom is placed between the transmitter and receiver Fig. 3 and 4 show the measured permittivity and calculated conductivity for different water to flour ratios in the range of 2-12 GHz. Signals correspond to various dielectric properties values are generated to train the NN. We have used different ratios of water to wheat flour to form the tumor with different dielectric properties as shown in Table 1.

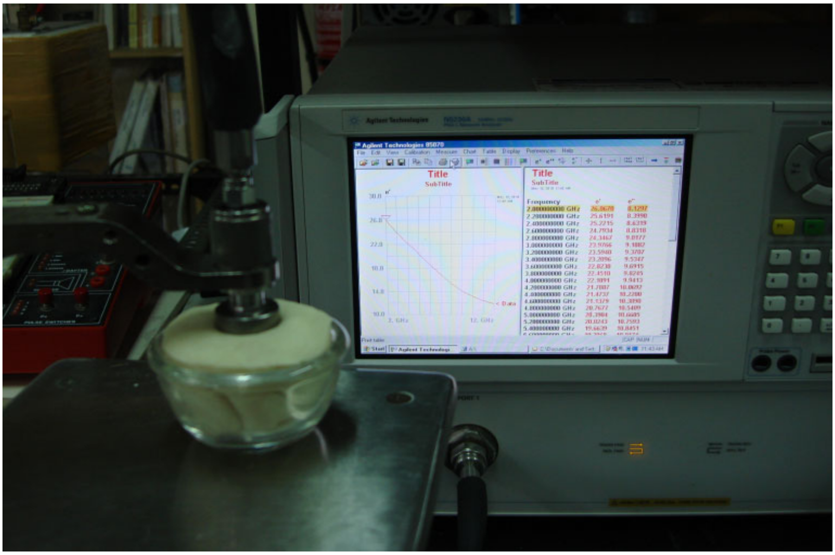


Fig. 1. Agilent N5230A VNA with HP 85070B dielectric probe

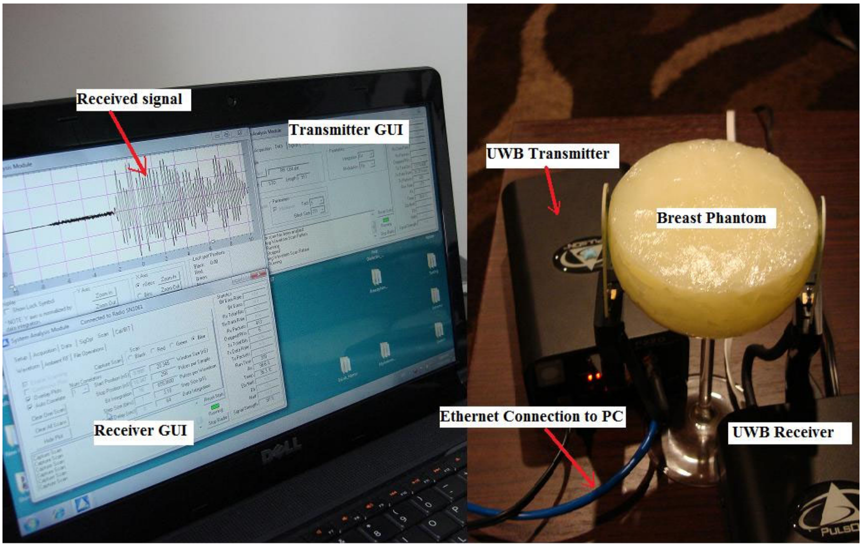


Fig. 2. Experimental system set-up scenario

PulsOn transmitter and receiver are used to transmit and receive the wanted UWB signals. This device operates at center frequency 4.7 GHz and with 3.2 GHz bandwidth [14]. The device has an option to show the waveform of the received signals. Fig. 5 shows two received UWB signals waveform and its calculated DCT components. It shows the difference between the received signals through the healthy breast without tumor and the one with a 7.5 mm tumor located at 70 mm from

Table 1. Dielectric properties of the used materials at 4.7 GHz

Breast phantom part	Material	Permittivity	Conductivity (S/M)
Fatty breast tissue	Vaseline	2.36	0.012
Tumor	Various water to flour ratio	15.2-37.3	2.1-4.0
Skin	Glass	3.5-10	negligible

transmitter. The experimental setup is performed first by placing a 50 mm (diameter) tumor at a distance 50 mm along the line of site between the transmitter and receiver. The tumor has the lowest possible dielectric constant value. After that, UWB signal is transmitted from one side and received on the other side. Finally, the water to flour ratio is changed 14 times to generate different dielectric properties values. The tumor size and location are kept constant.

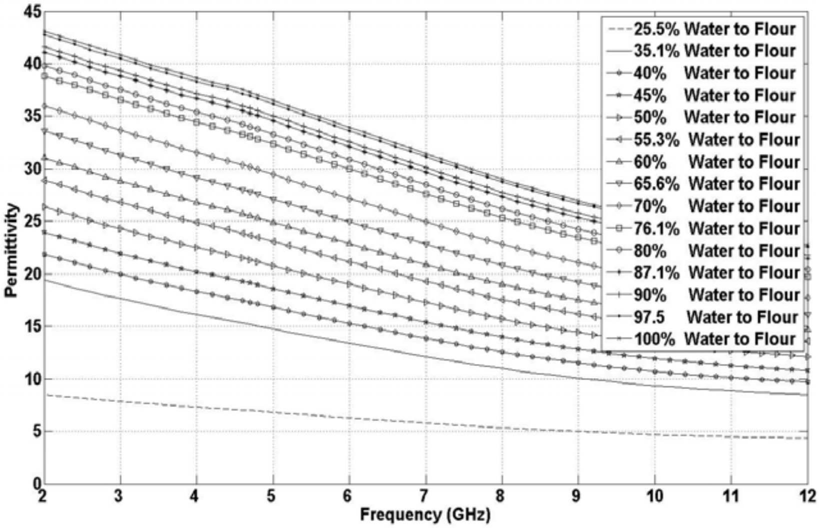


Fig. 3. Permittivity of tumor constructed using various water to flour ratios

Discrete Cosine Transform (DCT) is applied to all of the collected UWB signals. The DCT values between 50-300 are used to generate the training feature vector. This is because the transform shows higher values in this interval. We have used MATLAB software to construct a feed-forward back propagation NN model. It has one hidden layer with 3 nodes and an output layer with two nodes. The two nodes of the output layer produce the tumor permittivity and conductivity. Table 2 shows the NN parameters.

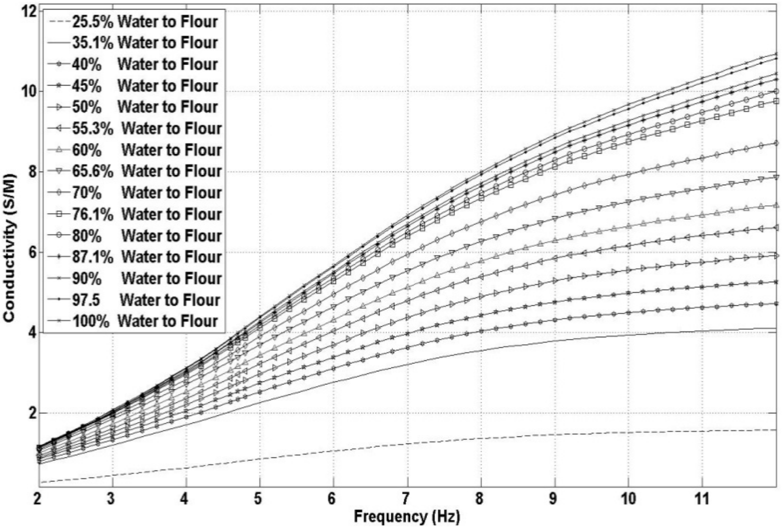


Fig. 4. Conductivity of tumor constructed using various water to flour ratios

Table 2. The NN parameters

NN parameters used in MATLAB	Parameters
Number of nodes in Input layer	251
Number of nodes in Hidden layer	3
Number of nodes in Output layer	2
Transfer function	tansig
Training function	traingdm
Learning rate	0.005
Momentum constant	0.6
Maximum number of Epochs	400000
Minimum performance gradient	1e-25

To calculate the mean error, the used formula is:

$$E = \frac{1}{N}\sum_n (t_n - y_n) \tag{1}$$

where t_n is the target vector, y_n is the predicted values by the NN and N is the number of samples.

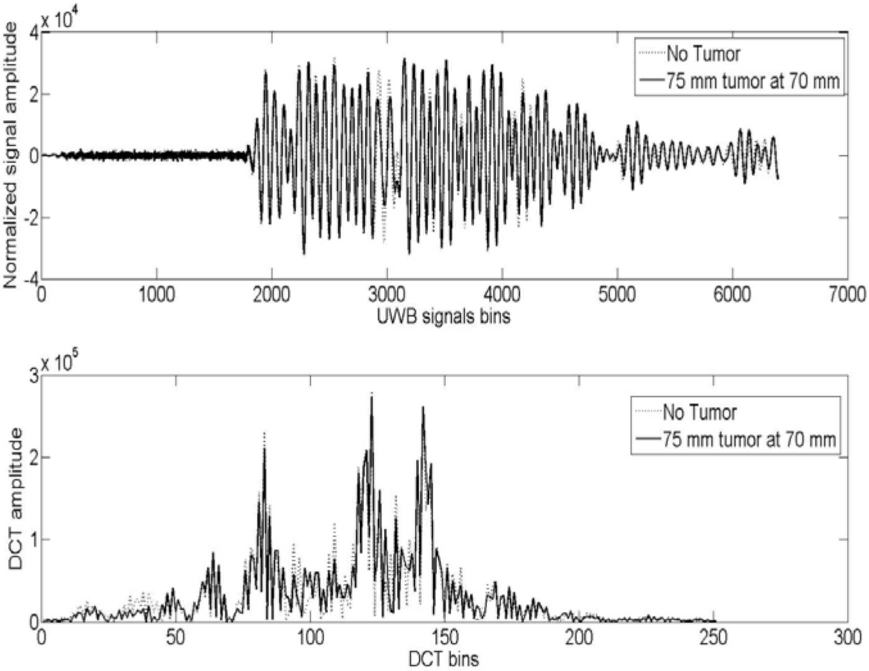


Fig. 5. Raw and DCT of UWB two received signals

3 Results

The experimental work produced performance accuracy in predicting tumor permittivity and conductivity of 98.6% and 99.5% respectively. In the experimental work, it is hard to generate many water to flour ratio values. This limits the number of training samples which is an important factor in successful NN. There is no sharp edge in the dielectric properties values between malignant and benign tumors. However, the output detection of this proposed system can be used to get intuition of the viscosity of the existing tumor. After that it is up to the physician to make the final decision. Table 3 shows experimental results for some of dielectric properties values used for NN training and testing.

Table 3. Tumor dielectric property found experimentally

Permittivity		Conductivity (S/M)	
Actual	NN output	Actual	NN output
15.18	15.45	0.21	0.29
17.28	16.45	0.23	0.28
19.12	19.56	0.25	0.29
21.33	21.87	0.25	0.28

Table 3. (*continued*)

23.68	23.97	0.27	0.30
25.54	25.66	0.30	0.30
25.54	25.08	0.32	0.30
27.80	28.05	0.32	0.30
30.15	31.8	0.34	0.35
30.15	25.82	0.36	0.32
33.07	33.28	0.38	0.28
33.07	33.25	0.38	0.36
34.02	33.95	0.38	0.38
35.29	35.08	0.39	0.38
35.29	35.1	0.39	0.31
35.80	36.08	0.39	0.38
35.80	36.82	0.39	0.36
36.95	37.12	0.40	0.37
36.95	35.64	0.40	0.40
37.28	37.9	0.40	0.43

It can be noticed that there are some permittivity and conductivity repeated values. This is because these values where used repeatedly but each with different transmitted and received UWB signal.

Fig. 6 and 7 show a comparison between some of actual and predicted permittivity and conductivity values. They agree very well and reside on the diagonal line indicating the efficiency of the NN model and the system.

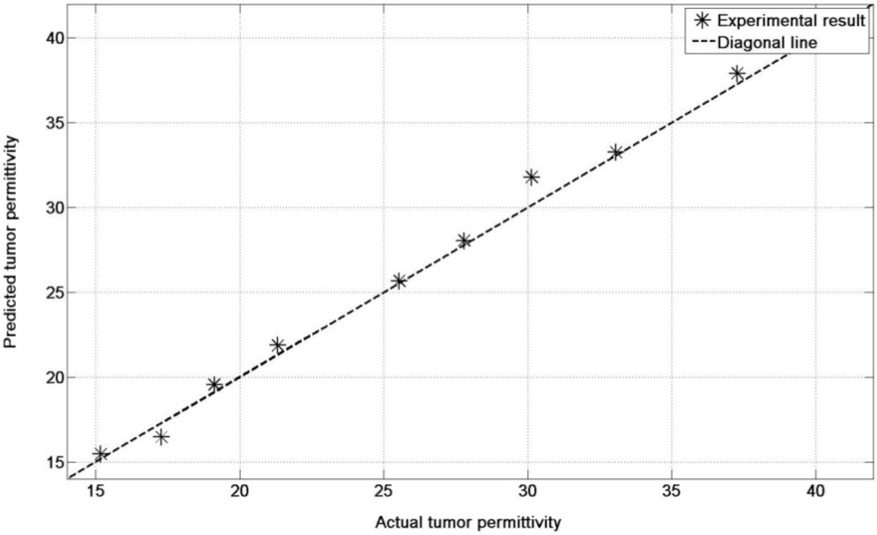


Fig. 6. Actual and predicted permittivity values comparison in experimental work

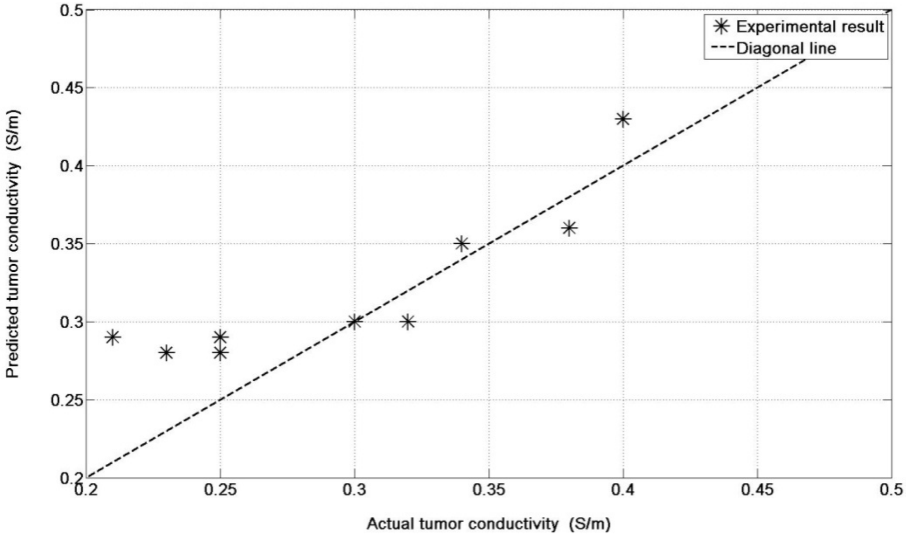


Fig. 7. Actual and predicted conductivity values comparison in experimental work

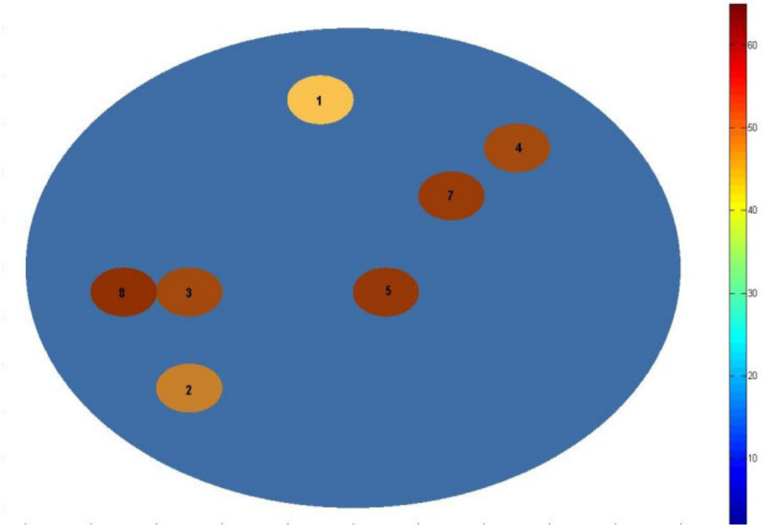


Fig. 8. Permittivity of test group. The darker the higher permittivity

In Fig. 8, some permittivity values of test group tumors are shown. The permittivity values are shown under each corresponding tumor object. The higher the permittivity, the darker the color of the tumor. So, if the tumor permittivity is less than 50, the decision would be that the tumor is benign. If the values is grater that 55 then the decision would be that the tumor is malignant and an immediate action has to be taken by doctors. If the values are in the range of 50-55, then urgent further medical diagnosis is also needed to determine its type.

We could not generate tumors with permittivity and conductivity values greater than 38 and 0.4 (S/m) respectively. Thais is because the water to wheat flour mix will be so sticky in a way that cannot be handled easily. However, the principle is still valid and the system will able to detect and show any tumor types.

4 Conclusion

A NN-based UWB system is developed to predict the dielectric values of tumor tissues experimentally using breast phantoms. Since there is strong relation between the types of tumor and its dielectric properties, this leads to possible discrimination between malignant and benign tumors with higher accuracy of around 98.6% and 99.5% in terms of tissue permittivity and conductivity respectively. Hence the superiority of the developed system is apparent. So, it is possible for doctors to decide whether the tumor is malignant or benign and take the appropriate action in an early stage. Our next step (presently under investigation) is to include the tumor shape and growth in the NN training data to reduce the soft margin and increase detection reliability.

References

1. Li, X., Davis, S.K., Hagness, S.C., Weide, D.W., Veen, B.D.: Microwave imaging via space-time beam forming: experimental investigation of tumor detection in multilayer breast phantoms. IEEE Trans. Microwave Theory Techniques. 52, 1856–1865 (2004)
2. Alshehri, S.A., Khatun, S.: UWB imaging for breast cancer detection using neural networks. Progress In Electromagnetic Research C 7, 79–93 (2009)
3. Sha, L., Ward, E.R., Story, B.: A review of dielectric properties of normal and malignant breast tissue. In: Proceedings IEEE SoutheastCon, pp. 457–462 (April 5-7, 2002)
4. Lai, J.C., Soh, C.B., Gunawan, E., Low, K.S.: Homogeneous and heterogeneous breast phantom for ultra-wideband microwave imaging applications. Progress In Electromagnetic Research 100, 377–415 (2010)
5. Time domain corporation, Comings Research Park, 330 Wynn Drive, Suite 300, Hantsville, Al 35805, USA.
6. Lazebnik, M., et al.: A large-scale study of the ultrawideband microwave dielectric properties of normal, benign and malignant breast tissues obtained from cancer surgeries. Phys. Med. Biol. 52, 6093–6115 (2007)
7. Rangayyan, R.M., El-Faramawy, N.M., Leo Desautels, J.E., Alim, O.A.: Measures of acutance and shape for classification of breast tumor. IEEE Transactions on Medical Imaging 16, 799–810 (1997)
8. Davis, S.K., Van Veen, B.D., Hagness, S.C., Kelcz, F.: Breast Tumor characterization based on ultrawideband microwave backscatter. IEEE Transactions On Biomedical Engineering 55, 237–246 (2008)
9. Insana, M.F., Pellot-Barakat, C., Sridhar, M., Lindfors, K.K.: Viscoelastic imaging of breast tumor microenvironment with ultrasound. Journal of Mammary Gland Biology and Neoplasia 9, 393–404 (2004)

10. Bindu, G., Lonappan, A., Thomas, V., Ananadan, C.K., Mathew, K.T.: Active microwave imaging for breast cancer detection. Progress In Electromagnetic Research 58, 149–169 (2006)
11. Bindu, G., Mathew, K.T.: Characterization of benign and malignant breast tissues using 2-D microwave tomographic imaging. Microwave And Optical Technology Letters 49, 2341–2345 (2007)
12. Hagness, S.C., Taflove, A., Bridges, J.E.: Three dimensional FDTD analysis of a pulsed microwave confocal system for breast cancer detection design of an antenna-array element. IEEE Transactions on Antennas And Propagation 47, 783–791 (1999)
13. Alshehri, S.A., Khatun, S., Jantan, A., Raja Abdullah, R.S.A., Mahmood, R., Awang, Z.: Experimental Breast Tumor Detection Using NN-Based UWB Imaging. Progress In Electromagnetic Research 111, 447–465 (2011)

Banking Deposit Number Recognition Using Neural Network

Bariah Yusob, Jasni Muhamad Zain, Wan Muhammad Syahrir Wan Hussin,
and Chin Siong Lim

Faculty of Computer Systems & Software Engineering,
Universiti Malaysia Pahang, Lebuhraya Tun Razak,
26300 Kuantan, Pahang, Malaysia
bariahyusob@ump.edu.my, jasni@ump.edu.my,
wmsyahrir@ump.edu.my

Abstract. During normal cash deposit process, the bank customer will fill in the account number, amount of cash and name of the account holder at the bank in slip, then key in the account number and amount manually into the computer. If there are numbers of customer at one time, the process will take times and sometime the banker will make errors during reading or keying the data. The recognition process was executed using integration of Artificial Intelligent techniques: image preprocessing and Neural Network. Image processing techniques were used to extract the written character on the slip. After that, the extracted characters were passed to the recognition phase, where Neural Network will identify the input character patterns. Results: We tested the proposed method using 40 cash deposit slip written with numbers to be tested. 3 neural networks with 40, 50 and 60 training data particularly were used to test the success rate of recognition. Through experiment, the proposed system had successfully recognizes at least 90% of the written character on cash deposit slips. Using the proposed approach, we developed an automatic banking deposit number recognition system which is able to recognize the handwritten account number and amount number on the cash deposit slip and thus automate the cash deposit process at bank counter.

Keywords: Image processing, Probabilistic Neural Network (PNN), handwritten, banking deposit, multiple checks, number location detection, image thresholding.

1 Introduction

A deposit slip is a printed form which accompanies bank deposits. The depositor fills out the deposit slip to indicate what types of funds are being deposited and which accounts they should be deposited into. In some cases, a bank will pre-print deposit slips with account information and include them in a checkbook. Deposit slips are used by a bank to keep track of the money deposited over the course of a business day and to ensure that no funds slip through the cracks. For bank clients, a deposit slip

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 342–352, 2011.

offers a form of protection, indicating that funds were counted and accepted by the bank. If the deposit is processed improperly, the deposit slip will provide a study trail.

Most people with a checking or savings account have interacted with a deposit slip and are familiar with the basic format which asks for the name of the client, the date and the account number. Fields underneath this basic information provide a space for the client to enter the type of funds: cash, coin, or check. If multiple checks are being deposited, the deposit slip usually has a space on the back to list and tally them all. The client adds up the funds to come up with a subtotal, indicates whether or not he or she would like cash back and then enters a final total of funds being deposited.

When a client enters a bank with a deposit slip and funds, the bank clerk or teller will count the funds to make sure that the total listed on the deposit slip is correct. The deposit slip is signed, stamped, or printed, depending on the bank, to indicate that it was accepted by a teller and the teller updates the listed total in the bank account to reflect the deposit. If the funds are in the form of cash and coin, the bank re-circulates them. If checks are included in the deposit, the bank sends them on to the issuing bank to be processed. The requested cash back, if any, is also provided [1].

Previously, there was a study done to extract the text area from an optical document image [2]. Apart from that, there are many applications that require the recognition of unconstrained handwritten numeric strings, such as reading bank checks, reading tax forms and interpretation of postal addresses [3]. The task becomes particularly challenging when adjacent digits in the numeric string are touching.

Xian, et. al [4] describe a holistic method for recognizing touching digits in numeric strings. The method supports intuitive knowledge that the central part of the pattern lacks information useful for classification. There is another approach for Numeral String Recognition using character touching type verification in [5]. The numeral character is classified as touching into six types. These touching types are easily detected by comparing length of a vertical block pixel run with that of the horizontal one. The verification unit consists of a pair of character codes and their touching types. The verification units which are impossible in actual character patterns are used to reject isolated character recognition results. This method could be improved for character segmentation efficacy and accuracy by adding new touching types and using character candidates obtained by character segmentation method [5].

There are many problems when dealing with numeric strings. The length of the numeric string is unknown or the length is known. There are distinctiveness and similarities of handwritten numerals. This issue is addressed in [6]. To solve the problem of Numeral recognition efficiently, some researchers have performed improvement of Polynomial classifiers [6]. A performance comparison of statistical and Neural network classifiers is also performed [6]. In theoretical aspect, the statistical classifiers have the fairly appealing advantage. However due to their assumptions and/or the required size of training set, the Multi Layer Perceptron or the 1-NN without any assumption behave better.

In order to retrieve text information on a printed document image, a grayscale image is chosen to be processed. Basically, 2 steps are done in the text detection process:

- Extracts regions of interest in a grayscale document image, using cumulative gradient considerations;
- Classify them in Text and non text zone on the basis of entropy criteria.

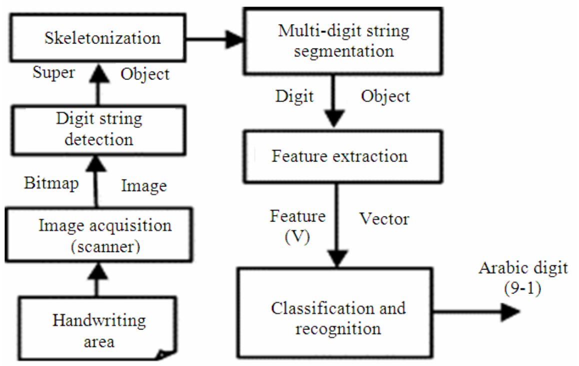


Fig. 1. Structure of the recognition system [7]

Then, the recognition of the handwritten Indian numerals one to nine (1-9) was proposed using a Probabilistic Neural Network (PNN) approaches [7]. The developed algorithm offers flexibility to deal with tilted and off-centered handwritten strings, otherwise known as object rotation and translation, respectively. The developed recognition system consists of the five major processing stages as shown in Fig. 1.

The handwritten digit string acquisition is done by scanning digitally to acquire a digital colored image. The handwritten numeric digit string is placed on a simple background, identifying the digit string is accomplished via a simple thresholding process to obtain a binary image that contains numeric digits on a uniform background. Then the Skeletonization (Thinning) process is done to and will obtain a super-object (multi-digit strings are treated as a single object). In order to run the Recognition and classification and process, the numeric digits is conducted on the individual digits independently, therefore, skeletonized super-object needs to be segmented into several one digit objects before further steps can be performed. Then the handwritten digit string is reduced to several single-bit-wide geometrical shapes through the Feature Extraction process. The Recognition and classification of the numeric digit string is done by probabilistic neural networks.

2 Proposed Methods

In this study, the application of image processing technique and neural network is proposed to recognize the handwritten account number and amount number on the cash deposit slip and thus automate the cash deposit process at bank counter.

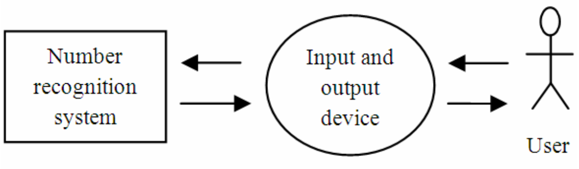


Fig. 2. Basic architecture of the automated banking deposit number recognition system

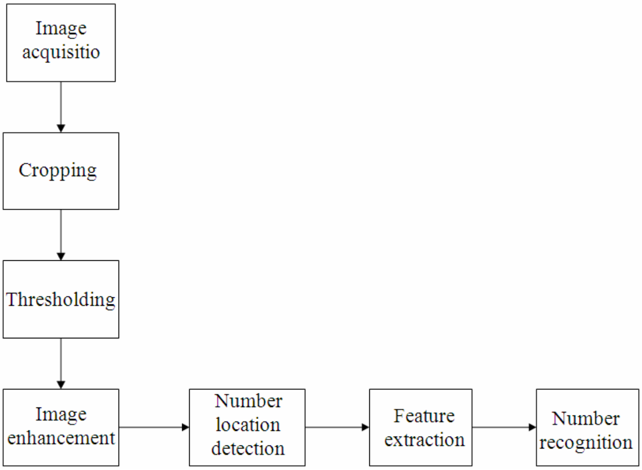


Fig. 3. Flow chart of the Main Number Recognition System

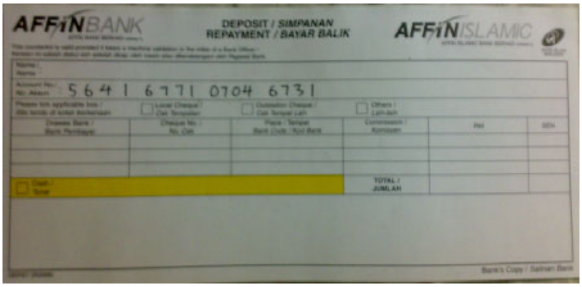
Proposed approach: We divide the system into 3 components as following (Fig. 2):

- Main number recognition system
- Input and output device
- The system user

Main number recognition system. The main number recognition system is the backbone system of the entire system. The recognition system will take input from the input device and do the recognition on the number written. Fig. 3 shows the flow chart of the main number recognition system.

The following process is implemented in order to successfully recognize the numbers written on the bank in slip:

- Image acquisition
- Image cropping
- Image thresholding
- Number location detection
- Image preprocessing
- Feature extraction
- Number recognition



AFFiNBANK

DEPOSIT / SIMPANAN
REPAYMENT / BAYAR BALIK

AFFiNISLAMIC

5641 6771 0704 6731

TOTAL /
JUMLAH

Fig. 4. Sample input image

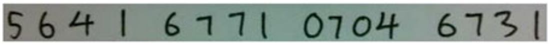


Fig. 5. Sample of cropped image

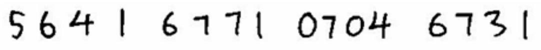


Fig. 6. Sample of binary image

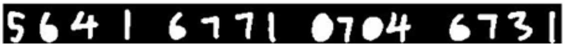


Fig. 7. Sample of dilated image

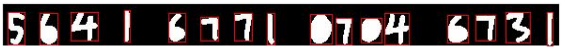


Fig. 8. Detected written number location

Image acquisition: The image acquisition is done by taking the image of the written bank-in slip with a webcam. An image is taken with the webcam from right top on the cash deposit slip (Fig. 4).

Image cropping: The image is then being crop on the part where the account number is written by setting the pixel location (Fig. 5).

Image thresholding: The cropped image is being converted to grayscale image and then converted to binary image (Fig. 6).

Image enhancement: The binary image will then be enchanted with few processes such as edge detection, image dilation and also image filling (Fig. 7).

Number location detection: The location of the written number will then be detected by detecting boundaries of each number with MATLAB (Fig. 8).

Image resize: All the segmented images will then be resized to become same size to meet network input requirement. All squares are resized to become 50×70 pixels as shown in Fig. 9.

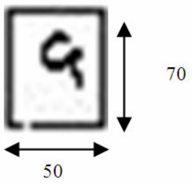


Fig. 9. Image resized to 50×70 pixel

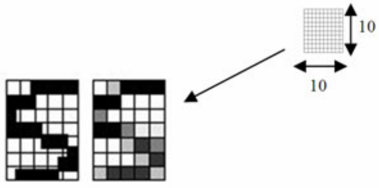


Fig. 10. Image resized again into 5×7 matrices

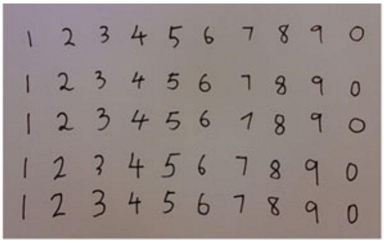


Fig. 11. Sample training data

Feature extraction: The image is then be resized again to meet the network input requirement, 5 by 7 matrices. Fig. 10 shows image with each box contains sum values of each 10×10. Then, the 5 by 7 matrices are concatenated into a stream so that it can be feed into network 35 input neurons. The input of the network is actually the value of the intensity of the relevant pixel.

Number recognition: The extracted data will then be recognized using back propagation neural network. All the data had been trained by the developer into the neural network for the recognition purpose. The training sample is taken from the different handwriting from 3 people. Fig. 11 shows a sample of 5 set of data written for each number for the training purpose. Each sample will go through the image preprocessing process to be feed into neural network.

Fig. 12 shows the neural network design for the recognition process. There are 3 layers in this network: input layer, output layer and 1 hidden layer.

The neural network needs 35 inputs and 10 neurons in its output layer to identify the numbers. The network is a 3 layer log-sigmoid network. The log-sigmoid transfer function was picked because its output range (0-1) is perfect for learning to output Boolean values.

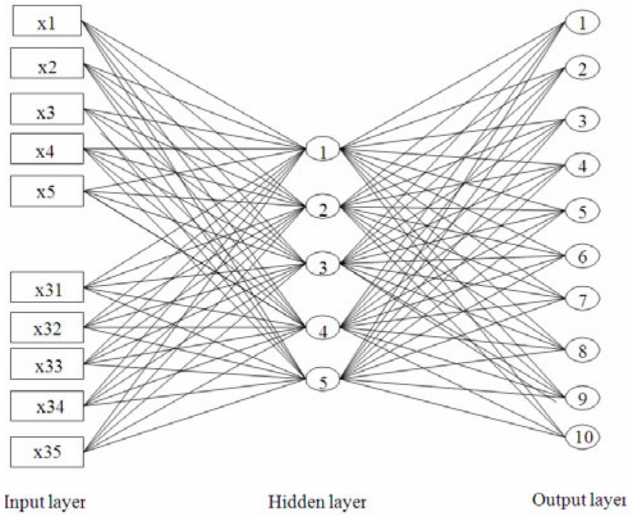


Fig. 12. Neural network for number recognition

Table 1. Desired output pattern for each number

Number	Desired output pattern
1	1000000000
2	0100000000
3	0010000000
4	0001000000
5	0000100000
6	0000010000
7	0000001000
8	0000000100
9	0000000010
0	0000000001

Table 2. Determining the number of hidden layers

No. of hidden layers	Result
None	Only capable of representing linear separable functions or decisions.
1	Can approximate arbitrarily while any functions which contains a continuous mapping from one finite space to another.
2	Represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy.

The performance function of the network is ‘sse’ (Sum squared error). The network will be trained for a maximum of 5000 epochs or until the network sum squared error falls beneath 0.1.

The input layer has 35 input neuron which taken from data extracted from previous phase, the output layer has 10 output neurons since the network is designed to recognize 10 numbers. The desired output pattern for each number is shown in Table 1.

There is only 1 hidden layer in this designed network based on the method to choose the number of hidden layer suggested in [8]. Differences between the numbers of hidden layers are summarized in Table 2 [9].

There are 5 neurons in the hidden layer. This is the initial network designed with 5 hidden neurons. The network will be trained and tested. Then the number of hidden neurons will be increased and the process repeated as the overall result of training and testing improved in the system maintenance phase.

In Automated Banking Deposit Number Recognition system, only the developer is in charge of training the network for recognition whiles the user of the system only using the system without direct access to the infrastructure of system.

Input and output device: The following are the input and output devices required for user to interact with the Automated Banking Deposit Number Recognition System:

- Webcam-To capture image from Cash Deposit Slip
- Keyboard and Mouse-Device used by user to operate the system in computer.

3 Results

We test the proposed method using 40 cash deposit slip written with numbers to be tested. 3 neural networks with 40, 50 and 60 training data particularly are use to test the success rate of recognition.

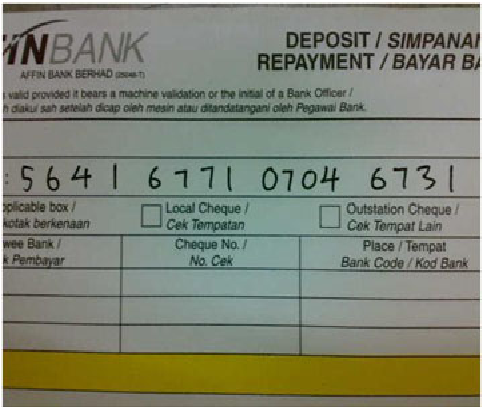
40 samples of cash deposit slip that were used in the testing process were written from 5 different students in ump. The image input of the cash deposit slip were all taken in the same environment to preserve the accuracy of the result. The camera was set to focus on the part of the cash deposit slip where the account number is written on. Figure 13 shows the original image taken from then cash deposit slip and Fig. 14 shows the part of number written obtained from the cropping process.

Image preprocessing: 2 image preprocessing process were implemented on the cropped image in this phase so that the interested information of written numbers is ready to be extracted for recognition purpose.

Image thresholding: The image was first be converted to become a grayscale image as shown in Fig. 15 and then converted to become a binary image as shown in Fig. 16.

Image enhancement: The process edge detection will detect the edges of the numbers in the image. Fig. 17 shows the result image after edge detection process.

Then the image is gone through the process of image dilation and image filling in order to enchant the image for the purpose of feature extraction. Figure 18 shows the result image after image dilation process and Fig. 19 shows the image after image filling process.



NBANK

AFFIN BANK BERHAD (25046-T)

DEPOSIT / SIMPANAN
REPAYMENT / BAYAR B

valid provided it bears a machine validation or the initial of a Bank Officer /
h diakui sah setelah dicap oleh mesin atau ditandatangani oleh Pegawai Bank.

: 5641 6171 0704 6731

plicable box / kotak berkenaan

□ Local Cheque / Cek Tempatan

□ Outstation Cheque / Cek Tempat Lain

wee Bank / k Pembayar

Cheque No. / No. Cek

Place / Tempat Bank Code / Kod Bank

Fig. 13. Original Image taken from the cash deposit slip

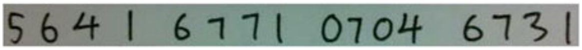


Fig. 14. Image cropping on the written account number

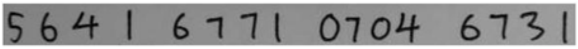


Fig. 15. Result image after converted to grayscale image

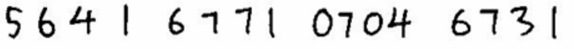


Fig. 16. Result image after converted to binary image

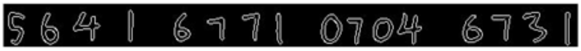


Fig. 17. Result image after edge detection process

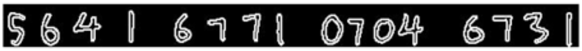


Fig. 18. Result image after image dilation process

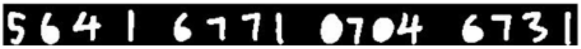
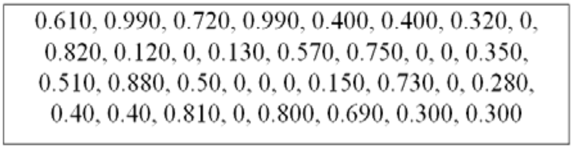


Fig. 19. Result image after image filling process

Feature extraction: Each number written was then been resize to become a 5×7 matrix image and then represent the information of the image in a single vector. The vector extracted from the image is ready to be used for training as well as the recognition purpose. Figure 20 shows the example vector of the number "1" in the image.



0.610, 0.990, 0.720, 0.990, 0.400, 0.400, 0.320, 0,
0.820, 0.120, 0, 0.130, 0.570, 0.750, 0, 0, 0.350,
0.510, 0.880, 0.50, 0, 0, 0, 0.150, 0.730, 0, 0.280,
0.40, 0.40, 0.810, 0, 0.800, 0.690, 0.300, 0.300

Fig. 20. Vector of the image number "1"

Table 3. Result of the testing process

Neural network	Successful rate of 100 (%) recognize slip
40 data trained	75
45 data trained	90
50 data trained	92

Testing result of number recognition process: The testing phase was done using 40 sample of cash deposit slip with 3 neural networks each with 40, 45 and 50 data trained. Table 3 shows the result of the recognition process of each network.

From the result shown in Table 3, neural network with 45 trained achieve rate of 90. Although neural network with 50 data trained achieve a higher rate, but it only shows a minor improvement of 2%, therefore the neural network with 45 data trained is most optimum for the recognition process in this project.

4 Discussion

The objectives of this research were to develop an application that will recognize handwritten account number on the cash deposit slip during bank-in process at bank and to automate the process of banker entering the deposit information. In order to achieve that, we conducted an experiment starting from using webcam to capture image from cash deposit slip until the recognition of the images.

Therefore, our approach was based on two phases, which are character extraction using image processing technique and recognition by neural network. During extraction phase, we used camera to capture the bank-in slip. Then the image was processed through grayscaling and thresholding. The detection of the hand-written number was done as a single object and independently before they can be converted into a set of input data to be fed to the neural network. The back propagation neural network was trained to associate outputs with input patterns.

From the experimental results, we found that neural network has the capability to recognize the input character patterns when integrated with the image processing technique.

We also found that the increasing number of data trained using this technique will increase the successful rate of recognizing the images. This is to confirm that in order to achieve good results in neural network training, sufficient amount of data is required.

5 Conclusion

In this study, we present an Automated Banking Deposit Number Recognition System, which is developed to automate the cash deposit process at bank counter by integrating image processing and neural network techniques. The system recognizes the numbers written by the bank customer on the cash deposit slip instead of key in by the banker in manual way, thus, shorten the cash deposit process. Through experiments conducted to 40 sample of cash deposit slip, more than 90% successful rate of accuracy is achieved. For future work, we suggest to use a computerized agent to choose the most suitable network design for the best recognition process. Later, the system will be enhanced to allow the system to be able to recognize numbers even they are written connected to each other. This is because based on the common nature of human writing style, there are very low possibility that all the number written is well sorted and divided clearly 1 by 1. Thus, this could make the Automated Banking Deposit Number Recognition system to be able to be implemented in a real banking environment. Finally, the image preprocessing processes could also be altered in order to get a better feature extraction result to be used in training and also recognition process in neural network.

References

1. Smith, S.E.: What is a deposit slip (2008), http://www.wisegeek.com/what-is-a-deposit-slip.htm
2. Yi, X., Hong, Y.: Text region extraction in a document image based on the Delaunay tessellation. Pattern Recognition, vol. 36, 799–809 (2003), doi:10.1016/S0031-3203(02)00082-1.
3. Prasad, J.R., Kulkarni, U.V.: Trends in Handwriting RecognitionThird. In: International Conference on Emerging Trends in Engineering and Technology, November 19-21, p. 491 (2010), doi:10.1109/ICETET.2010.92. Dsfdsf
4. Xian, W., Govindraju, V., Srihari, S.: Holistic Recognition of Touching Digits. In: Advances in Handwriting Recognition, pp. 359–369. World Scientific Publications, Singapore (1999)
5. Nishiwaki, D., Yamada, K.: New Numeral String Recognition Method Using Character Touching Type Verification. In: Advances in Handwriting Recognition, pp. 416–425. World Scientific Publications, Singapore (1999)
6. Lee, S.W.: Advances in Handwriting Recognition. Series in Machine Perception Artificial Intelligence, vol. 34. Word Scientific Publications, Singapore (1998), ISBN- 981-02-3715-4
7. Faruq, A.A., Omar, A.: Handwritten Indian Numerials Recognition System using Probabilistic Neural Network. Advanced Eng. Inform. 18, 9–16 (2004), doi:10.1016/j.aei.2004.02.001.
8. Heaton, J.: Introduction to Neural Networks for Java, 2nd edn., p. 380. Heaton Research Inc. (2005) ISBN: 978-0977320608
9. Heaton, J.: A Feed Forward Neural Network (2008), http://www.heatonresearch.com/articles/5/page2.html

Features Selection for Training Generator Excitation Neurocontroller Using Statistical Methods

Abdul Ghani Abro, Junita Mohamad Saleh, and Syafrudin bin Masri

School of Electrical and Electronic Engineering
Engineering Campus Universiti Sains Malaysia
14300 Nibong Tebal, Seberang Perai Selatan
Penang, Malaysia
aga10_eee097@student.usm.my, jms@eng.usm.my, syaf@eng.usm.my

Abstract. Essentially, control system requires suitable control signal for yielding desired response of a physical process. Control of synchronous generator has always remained very critical in power system operation and control. For certain well known reasons power generators are normally operated well below their steady state stability limit. This raises demand for efficient and fast controllers. Artificial intelligence has been reported to give revolutionary outcomes in the field of control engineering. The capability of Artificial Neural Network (ANN) to map any nonlinear function satisfactorily based on input-output data has been widely established in intelligent control. Selecting optimum features to train a neurocontroller is very critical because correlation between features of parameters may avert learning capability of an ANN. In this work statistical methods are employed to select independent factors for ANN training.

Keywords: generator excitation, neural network, regression analysis.

1 Introduction

In recent years it has been recognized that to realize more flexible control systems it is necessary to incorporate other elements, such as logic, reasoning and heuristics into an algorithmic techniques of conventional control theory [1]. Artificial Neural Network (ANN) is not competitive with artificial intelligence, however it is complementary. The use of an ANN with its learning ability avoids complex mathematical analysis in solving control problems when plant dynamics are unpredictably complex and highly nonlinear [2]. This is a distinctive advantage over the traditional nonlinear control methods.

ANNs are parallel distributed processing systems capable of synthesizing a complex and highly nonlinear mapping from input feature space to output space [3]. The parallel processing element distribution not only gives higher degree of tolerance but also higher capability of fast information processing. Another important feature of ANN is learning and adaptation. A well-trained ANN has the ability to generalize training pattern. Networks can also be adopted online [4]. The neural network has many different topologies.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 353–364, 2011.

Multi Layer Perceptron (MLP) is the most commonly used neural network topology as it is a simple type of feed-forward network. MLP has found many applications in intelligent control due to its simplicity and capability to approximate any function with less number of inputs. Another ANN model is the Radial Basis Function (RBF) network. MLP is a global approximator, most suitable for offline training, whereas RBF is a local approximator suitable for online training. Both contain one or many hidden layers. Hidden layer adds nonlinear approximation capability into this architecture. Usually one hidden layer with sigmoid transfer function is sufficient to approximate any nonlinear function [3].

MLPs are very efficient approximators in high dimensional spaces. Due to the nonlinear and highly dynamic nature of power system, artificial intelligence particularly neural networks are finding wide variety of applications in operation and control of power system [5-8]. It's a well established fact that the overall performance of MLP depends upon its input features. Hence, a highly challenging task in training ANN for power system control and operation is the selection of input features. The literature review of [2, 9-11] reveals that variables given in TABLE I were used for ANN training to control excitation of synchronous generator. The output of the excitation system is called excitation voltage (V_f) and it is a dependent parameter. Detailed explanation of excitation system's impact on generator operation is given in next section.

Table 1. Input features used in ANN training

Symbol	**Parameter**
ΔV_T	Terminal Voltage
ω	Rotor Speed
P	Active Power
Q	Reactive Power

ΔV_T is deviation of terminal voltage from reference voltage i.e. V_{REF}-V_T

No proper procedure has been reported for selecting input features for generator excitation neurocontroller training. More input features may require many processing elements and hence more information processing time. In addition, multicollinearity between input features may inhibit a neurocontroller's learning capability. Also, uncorrelated input and output space make the mapping very complex. Furthermore, generator terminal voltage can be sensed by different combinations of parameters in TABLE I even with few additional factors.

The field of statistics deals with the collection, presentation, analysis and use of data in making decisions. Statistical methods are used to assist in describing and appreciating variability. Variability, meaning successive observations of a system or phenomenon, does not always produce exactly the same result. Hence statistical thinking gives us a useful way to incorporate variability into decision making process.

Statistical methods and ANN have been used for prediction and regression, with ANN giving higher accuracy in high dimensional problems [12]. In fact, MLP are nothing more than nonlinear regressor. By using statistical methods optimum parameters can be found for enhancing MLP neurocontroller learning capability while the generator dynamics remain unaffected.

This paper is divided into two parts. The following section discusses the model considered for data generation. The last section presents data analysis based on statistical methods followed by conclusion.

2 Power System Modelling

Power system comprises of generation, transmission and distribution sections. The primary element of generation section is synchronous generator also termed as alternator. Synchronous generator consists of stator called armature and rotor also known as field. Field is responsible for spreading magnetic flux in an air gap. For proper operation of synchronous generator, the key condition is synchronism between armature and field. The strength of synchronism largely depends upon the strength of air gap magnetic flux and the excitation system is responsible for keeping air gap flux strength constant. Synchronism can be jolted by faults induced anywhere in a power system, but the extreme disturbance is fault introduced at terminals of a generator. Fault deteriorates the strength of magnetic flux as explained by armature reaction phenomenon and so has the effect on synchronism. The mechanical angle between rotor magnetic field and armature magnetic flux of a generator is known as the load angle or power angle, δ.

The ability of power system to regain a state of operating equilibrium after being subjected to a physical disturbance or fault is called power system stability. In addition, neither a unit at generating station nor a portion of power system should lose synchronism with respect to the generating station or the power system [13]. The excitation system's output is based on the difference (ΔV) between reference voltage and terminal voltage. The fault causes a decrease in air gap flux density, depending upon the direct and quadrature axis sub-transient and transient time constant. Moreoover duration of fault, and decrease in terminal voltage have great influence on air gap flux density reduction. This leads to increase in ΔV, so the output of generator excitation will shoot up to compensate error. Stability may be enhanced by rapidly increasing excitation current [14]. ANN requires quite considerable time to tune weights but it is fast and accurate once tuned properly. In this research work besides variables given in TABLE I, the effect of one more new variable, deviation of quadratic voltage from reference voltage i.e. $\Delta V_q = V_{ref} - V_q$ was analyzed on excitation voltage (V_f). Terminal voltage is the vector sum of direct and quadratic voltage components. Quadratic voltage was preferred over direct axis voltage component because of its higher correlation coefficient.

Power system stability enhancement is referred to as reducing risk of losing stability by inserting additional signals into the system to smooth out the system dynamics. During steady state excitation system should be driven by only voltage difference. While, during transient state rotor swings and ΔV undergoes oscillations caused by change in rotor angle. It is compulsory to add additional information to a neurocontroller for damping out oscillations. Rotor speed, active power or both are usually used variables for generating stabilizing signals [5, 15, 16]. In this research work one more parameter, load angle (δ) is also included for analyzing learning performance based on its correlation with excitation voltage and active power. The selection of load angle will not negatively affect the generator dynamics because active power and load angle are proportional as evident from equation

$$\mathrm{P} = \frac{E_f * V_t}{X_S} sin\delta \tag{1}$$

where P is active power, E_f is internal generated voltage, V_T is terminal voltage, X_S is synchronous reactance of generator and δ is load angle. It is important to bear in mind that these stabilizing signals are auxiliary signals, whereas terminal voltage and quadratic voltage component are primary signals fed to regulate voltage at load terminals.

In this work single machine infinite bus (SMIB) power system model as shown in Figure 1, was considered for generating data. This model simulates a generator connected with the rest of the power system. Simulation of the model was carried out on Matlab/Simulink with generator rating 13.8KV, 150MVA, 50Hz at load (0.09+j0.056) Ω. The generator parameters and excitation system parameters are given in Tables 2 and 3, respectively. A three-phase to ground fault was simulated to analyze system transient stability. Figs 2 and 3 show terminal voltage and load angle behavior after simulation of 120ms fault at generator terminals. Both figures depict stable behavior of the generator. This implies that data were collected from a stable system.

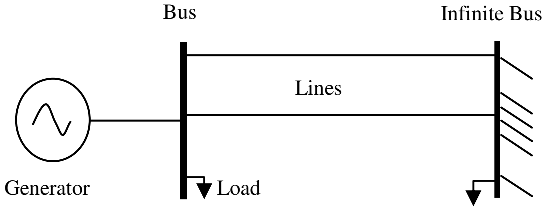


Fig. 1. A single machine-infinite bus system

Table 2. Synchronous generator parameters

X_d	=	1.83	X_q	=	1.7	R_{Stator}	=	0.003
X'_d	=	0.24	X'_q	=	0.43	Inertia	=	3.6
X''_d	=	0.20	X''_q	=	0.26	Hz	=	50
T'_d	=	0.3s	T''_d	=	0.04s	T''_q	=	0.031s

Table 3. Excitation system parameters

Ka	=	1	Ta	=	0.001s
Ke	=	2	Te	=	0.3s
Kf	=	1	Tf	=	0.003s

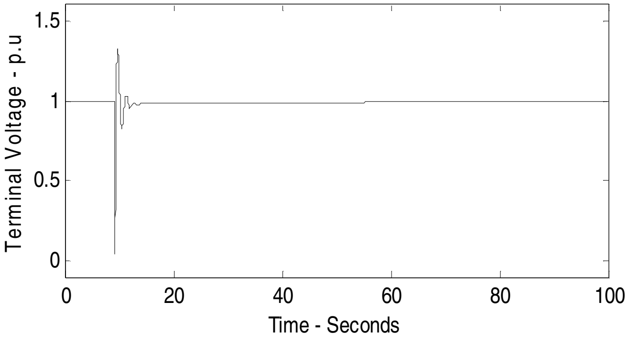


Fig. 2. Terminal voltage after 120ms fault at generator terminals

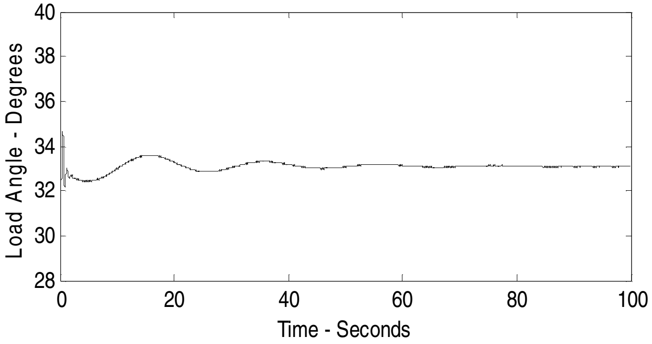


Fig. 3. Load angle (δ) transition after fault

Statistical analysis of the results was carried out using Minitab software. In statistical modeling data generation plays an important role in model acceptance. In this work aforesaid model simulation include ±10% change in Vref, three-phase to ground fault at generator terminals and transmission line tripping and addition as the types of disturbances. Data were sampled at 200Hz sampling frequency. Generated data were randomized using Matlab program, and then fifty random samples were taken for further analysis. Through graphical analysis of chosen data sample care was taken that sample should contain effect of every event.

3 Data Analysis

The statistical modeling process involved three steps: (i) correlation analysis, (ii) regression analysis, and (iii) model assessment [17]. These steps are discussed.

(a) *Correlation* is the process for determining the strength of relation between dependent and independent variables. Table 4 shows the correlation between various independent parameters and the dependent factor, excitation voltage (V_f). The table also shows the significance, also called probability value (P-value). The table contains Pearson correlation coefficient. Pearson correlation was chosen because the data is a scale type i.e value varies from -∞ to +∞.

Table 4. Statistical correlation output between Excitation Voltage (V_f) and different parameters

Independent Variables	Correlation Coefficient	Significance P-Value
ΔV_T	0.587	0.000
ω	-0.06	0.686
P	0.648	0.000
Q	0.635	0.000
ΔVq	0.759	0.000
δ	-0.392	0.005

Significance-value of less than 0.05 is considered statistically meaningful for 95% confidence level. The table suggests that the strongest correlation is between quadratic voltage (ΔV_q) and excitation voltage (V_f), followed by active (P_t) and reactive power (Q_t). The results suggest no relationship between rotor speed and excitation voltage. Rotor speed was not excluded from further analysis keeping in view that it is being used currently as auxiliary stabilizing signal. Analysis of combining each active power (Pt) and reactive power (Qt) to map Vf were carried out with each terminal voltage and quadratic voltage. However, results depicted higher VIF factor and lower R^2 value as well. Hence they were barred from further comparison.

(b) *Regression Analysis* gives the prediction of dependent variable based on independent factors of an empirical model. The regression equation is given as below [18]

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i \quad X_i + \epsilon_i \tag{2}$$

where β are constants and X is the independent variable and Y is the dependent variable. The random error term is assumed to have zero mean, constant variance σ^2 and normally distributed [18].

The accuracy of regression model is determined by the coefficient of multiple determination i.e. R^2. Higher R^2 value indicates good prediction of model. Using only R^2 is not always a good indicator of model adequacy as R^2 increases with addition of another regressor variable irrespective statistical significance of additional variable. The adjusted coefficient of multiple determination i.e. R^2(Adj) is a better reflection of the model adequacy along with R^2. R^2(Adj) will increase only when additional factor is statistically significant. Lower standard deviation is also conceived, which is an indicator for better performance of the model.

Table 5 gives the regression analysis results of the models considered. Model 1 to Model 4 consisting of different combinations of quadratic voltage (Vq), terminal voltage (Vt), reference voltage (Vref), rotor speed (ω) and load angle (δ). Prediction accuracy of models containing quadratic voltage is higher than models comprising variables in combination with terminal voltage. Whereas value of R^2(Adj) is also higher for model 3 and model 1. Therefore, it can be concluded that the set of input parameters containing quadratic voltage has higher prediction accuracy than the set consisting of terminal voltage. Hence it is expected that neurocontroller trained on model 1 or 3 may give less error than model 2 or 4.

Table 5. Regression analysis output of different models

		Model 1	Model 2	Model 3	Model 4
Constant		-6.025	-1.825	-9.230	-6.202
Vq(Coefficient)		26.167	-	29.697	-
Significance		0.000	-	0.000	-
VIF		1.5	-	2.4	-
Vt(Coefficient)		-	36.522	-	36.907
Significance		-	0.000	-	0.000
VIF		-	2.9	-	2.1
Vref(Coefficient)		-19.13	-31.84	21.85	-30.09
Significance		0.000	0.000	0.000	0.000
VIF		1.5	2.0	1.4	2.1
δ(Coefficient)		-	0.0419	0.2995	-
Significance		-	0.744	0.005	-
VIF		-	1.6	1.9	-
ω(Coefficient)		4.308	-	-	3.222
Significance		0.016	-	-	0.159
VIF		1.3	-	-	1.3
S		0.9291	1.2451	0.9073	1.2197
R^2		78.2%	60.9%	79.2%	62.5%
R^2(Adj)		76.8%	58.3%	77.9%	60.0%
ANN	**MSE**	0.6313	0.7687	0.6806	0.7496
	MAE	0.3444	0.39691	0.3672	0.4149

(c) *Model Assessment* step is carried out after the regression model has been developed. Acceptability and reliability are carried out in this step. Fitting a regression model requires few assumptions, meeting them tells credibility of the model. Performance analysis of models 2 and 4, using steps suggested in model assessment depicts pretty poor picture. However their comparison is not shown here. Instead assessment comparison between models 1 and 3 is described here.

It is assumed while fitting data that the residuals are randomly distributed and lie within ±2. The residuals from a regression model are given by

$$e = y_{des}-y_{est} \tag{3}$$

where e is the error, y_{des} is the desired output and y_{est} is the estimated output. Analysis of residuals is helpful in checking the assumption that the errors are approximately normally distributed with constant variance. Moreover, the final step for checking model adequacy is to know the behavior of model residuals.

Figs 4, 5, and 6 are related with residuals of model 1 while Figs 7, 8 and 9 are associated with residuals of model 3. Residual plot of model 1 shows more random approach than the residual plots of model 3. More than 95% residuals of model 1 lie in the range of ±2, which indicates that the assumptions of randomly distributed residual is satisfied [17], as shown by Figs 5 and 6. Whereas less than 95% of residuals of model 2 lie within range of ± 2, as indicated by Figs 8 and 9. Residual plots comparison, Figs 4 and 7, of both models exhibit that the residuals of model 1 are more random. The standard deviation (S) of model 2 is a bit higher than model 3 shown in Table 5. The low S means that data set tends to be very close to mean, whereas the assumed mean here is zero.

The distribution of residuals along regression fit is shown in Figs 5 and 8 for model 1 and model 2, respectively. The comparison of both plots expose that the distribution of model 1 residuals is approaching normality more than model 2 residuals.

The final evidence to prophesy better performance of models 1 and 2 is indicated by the variance inflation factor (VIF). It predicts co-linearity among predictors. Lesser VIF value means better performance. VIF value of model 1 is less than VIF value of model 3 as given in TABLE V.

Table 5 also exhibits the ANN results. MLP was trained on vectors of model 1 to 4. The performance of MLP was analyzed on the basis of mean square error (MSE) and mean absolute error (MAE). The MLP output is almost in proportion to the statistical methods. Models containing terminal voltage have higher error than models with quadratic voltage. However the difference between errors of different sets is not in proportion to the difference obtained using statistical methods. This manifests ANN's ability to efficiently map highly complex functions. Whereas ANN output for models 1 and 3 give different output in comparison to the statistical methods. It is easy to map excitation voltage for ANN trained on model 1 than model 3, as it is clearly evident from ANN results. Authors strongly believe that it is because of higher VIF factor of model 3 and poor model performance analysis described in model assessment section.

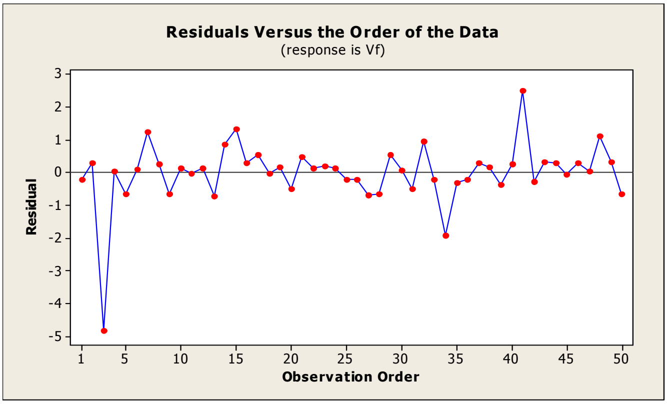


Fig. 4. Residual plots of model 1

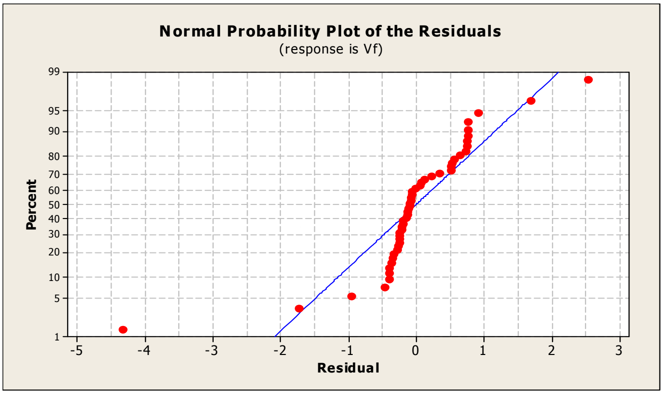


Fig. 5. Normal distribution of residual of model 1

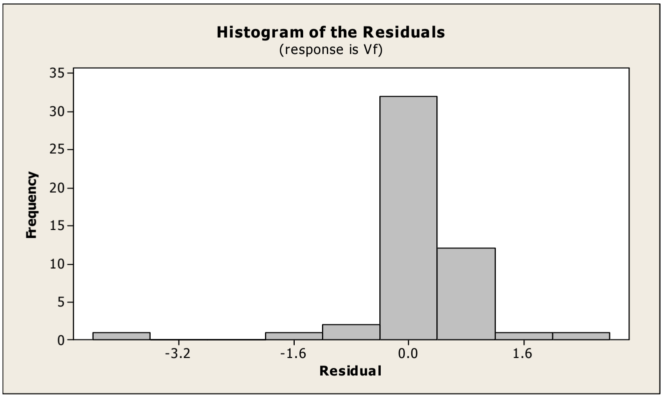


Fig. 6. Histogram of model 1 residuals

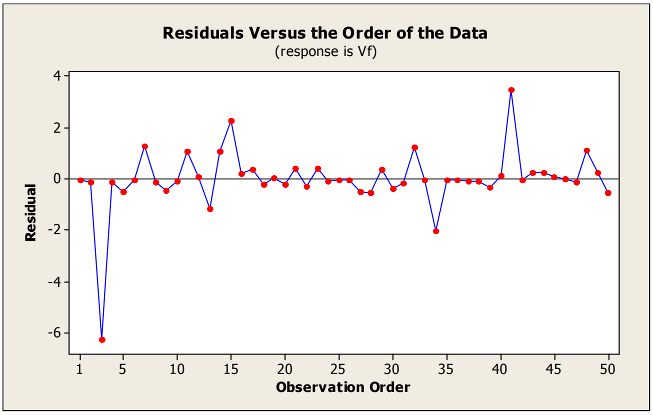


Fig. 7. Residual plots of model 3

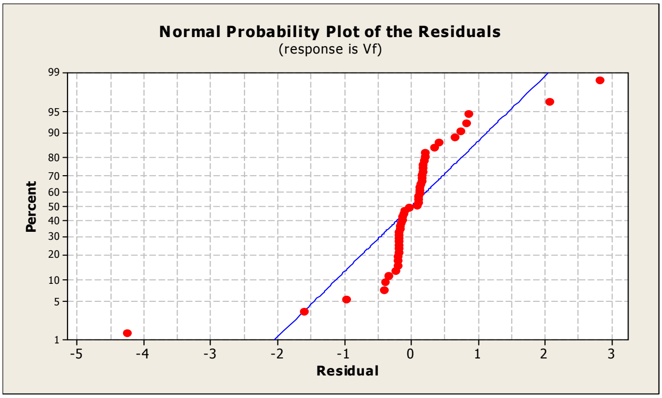


Fig. 8. Normal distribution of residual of model 3

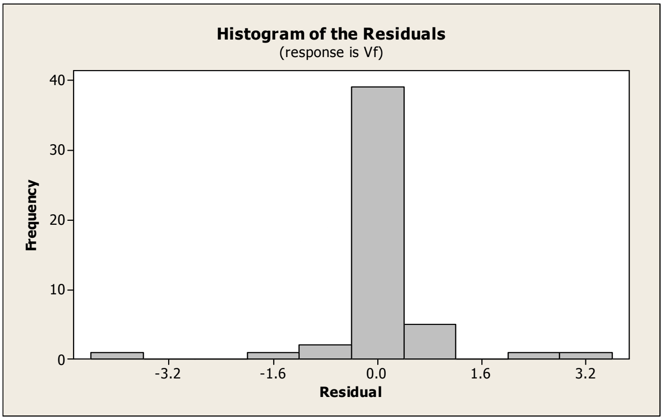


Fig. 9. Histogram of model 3 residual

4 Conclusion

With the help of statistical analysis, it is revealed that multi co-linearity and uncorrelated input-output space inhibit ANN learning capability. Instead of using terminal voltage, quadratic voltage is a better feature for training a neurocontroller for synchronous generator excitation system. Results comparison imparts that the performance of ANN is superior to linear regression.

Acknowledgement

The authors greatly acknowledge the support from the Universiti Sains Malaysia fellowship scheme for the financial support.

References

1. Linkens, D.A., Nyongesa, H.O.: Learning systems in intelligent control: an appraisal of fuzzy, neural and genetic algorithm control applications. IEE Proceedings Control Theory and Applications 143(4), 367–386 (1996)
2. Venayagamoorthy, G.K., Harley, R.G.: A continually online trained neurocontroller for excitation and turbine control of a turbogenerator. IEEE Transactions on Energy Conversion 16(3), 261–269 (2001)
3. Basheer, I.A., Hajmeer, M.: Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 43(1), 3–31 (2000)
4. Hunt, K.J., Sbarbaro, D., Zbikowski, R., Gawthrop, P.J.: Neural networks for control systems-A survey. Automatica 28(6), 1083–1112 (1992)
5. Nguyen, T.T., Gianto, R.: Neural networks for adaptive control coordination of PSSs and FACTS devices in multimachine power system. IET Generation, Transmission & Distribution 2(3), 355–372 (2008)
6. Park, J.-W., Harley, R.G., Venayagamoorthy, G.K., Gilsoo, J.: Dual heuristic programming based nonlinear optimal control for a synchronous generator. Engineering Applications of Artificial Intelligence 21(1), 97–105 (2008)
7. Sharaf, A.M., Lie, T.T.: Artificial neural network pattern classification of transient stability and loss of excitation for synchronous generators. Electric Power Systems Research 30(1), 9–16 (1994)
8. Park, J.-W., Harley, R.G., Venayagamoorthy, G.K.: Decentralized optimal neurocontrollers for generation and transmission devices in an electric power network. Engineering Applications of Artificial Intelligence 18(1), 37–46 (2005)
9. Venayagamoorthy, G.K., Harley, R.G.: Two separate continually online-trained neurocontrollers for excitation and turbine control of a turbogenerator. IEEE Transactions on Industry Applications 38(3), 887–893 (2002)
10. Swidenbank, E., McLoone, S., Flynn, D., Irwin, G.W., Brown, M.D., Hogg, B.W.: Neural Network based control for synchronous generators. IEEE Transactions on Energy Conversion 14(4), 1673–1678 (1999)
11. Djukanovic, M., Novicevic, M., Dobrijevic, D., Babic, B., Sobajic, D.J., Pao, Y.-H.: Neural-net based coordinated stabilizing control for the exciter and governor loops of low head hydropower plants. IEEE Transactions on Energy Conversion 10(4), 760–767 (1995)

12. Kim, Y.S.: Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. Expert Systems with Applications 34(2), 1227–1234 (2008)
13. Kundur, P., Paserba, J., Ajjarapu, V., Andersson, G., Bose, A., Canizares, C., Hatziargyriou, N., Hill, D., Stankovic, A., Taylor, C., Van Cutsem, T., Vittal, V.: Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. IEEE Transactions on Power Systems 19(3), 1387–1401 (2004)
14. Wang, Y., Hill, D.J., Middleton, R.H., Gao, L.: Transient stability enhancement and voltage regulation of power systems. IEEE Transactions on Power Systems 8(2), 620–627 (1993)
15. Abdelazim, T., Malik, O.P.: Fuzzy logic based identifier and pole-shifting controller for PSS application. IEEE Power Engineering Society General Meeting 3, 1680–1685 (2003)
16. Chaturvedi, D.K., Malik, O.P.: Generalized neuron-based adaptive PSS for multimachine environment. IEEE Transactions on Power Systems 20(1), 358–366 (2005)
17. Wu, Y., Zhou, Q., Chan, C.W.: A comparison of two data analysis techniques and their applications for modeling the carbon dioxide capture process. Engineering Applications of Artificial Intelligence 23(8), 1265–1276 (2010)
18. Chaloulakou, A., Saisana, M., Spyrellis, N.: Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. The Science of The Total Environment 313(1-3), 1–13 (2003)

Intelligent Decision Support Model Based on Neural Network to Support Reservoir Water Release Decision

Wan Hussain Wan Ishak[1], Ku Ruhana Ku-Mahamud[1], and Norita Md Norwawi[2]

[1] College of Arts and Sciences, Universiti Utara Malaysia, UUM Sintok, Kedah, Malaysia
{hussain, ruhana}@uum.edu.my
[2] Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai, Ng Sembilan, Malaysia
norita@usim.edu.my

Abstract. Reservoir is one of the emergency environments that required fast an accurate decision to reduce flood risk during heavy rainfall and contain water during less rainfall. Typically, during heavy rainfall, the water level increase very fast, thus decision of the water release is timely and crucial task. In this paper, intelligent decision support model based on neural network (NN) is proposed. The proposed model consists of situation assessment, forecasting and decision models. Situation assessment utilized temporal data mining technique to extract relevant data and attribute from the reservoir operation record. The forecasting model utilize NN to perform forecasting of the reservoir water level, while in the decision model, NN is applied to perform classification of the current and changes of reservoir water level. The simulations have shown that the performances of NN for both forecasting and decision models are acceptably good.

Keywords: Emergency Management, Intelligent Decision Support System, Neural Network, Forecasting.

1 Introduction

In emergency situation, “decisions must be made in human perceptual timeframes under pressure to respond to dynamic uncertain conditions” [1]. A failure to response early to the emergency situation could cause severe damages and possibility of loss of life. Flood for example, could strike without warning due to natural circumstances such as climate changes. This hazard can cause direct effect to the society such as loss of human lives and damage to the physical structures.

Moreover, “information can be inaccurate or obtained from multiple sources that are inconsistent with each other, resulting in uncertainty and information overload for the user” [2]. Uncertain conditions are very difficult to interpret as the data are vague and incomplete. Access to the data is difficult [2] and the presentation of the data is quite poor; difficult to interpret and understand [1]. In addition, the interdependence between subsequent decision and the time constraint [3] increase the complexity of

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 365–379, 2011.

the decision process. The complexity of the problem required experience decision maker to make an accurate decision [4].

An intelligent decision model, namely intelligent decision support system (IDSS) is one of the potential solutions to support decision maker in emergency situation. IDSS is an integration of DSS and Artificial Intelligence (AI). Reservoir management has been one of the potential applications in IDSS due to the complexity of the operation, expert knowledge requirement and intelligent judgement [5].

Reservoir plays an important function in water resources planning and management. Typically two categories of reservoir have been established around the world namely single and multi-purpose reservoir. Reservoir operation is influenced by it purposes [6]. The operation problem for a single-purpose reservoir is to decide the adequate release volume so that the benefits for that purpose are maximized. The operation of multi-purpose reservoir inherit the same problem, additionally, the release need to be optimally allocated among purposes. The compatibility of the purposes will affect the coordination effort and thus will increase the complexity of the reservoir operation.

The operation of multi-purpose reservoir with flood control and water supply purposes is one of the complex and dynamic problems in reservoir operation. During heavy rain, reservoir needs to impound water and release them gradually to maintain downstream discharges within the safe carrying capacity of the channel [6] and minimize the downstream damages and to ensure dam safety [7].

Conversely, during less intense rainfall, the reservoir has to impound adequate water to maintain its water level without affecting its release for water supply. Therefore, the realistic reservoir operating policies for water allocation and the optimal reservoir releases need to be established [8]. The computer based decision support system also needs the capability to adapt to the changes [2] as climate change is one of the greatest threats of this century [9].

2 Reservoir as Structural Mitigation during Emergency Situation

The reservoir is a physical structure such as pond or lake either natural or artificially developed to impound and regulate the water. It has been used as one of the structural approaches for flood defence and water storage. Flood defence is a mechanism use to modify the hydrodynamic characteristics of river flows in order to reduce the flood risk downstream [6]. Water storage is to contain water in order to maintain water supply for it use such as in agriculture, domestic and industry.

Mitigation is one of the tasks in emergency management cycle. Mitigation is a process of reducing the risk of the disaster. Mitigation related to water disaster, can be divided into structural and non-structural approaches [10]. Structural approach such as defence mechanism [6] is related to physical control of the emergency situation. Reservoir is one of the defence mechanism for both flood and drought disaster.

Defence mechanism is use to modify the hydrodynamic characteristics of river flows and coastal waters [6]. The defence can be achieved by traditional water engineering methods and by water abatement methods. Traditional water engineering methods using ‘hard’ defences include river channel modifications or using artificial materials like concrete which specifically shaped and designed to control water flows. While, water abatements methods using ‘soft’ defences rely on essentially natural materials, whether of geological or biological origin and existing environmental processes. One of the popular defence mechanisms that are currently being used by many countries in the world is dam. The use of dam for flood mitigation is aim to impound water in a reservoir during periods of high flow in order to maintain safe downstream discharges [6]. The opening of the dam’s spillway gate must be adequate to ensure that the reservoir capacity will not over its limits and the discharges will not cause overflow downstream. During drought, the reservoir needs to impound water and release adequately to fulfil its purposes.

A reservoir system can be divided into four components namely, upstream, reservoir catchment, the spillway gate, and downstream (Figure 1). The upstream consists of one or several rivers that carry the water into the reservoir. The water is stored in the reservoir catchment before releases through the spillway gate to the downstream. This kind of system is designed to ensure that during heavy rainfall, the upstream water flow does not directly flow to the downstream. The reservoir system will control the water flow and the releases within the safe carrying capacity of the downstream river [6], thus minimize the downstream damages [7].

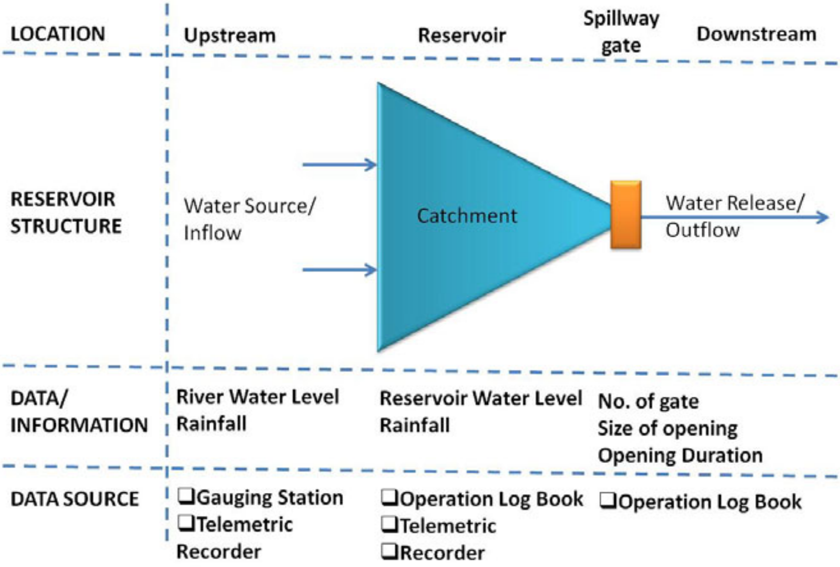


Fig. 1. Conceptual Model of Reservoir System

As shown in Figure 1, each component of the reservoir system is associated with data or information. The water level and rainfall are prevalence in both upstream and the reservoir catchments. These data are recorded hourly using the telemetric recorder situated at the strategic location of both upstream river and reservoir.

Additionally, manual reading of the rainfall also recorded through the gauging stations. At the spillway gate, the typical data are number of gate opened, the size of opening, and the opening duration. These data are recorded manually by the reservoir operator in the operation log book.

During both flood and drought situations, the decision to open and close the water gate is a critical action need to be undertaken by the dam operator as late decision will not only causing flood downstream but also will damage the dam structure. Releasing the water earlier before the reservoir reaching its full capacity might reduce the flood risk downstream. However, one cannot be sure that the water release will be replaced by the new one to serve it usage during less intense rainfall. As for multi purpose dam low water in the reservoir will cause conflict on its usage. Researchers believe that the use of forecasting and warning system might improve the dam operation and decision [10].

Forecasting and warning system are example of the non-structural approach. Non-structural approach is non-physical control in which the emergency is control using a procedure. Example of the non-physical control is flood insurance [11], flood zoning [12] and flood forecasting [13]. In term of implementation, structural solution cost higher compare to non-structural solution. However, non-structural solution is constraint to its political implementation [10].

The combine implementation of structural and non-structural approach is vital to avoid casualty and false sense of security at storage [14]. Structural approach such as dam is a solid structure that holds water for a certain period or at maximum reservoir water level. In practice, the water release or the gate opening decision depends on the operating rules [15]. These rules are static and do not consider the dynamic nature of the hydrology systems. Therefore, non-structural approach such as forecasting is vital to support the water release or the gate opening decision. The dynamic of the forecasting system will be able to cope with the event frequency and triggered alert to the authority when the situation is at the severe level. Flood forecasting is significant to cope with the great floods [16].

3 Intelligent Decision Support System Framework

Intelligent Decision Support System is an integration of DSS and artificial intelligence (AI) technology combining the basic function of DSS and reasoning capabilities of AI techniques [17]. AI is viewed as a system that has the ability to “think” and “act” [18]. Based on discussion in Russell and Norvig, AI definitions can be viewed into two dimensions (Figure 2). In the first dimension, AI can be regarded as a system that think like humans or that thinks rationally. In second dimension, AI is viewed as a system that acts like humans or that act rationally.

Guerlain et al. [19] has identified six characteristics of successful IDSS; interactivity, representation aiding, event and change detection, error detection and recovery, information extraction and predictive capabilities. Interactivity is the interaction between the system and user. The IDSS is expected to support interactivity, in which user can present the input and receive the output as the feedback. The presentation of the information on the interface should be readable and understandable by the user. The system should be aid the user and explain the output or how it derives the conclusion. The intelligent capability of the system should be intelligent enough to detect and adapt the changes in user input or the surroundings which might influence the operation or the processes in the system decision making. The system should be prone to error by integrating the error detection and recovery facility.

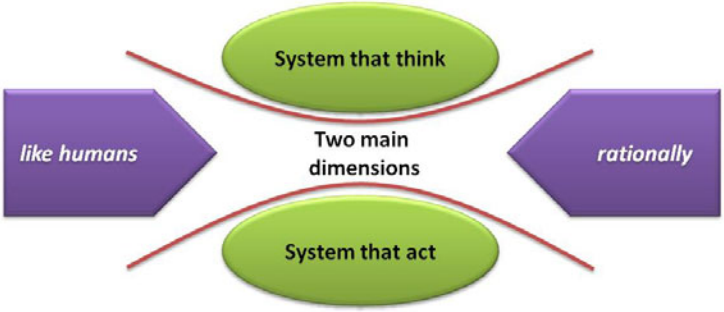


Fig. 2. Dimension in AI Technology

The most important component of the IDSS is the information extraction and predictive capability. Information extraction is the capability to extract the useful information from the abundance of information. This information will serve as the input to the IDSS or to be represented to the user in a meaningful format. The IDSS predictive capability will use and analyze the information into a pattern by which represents a trend of the event. This trend will be learned and to be used to predict the future event. These facilities exhibit intelligent behaviour of the IDSS and very useful in assisting the decision maker.

The theoretical framework in Figure 3, shows the mapping between the conceptual and computational level, and the relationship between the emergency environment and the real world practice. The emergency situations inherit several characteristics namely, dynamic, urgency, uncertain, complex, high risk, and previous action dependent. These characteristics are also a part of the problem that solved through naturalistic decision making (NDM). This problem can be solved using adaptive and dynamic system approach: Intelligent Decision Support System (IDSS). The proposed IDSS consists of three main sub-models: situation assessment model, expectancy forecasting, and decision model. Situation assessment model can utilized data mining technique to extract temporal data sets from the data sources. Expectancy forecasting model is to forecast the future effect of the known factors. Neural network can be used as the forecasting technique. The decision model will produced the final output of the IDSS. In this model, expert system, fuzzy logic or neural network can be utilized as the decision engine.

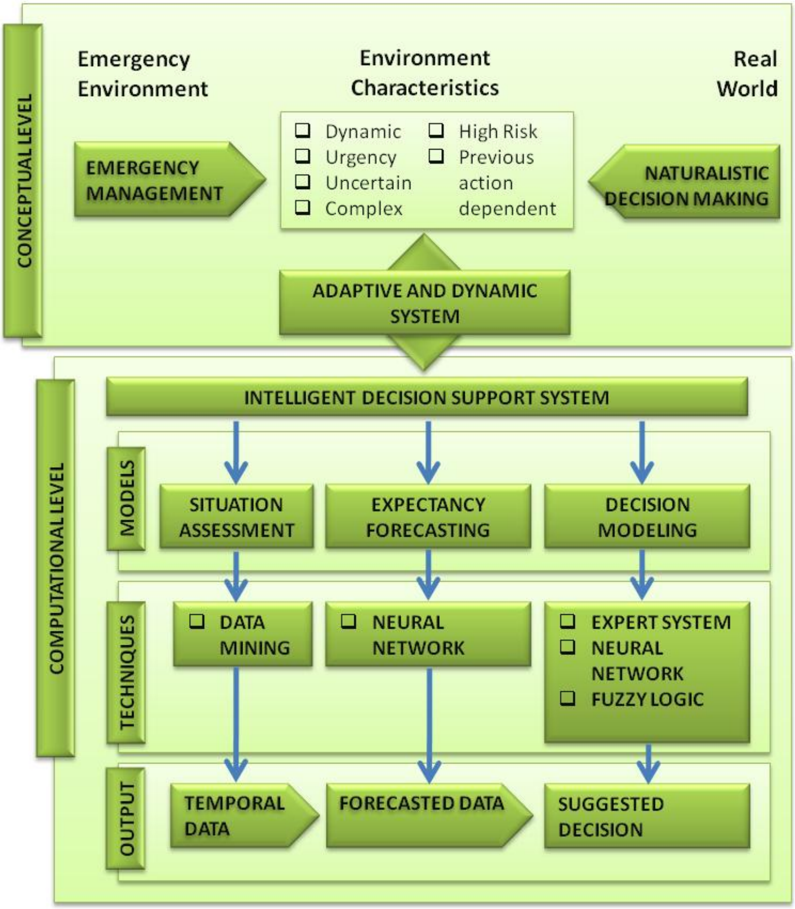


Fig. 3. Theoretical Framework

Figure 4 shows the conceptual of model IDSS for reservoir operation. As shown in Figure 4, data mining will combine both hydrological and operational data and extract the temporal data that maintain the temporal relationship of the data. The extraction process will include data integration, data preprocessing, temporal data mining, and post processing. The extracted data will be feed into water level forecasting model, which will calculate the probability of the rising of reservoir water level using neural network. The result of this model is the forecasted water level at time $t+1$. The forecasted data will be used in the decision model. Finally, the gate opening decision will be produced.

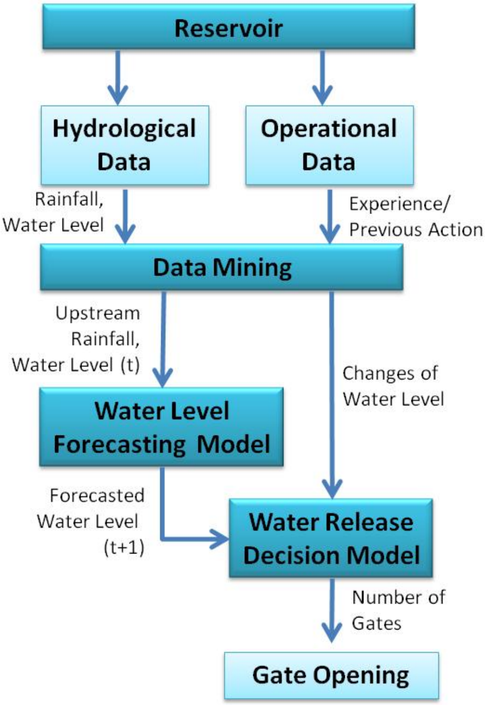

Reservoir
Hydrological Data
Operational Data
Rainfall, Water Level
Experience/ Previous Action
Data Mining
Upstream Rainfall, Water Level (t)
Changes of Water Level
Water Level Forecasting Model
Forecasted Water Level (t+1)
Water Release Decision Model
Number of Gates
Gate Opening

Fig. 4. Conceptual Model of IDSS for Reservoir Operation

4 Neural Network Application in Reservoir Operation

Neural network (NN) is a mathematical computational model that imitates the biological neuron capability. One of the main features of neural network is it be able to learn a pattern and apply the "knowledge" to the similar pattern. This capability enable neural network to be used to solve wide range of problems including forecasting and classification problem. In the application of reservoir operation and management, NN has been applied for various simulation and optimization problem. Table 1 summarizes some of the related studies and NN model implemented.

Table 1. Related Studies and NN Application in Reservoir Operation and Management

Studies	Application	NN Model
Hu et al., [20]	River Flow Prediction	Range-Dependent NN(RDNN)
Dibike and Solomatine [21]	River Flow Forecasting	Multi-Layer Perceptron Network (MLP) & Radial Basis Function Network (RBF)
Chang and Chen [22]	Streamflow Prediction	Counterpropagation Fuzzy-NN (CFNN)

Table 1.(*continued*)

Kisi [23]	Streamflow Prediction	Backpropagation NN
Coulibaly et al. [24]	Multivariate Reservoir Inflow Forecasting	Temporal NNs
Coulibaly et al. [25]	Daily Reservoir Inflow Forecasting	Multi-layer Feed-Forward NN (FNN)
Chang and Chang [26]	Prediction of Reservoir Water Level	Adaptive Network-Based Fuzzy Inference System (ANFIS)
Lobbrecht and Solomatine [27]	Controlling the Polder Water Levels	ANN and Fuzzy Adaptive Systems (FAS)
Solomatine and Xue [28]	Flood Forecasting	Multilayer Perceptron & Hybrid (M5 & MLP)
Kumar et al. [29]	Flood Control Operation and Conservation Operation	Standard Backpropagation Algorithm
Chaves and Chang [30]	Intelligent Reservoir Operation System	Evolving ANN

5 Method

In this study, standard backpropagation neural network with bias, learning rate and momentum are used in both forecasting and decision model. In forecasting model, neural network is used to train the rainfall data (at *t*) and to create a mapping with the reservoir water level at *t+1*. In the decision model, neural network is used to train the water level (at *t* and t+1) and the changes of water level. The output produce by the decision model is the number of gate to be opened. The temporal information of the rainfall and water level data are preserve by using sliding window technique. Once data has been prepared, the training was conducted base on the standard training procedure.

5.1 Case Study: Timah Tasoh Reservoir

In this study, Timah Tasoh reservoir was used as a case study. Timah Tasoh reservoir is one of the largest multipurpose reservoirs in northern Peninsular Malaysia. Timah Tasoh located on Sungai Korok in the state of Perlis, about 2.5km below the confluence of Sungai Timah and Sungai Tasoh. Timah Tasoh reservoir covered the area of 13.33 Km^2 with the catchment area 191.0 Km^2. Its maximum capacity is 40.0 Mm^3. Timah Tasoh reservoir serves as flood mitigation in conjunction to other purposes: water supply and recreation. Water from Timah Tasoh is used for domestic, industrial and irrigation.

5.2 Data Preparation

Reservoir water level is influence by a number of factors such as upstream rainfall, water flow, heat and temperature, and evaporation rate. However, technological and political constraints have limited the availability of the data. In this study, a total of 3041 daily data from Jan 1999 – April 2007 were gathered from the Timah Tasoh reservoir operation record. Timah Tasoh upstream rainfall was manually recorded through 5 upstream gauging stations. Rainfall observed from these stations will eventually increase the reservoir water level.

For the forecasting model, rainfall data from these stations and the current reservoir water level (*t*) are used as the input data and the reservoir water level at time *t+1* is used as the target. In the decision model the current water level (*t*), tomorrow water level (*t+1*), and the changes of water level at *t*, *t-1*, …, *t-w* were used as the input data, while the gate opening/closing at *t* is used as the target. The constant *t* and *w* represent time and days of delays (which later represented as window size). Gate opening/closing value is in range of zero to six. Zero indicates gate is closed and values from one to six indicate the number of gates that are open. The change of this value implies the decision point. At this point window slice will be formed begin from that point and preceding to *w* days according the window size.

Sliding window technique is used to capture the time delay within the data set. Sliding window technique was proven able to detect patterns from temporal data [31]. This process is called segmentation process. For both forecasting and decision model, nine data sets have been formed. Each data set represents different sliding size. Each sliding size represent time duration of the delays. For example, sliding size 2 represents two days of delays. Table 2 summarizes the number of instances extracted for each data set. Segmentation process for decision model will return a total of 124 instances. Redundant and conflicting instances are then removed.

Table 2. Data set and the number of instances

Data Set	Sliding Size	Number of Instances	
		Forecasting Model	Decision Model
1	2	2075	43
2	3	2408	54
3	4	2571	71
4	5	2668	82
5	6	2732	95
6	7	2774	109
7	8	2805	113
8	9	2826	118
9	10	2844	119

Each data set consists of *N* number of input columns and 1 output column. The output consists of 4 classes. The input is then normalized using Min-Max method (Equation 1) to transform a value *x* to fit in the range [*C*,*D*]. Where, *C* is the new minimum (-1) and *D* is the new maximum (1) values. In this study the new value is set in range of [-1,1]. The output is encoded based on Binary-Coded-Decimal (BCD) scheme. BCD is preferably as the total number of output nodes can be reduced to the integer of $Log_2\ M$, where *M* is the number of classes [32].

$$New(x) = \left[\frac{x - \min(x)}{\max(x) - \min(x)}\right] * (D - C) + C \tag{1}$$

Each data set is then divided randomly into three data sets: training set (80%), validation set (10%) and testing set (10%). Training set is used in the training phase of neural network, while validation set is used to validate the neural network performance during the training. Testing set is used to test the performance of neural network after the training has completed.

5.3 Neural Network Modelling

The aim of neural network modelling is to create a mapping between the input data and the target output. This mapping was established by training the neural network to minimize the error between the network output and the target (Equation 2). In this study, nine neural network models were developed for both forecasting and decision model. Each neural network model is trained with one data set. Each model is trained with different combination of hidden unit, learning rate and momentum. The training is control by three conditions (1) maximum epoch (2) minimum error, and (3) early stopping condition. Early stopping is executed when the validation error continue to arises for several epochs [33]. Fig. 5 shows the procedure for the neural network training. The aim of this procedure is get the neural network model that gives the best result.

$$SE = \frac{1}{2}\sum_{k=1}^{m}(t_k - y_k)^2 \qquad (2)$$

```
for each hidden unit (HU)
where HU = {3,5,7,9,11,13,15,17,19,21,23,25}

     for each learning rate (LR)
     where LR = {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}

        for each momentum (Miu)
        where Miu = {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}

          Training:
              Feedforward()
              Backpropagation of error()
              Weight update()

          Validation()

        end loop (Miu)
     end loop (LR)
end loop (HU)
```

Fig. 5. Pseudo Code for Neural Network Training

6 Findings

6.1 Forecasting Model

Table 3 shows the results for each data set after training and testing for the forecasting model. Overall the minimum training, validation and testing error are 0.461878, 0.41825 and 0.416571 respectively. The best result achieved for training, validation and testing are 89.99%, 91.34% and 91.52% respectively. There is a small difference between the highest and lowest results achieve from training, validation and testing. The difference shows that neural network has learned the data quite well. Based on the results, data set 7 is chosen as the best data set for reservoir water level forecasting model. The result for training, validation and testing are 89.61, 91.34 and 90.75. Data set 7 was formed using sliding size 8 which contains 2805 instances.

Table 3. Results of Training, Validation and Testing

Data Set	Training		Validation		Testing	
	(%)	Error	(%)	Error	(%)	Error
1	87.48	0.785791	86.22	0.860958	89.26	0.667375
2	87.92	0.58714	87.00	0.573727	87.56	0.586856
3	87.65	0.599483	89.75	0.457907	89.36	0.490453
4	89.45	0.492463	88.52	0.502691	90.76	0.444052
5	89.50	0.483055	89.87	0.50378	90.36	0.503575
6	89.43	0.480323	90.74	0.421007	89.05	0.534949
7	89.61	0.474844	91.34	0.41825	90.75	0.443816
8	89.99	0.461878	89.52	0.474101	91.52	0.416571
9	89.77	0.467551	90.85	0.430233	90.73	0.4428
Min	87.48	0.461878	86.22	0.41825	87.56	0.416571
Max	89.99	0.785791	91.34	0.860958	91.52	0.667375

Values for the network parameters that were achieved from the training phase are shown in Table 4. As for data set 7, the total epoch is 21 and the best result achieved was with both learning rate (LR) and momentum (Mom) equal to 0.2. The best network architecture achieved is 24-15-3.

Table 4. Neural Network Parameters

Data Set	Epoch	#Input	#hidden unit	#output unit	LR	Mom
1	88	6	31	3	0.7	0.5
2	91	9	35	3	0.4	0.4
3	39	12	21	3	0.5	0.2
4	21	15	7	3	0.3	0.1
5	46	18	3	3	0.3	0.1
6	21	21	5	3	0.3	0.1
7	21	24	15	3	0.2	0.2
8	21	27	23	3	0.1	0.3
9	21	30	21	3	0.2	0.1

6.2 Decision Model

The results of neural network training, validation, and testing for the decision model are shown in Table 5. Overall, the lowest error achieve for training, validation and testing was 0.065795, 1.59E-07, and 9E-10 respectively. The best results of training, validation, and testing was 98.35%, 100%, and 100% respectively. These results show that neural network classifier has performed very well on temporal data set. Based on the results in Table 3, data set 4 is chosen to be the best data set. Neural network train with data set 4 achieves 93.94% of training performance and 100% of validation and testing performance. The error was 0.23505, 0.023383, and 0.007085 respectively. Data set 4 was formed with window size 5 with 82 instances.

Table 5. Results of Training, Validation and Testing

Data Set	Training		Validation		Testing	
	%	Error	%	Error	%	Error
1	90.00	0.39996	87.50	0.5	100.00	9E-10
2	90.91	0.362563	100.00	0.007216	100.00	6.13E-05
3	95.62	0.147186	85.72	0.626408	100.00	0.034537
4	93.94	0.23505	100.00	0.023383	100.00	0.007085
5	89.34	32.00295	100.00	1.59E-07	100.00	1.4E-07
6	97.70	0.092475	95.46	0.188657	100.00	0.002146
7	98.35	0.065796	100.00	0.032103	95.46	0.191186
8	93.09	0.276602	95.84	0.166669	95.84	0.168359
9	97.37	0.104647	95.84	0.171619	100.00	0.003985
Min	89.34	0.065795	85.72	1.59E-07	95.455	9E-10
Max	98.35	32.00295	100	0.626408	100	0.191186

Values for the network parameters that were achieved from the training phase are shown in Table 6. As for data set 4, the total epoch is 86 and the best result achieved was with learning rate (LR) 0.8 and momentum (Mom) 0.2. The best network architecture achieved is 8-23-2.

Table 6. Neural Network Parameters

Data Set	Epoch	#Input	#Hidden Unit	#Output Unit	LR	Mom
1	77	5	25	2	0.9	0.4
2	42	6	23	2	0.8	0.4
3	33	7	17	2	0.7	0.3
4	86	8	23	2	0.8	0.2
5	31	9	9	2	0.9	0.8
6	31	10	7	2	0.7	0.5
7	54	11	5	2	0.5	0.5
8	42	12	25	2	0.4	0.8
9	27	13	9	2	0.4	0.6

7 Discussion

The sliding window technique has been successfully applied on reservoir water level data to extract and segment the data to preserve the temporal relationship of the data. It was shown in Table 1 that the size of window has influence the number of usable instances. The bigger the window size the larger the usable instances. The large number of usable instances will contains large number of temporal patterns that can be used for neural network modeling. The large size of data is vital as the performance of neural network model is highly influenced by the size of data set. However, as the data size increase the number of input also increase. The large number of input unit will increase the complexity of the neural network modeling.

The finding of this study also suggests that 8 days are the best time duration for the delay. This suggests that 8 days observation of the upstream rainfall will significantly increase the water level at the reservoir. Additionally, 5 days of observed water level changes has been found to be significant of the reservoir water release decision. This information is vital for reservoir management to plan early water release.

The reservoir water level data typically the current, the (expected) tomorrow water level and the changes of water level are extracted from the reservoir operation record. In actual reservoir operation and decision making, the current water level represent the current stage of reservoir water level (*t*), while the tomorrow water level is water level that is expected for tomorrow at *t+1*. As shown in this paper, the water level can be forecasted based hydrological variables. The changes of reservoir water level represent the increase or decrease of reservoir water level. Observing the changes of reservoir water level at time *t* and the preceding *t-1, t-2, ..., t-w* will give an insight on when to release the reservoir water.

8 Conclusion

The findings of this study can be used to aid reservoir water release decision. Typically, reservoir water release decision was influenced by the upstream rainfall. Since upstream rainfall was recorded through upstream gauging stations which are located quite far from the reservoir and river water might be lost due to environmental factors, the time delay is expected before the rain water can give effect to the reservoir water level. In this study, window sliding has been shown to be a successful approach to model the time delays, while neural network was shown as a promising modelling technique.

Manually, reservoir operator monitors the changes of water level and consults the superior officer before taking the appropriate action. Having unpredicted circumstances of the weather, early decision of the reservoir water release is always a difficult decision. Information on the delay and the forecasted reservoir water level can be used by reservoir operator to decide early water release. Early water release of the reservoir will reserve enough space for incoming inflow due to heavy upstream rainfall. In addition, the water release can be controlled within the capacity of the downstream river. Thus flood risk downstream due to extreme water release from the reservoir can be reduced.

Acknowledgments. The authors' most appreciation to the Perlis Department of Drainage and Irrigation for permission and supplying Timah Tasoh reservoir operational data.

References

1. Gaynor, M., Seltzer, M., Moulton, S., Freedman, J.: A Dynamic, Data-Driven, Decision Support System for Emergency Medical Services. In: Sunderam, V.S., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2005. LNCS, vol. 3515, pp. 703–711. Springer, Heidelberg (2005)
2. Philips-Wren, G.: Adaptive Decision Support for Dynamic Environments. New Ad-vances in Intelligent Decision Technologies 199, 235–243 (2009)
3. Feigh, K., Pritchett, A.: Design of Support Systems for Dynamic Decision Making in Airline Operations. In: Proceedings of the 2006 Systems and Information Engineering De-sign Symposium, pp. 136–141 (2006)
4. Sinha, R.: Impact of Experience on Decision Making in Emergency Situation. Psychol-ogy C/D: Extended Essay. Department of Human Work Sciences, Lulea University of Technology (2005)
5. Simonovic, S.P.: Decision support system for flood management in the Red River Basin. Canadian Water Resources Journal 24(3), 203–223 (1999)
6. Smith, K., Ward, R.: Floods: Physical Processes and Human Impacts. John Wiley, England (1998)
7. Jain, S.K., Singh, V.P.: Reservoir Operation. In: Jain, S.K., Singh, V.P. (eds.) Water Resources Systems Planning & Management, ch.11, vol. 51, pp. 615–679. El-sevier B. V, Amsterdam (2003)
8. Chang, T.J., Kleopa, X.A., Teoh, C.B.: Use of Flood-Control Reservoirs for Drought Management. Journal of Irrigation and Drainage Engineering 121(1), 34–42 (1995)
9. IFRC Annual Report 2007. International Federation of Red Cross and Red Crescent Societies (2007)
10. Tucci, C.E.M.: Flood Flow Forecasting. Presented at 54th Session of Executive Coun-cil of WMO World Meteorological Organization, Geneva (2002)
11. Federal Emergency Management Agency. Flood Insurance: The Right Choice: NFIP Fact Sheet 2008. Federal Emergency Management Agency, U.S. Department of Home-land Security (2008)
12. Baldwin County Planning and Zoning Department. Exploring the Baldwin County Flood Zoning Plan and the Benefits of Flood Hazard Mitigation. White Paper. Baldwin County Commission, Alabama (2007)
13. Manusthiparom, C., Apirumanekul, C., Mahaxay, M.: Flood Forecasting and River Monitoring System in the Mekong River Basin. Second Southeast Asia Water Forum (2005)
14. Technical Support Unit. Integrated Flood Management. APFM Technical Document No. 1 (2nd), The Associated Programme on Flood Management. (2004)
15. Wurbs, R.A.: Reservoir-System Simulation and Optimization Models. Journal of Water Resources Planning and Management 119(4), 455–472 (1993)
16. Calenda, G., Mancini, C.P.: The Role of the Corbara Reservoir on the Tiber River in the Flood Protection of the Town of Rome, Italy. In: Brookshie, P.A. (ed.) Proceedings of Waterpower Conference, ASCE Research Library (1999)

17. Zhou, F., Yang, B., Li, L., Chen, Z.: Overview of the New Types of Intelligent Decision Support System. In: Proceedings of International Conference on Innovative Computing Information and Control, p. 267 (2008)
18. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 2nd edn. Pearson Education Inc., New Jersey (2003)
19. Guerlin, S., Brown, D.E., Mastrangelo, C.: Intelligent Decision Support Systems. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 1934–1938 (2000)
20. Hu, T.S., Lam, K.C., Ng, S.T.: River flow time series prediction with a range-dependent neural network. Hydrohgical Sciences 46(5), 729–745 (2001)
21. Dibike, Y.B., Solomatine, D.P.: River flow forecasting using artificial neural networks. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26(1), 1–7 (2001)
22. Chang, F.-J., Chen, Y.-C.: A Counterpropagation Fuzzy-Neural Network Modeling Approach to Real Time Streamflow Prediction. Journal of Hydrology 245, 153–164 (2001)
23. Kisi, O.: River Flow Modeling Using Artificial Neural Networks. Journal of Hydrologic Engineering 9(1), 60–63 (2004)
24. Coulibaly, P., Anctil, F., Bobee, B.: Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. Journal of Hydrology 230, 244–257 (2000)
25. Coulibaly, P., Anctil, F., Bobee, B.: Multivariate Reservoir Inflow Forecasting using Temporal Neural Networks. Journal of Hydrologic Engineering 6(5), 367–376 (2001)
26. Chang, F.-J., Chang, Y.-T.: Adaptive Neuro-Fuzzy Inference System for Prediction of Water Level in Reservoir. Advances in Water Resources 29, 1–10 (2006)
27. Lobbrecht, A.H., Solomatine, D.P.: Control of water levels in polder areas using neural networks and fuzzy adaptive systems. Water Industry Systems: Modelling and Optimization Applications 1, 509–518 (1999)
28. Solomatine, D.P., Xue, Y.: M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. Journal of Hydrologic Engineering 9(6), 1–10 (2004)
29. Kumar, A.R.S., Jain, S.K., Agarwal, P.K.: Application of Artificial Neural Networks (ANN) in Reservoir Operation. Technocal Report No. TR/BR-6/1999-2000, National Institute of Hydrology, India (1999)
30. Chaves, P., Chang, F.-J.: Intelligent Reservoir Operation System Based on Evolving Artificial Neural Networks. Advances in Water Resources 31, 926–936 (2008)
31. Ku-Mahamud, K.R., Zakaria, N., Katuk, N., Shbier, M.: Flood Pattern Detection Using Sliding Window Technique. In: Third Asia International Conference on Modeling & Simula-tion, pp. 45–50 (2009)
32. Chong, C.C., Jia, J.C.: Assessments of neural network output codings for classification of multispectral images using Hamming distance measure. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 2, pp. 526–528 (1994)
33. Sarle, W.: Stopped Training and Other Remedies for Overfitting. In: Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics, pp. 352–360 (1995)

The Effect of Adaptive Momentum in Improving the Accuracy of Gradient Descent Back Propagation Algorithm on Classification Problems

M.Z. Rehman and N.M. Nawi

Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia (UTHM).
P.O. Box 101, 86400 Parit Raja, Batu Pahat, Johor Darul Takzim, Malaysia
hi090004@siswa.uthm.edu.my, nazri@uthm.edu.my

Abstract. The traditional Gradient Descent Back-propagation Neural Network Algorithm is widely used in solving many practical applications around the globe. Despite providing successful solutions, it possesses a problem of slow convergence and sometimes getting stuck at local minima. Several modifications are suggested to improve the convergence rate of Gradient Descent Back-propagation algorithm such as careful selection of initial weights and biases, learning rate, momentum, network topology, activation function and 'gain' value in the activation function. In a certain variation, the previous researchers demonstrated that in "feed-forward algorithm", the slope of activation function is directly influenced by 'gain' parameter. This research proposed an algorithm for improving the current working performance of Back-propagation algorithm by adaptively changing the momentum value and at the same time keeping the 'gain' parameter fixed for all nodes in the neural network. The performance of the proposed method known as 'Gradient Descent Method with Adaptive Momentum (GDAM)'is compared with the performances of 'Gradient Descent Method with Adaptive Gain (GDM-AG)' and 'Gradient Descent with Simple Momentum (GDM)'. The learning rate is kept fixed while sigmoid activation function is used throughout the experiments. The efficiency of the proposed method is demonstrated by simulations on three classification problems. Results show that GDAM is far better than previous methods with an accuracy ratio of 1.0 for classification problems and can be used as an alternative approach of BPNN.

Keywords: gradient descent, neural network, adaptive momentum, adaptive gain.

1 Introduction

Artificial Neural Networks (ANN) are modeled to mimic the biological neurons in a human brain. Just like a human brain, ANN consists of processing units known as Artificial Neurons that can be trained to perform complex calculations. Unlike traditional methods in which an output is based on the input it gets, a neuron can be

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 380–390, 2011.

trained to store, recognize, estimate, and adapt to new patterns without having the information about the form of function. Since its advent, ANN has shown its mettle in solving many complex real world problems such as predicting future trends based on the huge historical data of an organization. ANN have been successfully implemented in all engineering fields such as biological modeling, decision and control, health and medicine, engineering and manufacturing, marketing, ocean exploration and so on [1-5].

A standard multilayer feed-forward neural network consists of an input layer, hidden layer and an output layer of neurons. Every node in a layer is connected to every other node in the neighboring layer. Back-Propagation Neural Network (BPNN) algorithm is the most popular and the oldest supervised learning multilayer feed-forward neural network algorithm proposed by Rumelhart, Hinton and Williams [6]. The BPNN learns by calculating the errors of the output layer to find the errors in the hidden layers. Due to this ability of Back-Propagating, it is highly suitable for problems in which no relationship is found between the output and inputs. Due to its flexibility and learning capabilities it has been successfully implemented in wide range of applications [7]. Although BPNN has been used successfully it has some limitations. Since it uses gradient descent learning rule which requires careful selection of parameters such as network topology, initial weights and biases, learning rate value, activation function, and value for the gain in the activation function should be selected carefully. An improper choice of these parameters can lead to slow network convergence, network error or failure. Seeing these problems, many variations in gradient descent BPNN algorithm have been proposed by previous researchers to improve the training efficiency. Some of the variations are the use of learning rate and momentum to speed-up the network convergence and avoid getting stuck at local minima. These two parameters are frequently used in the control of weight adjustments along the steepest descent and for controlling oscillations [8].

2 BPNN with Momentum Coefficient (α)

Momentum-coefficient (α) is a modification based on the observation that convergence might be improved if the oscillation in the trajectory is smoothed out, by adding a fraction of the previous weight change [6], [9]. Thus, the addition of momentum-coefficient can help smooth-out the descent path by preventing extreme changes in the gradient due to local anomalies [10]. So it is essential to suppress any oscillations that results from the changes in the error surface [11].

In initial studies, momentum-coefficient was kept fixed but later studies on static momentum-coefficient revealed that Back-propagation with Fixed Momentum (BPFM) shows acceleration results when the current downhill of the error function and the last change in weights are in similar directions, when the current gradient is in an opposing direction to the previous update, BPFM will cause the weight direction to be updated in the upward direction instead of down the slope as desired, so in that case it is necessary that the momentum-coefficient should be adjusted adaptively instead of keeping it fixed [12], [13].

Over the past few years several adaptive-momentum modifications are proposed by researchers. One such modification is Simple Adaptive Momentum (SAM) [14], proposed to further improve the convergence capability of BPNN. Sam works by scaling the momentum-coefficient according to the similarities between the changes in the weights at the current and previous iterations. If the change in the weights is in the same ‘direction’ then the momentum-coefficient is increased to accelerate convergence to the global minima otherwise momentum-coefficient is decreased. SAM is found to lower computational overheads then the Conjugate Gradient Descent and Conventional BPNN. In 2009, R. J. Mitchell adjusted momentum-coefficient in a different way than SAM [14], the momentum-coefficient was adjusted by considering all the weights in the Multi-layer Perceptrons (MLP). This technique was found much better than the previously proposed SAM [14] and helped improve the convergence to the global minima possible [15].

In 2007, Nazri *et al.* [16] found that by varying the gain value adaptively for each node can significantly improve the training time of the network. Based on Nazri *et al.* [16] research, this paper propose further improvement on the current working algorithm that will change the momentum value adaptively by keeping the gain value fixed.

3 The Proposed Algorithm

There are certain reasons for the slow convergence of the Gradient Descent Back propagation algorithm. Mostly, the magnitude and direction components of the gradient vector play a part in the slow convergence. When the error surface is fairly flat along a weight dimension, the derivative of the weight is small in magnitude. Therefore many steps are required and weights are adjusted by a small value to achieve a significant reduction in error. On the other hand, if the error surface is highly curved along a weight dimension, the derivative of the weight is large in magnitude. Thus, large weight value adjustments may overshoot the minimum of the error surface along that weight dimension. Another reason for the slow rate convergence of the gradient descent method is that the direction of the negative gradient may not point directly toward the minimum error surface [17].

In-order to increase the accuracy in the convergence rate and to make weight adjustments efficient on the current working algorithm proposed by Nazri *et al.* [16], a new Gradient Descent Adaptive Momentum Algorithm (GDAM) is proposed in the following section.

3.1 Gradient Descent Adaptive Momentum (GDAM) Algorithm

The Gradient Descent Back propagation uses two types of training modes which are incremental mode and batch mode. In this paper, batch mode training is used for the training process in which momentum, weights and biases are updated for the complete training set which is presented to the network. The following iterative algorithm known as Gradient Descent Adaptive Momentum Algorithm (GDAM) is proposed which adaptively changes the momentum while it keeps the gain and learning rate fixed for each training node. Mean Square Error (MSE) is calculated after each epoch and compared with the target error. The training continues until the target error is achieved.

```
   For each epoch,
        For each input vector,
      Step-1:
         Calculate the weights and biases using the previous mo-
      mentum value
      Step-2:
         Use the weights and biases to calculate new momentum
      value.
        End input vector
   IF Gradient is increasing, increase momentum
   ELSE decrease momentum
   Repeat the above steps until the network reaches the desired
value.
   End epoch
```

The gradient descent method is utilized to calculate the weights and adjustments are made to the network to minimize the output error. The output error function at the output neuron is defined as;

$$E = \frac{1}{2}\sum_{k=1}^{n} (t_k - o_k(\alpha_k))^2 \tag{1}$$

An activation function plays a vital role in limiting the amplitude of the output neuron and generates an output value in any predefined range. Back propagation supports a lot of activation functions such as tangent, linear, hyperbolic and log-sigmoid function etc. This research will use log-sigmoid activation function which has a range of [0,1] in finding the output on the jth node;

$$O_j = \frac{1}{1 + e^{-a_{net,j}}} \tag{2}$$

where,

$$a_{net,j} = \left[\sum_{i=1}^{l} w_{ij} O_i\right] + \theta j \tag{3}$$

where,

n : number of output nodes in the output layer.
t_k : desired output of the k^{th} output unit.
o_k : network output of the k^{th} output unit.
O_j : Output of the *jth* unit.
O_i : Output of the *ith* unit.

W_{ij} : weight of the link from unit *i* to unit *j*.

$a_{net,j}$: net input activation function for the *jth* unit.

θ_j : bias for the *jth* unit.

$\frac{\partial E}{\partial \alpha_k}$, needs to be calculated for the output units and $\frac{\partial E}{\partial \alpha_j}$ is also required to be calculated for hidden units, so that the respective momentum value can be updated in the Equation (6):

$$\Delta \alpha_k = \left(- \frac{\partial E}{\partial \alpha_k} \right) \tag{4}$$

$$\Delta \alpha_j = \left(- \frac{\partial E}{\partial \alpha_j} \right) \tag{5}$$

$$\frac{\partial E}{\partial \alpha_k} = (t_{k-}O_k)O_k(1 - O_k)(\sum w_{jk}O_j + \theta_k) \tag{6}$$

The momentum update expression from input to output nodes becomes;

$$\Delta \alpha_k(n+1) = (t_{k-}O_k)O_k(1 - O_k)(\sum w_{jk}O_j + \theta_k) \tag{7}$$

$$\frac{\partial E}{\partial \alpha_j} = \left[-\sum_k \alpha_k w_{jk}(t_{k-}O_k)O_k(1 - O_k) \right] O_j(1 - O_j) \left[\left[\sum_j w_{ij}O_i \right] + \theta_j \right] \tag{8}$$

Therefore, the momentum update expression for the hidden units is:

$$\Delta \alpha_j(n+1) = \left[-\sum_k \alpha_k w_{jk}(t_{k-}O_k)O_k(1 - O_k) \right] O_j(1 - O_j) \left[\left[\sum_j w_{ij}O_i \right] + \theta_j \right] \tag{9}$$

Weights and biases are calculated in the same way, the weight update expression for the links connecting to the output nodes with a bias is;

$$\Delta w_{jk} = (t_{k-}O_k)O_k(1 - O_k)\alpha_k O_j \tag{10}$$

Similarly, bias update expression for the output nodes will be;

$$\Delta \theta_k = (t_{k-}O_k)O_k(1 - O_k)\alpha_k \tag{11}$$

The weight update expression for the input node links would be:

$$\Delta w_{ij} = \left[\sum_{k} \alpha_k w_{jk} (t_{k-} O_k) O_k (1 - O_k) \right] \alpha_j O_j (1 - O_j) O_i \tag{12}$$

And, finally the bias update expression for hidden nodes will be like this;

$$\Delta \theta_j = \left[\sum_{k} \alpha_k w_{jk} (t_{k-} O_k) O_k (1 - O_k) \right] \alpha_j O_j (1 - O_j) \tag{13}$$

4 Results and Discussions

Basically, the main focus of this research is to improve the Accuracy of the network convergence. Before discussing the simulation test results there are certain things that need be explained such as tools and technologies, network topologies, testing methodology and the classification problems used for the entire experimentation. The discussion is as follows:

4.1 Preliminary Study

The Workstation used for carrying out experimentation comes equipped with a 2.33GHz Core-2 Duo processor, 1-GB of RAM while the operating system used is Microsoft XP (Service Pack 3). The improved version of the proposed algorithms by Nazri [17] is used to carry-out simulations on MATLAB 7.10.0 software. For performing simulations, three classification problems like Breast Cancer (Wisconsin) [18], IRIS [19], and Australian Credit Card Approval [20] are selected. The following three algorithms are analyzed and simulated on the problems:

1) Gradient Descent with Simple Momentum (GDM),
2) Gradient Descent Method with Adaptive Gain (GDM-AG), [18] and
3) Gradient Descent with Adaptive Momentum(GDAM)

Three layer back-propagation neural networks is used for testing of the models, the hidden layer is kept fixed to 5-nodes while output and input layers nodes vary according to the data set given. Global Learning rate of 0.4 is selected for the entire tests and gain is kept fixed. While log-sigmoid activation function is used as the transfer function from input layer to hidden layer and from hidden layer to the output layer. In this research, the momentum term is varied adaptively between the range of [0,1] randomly . For each problem, trial is limited to 3000 epochs. A total of 15 trials are run for each momentum value. The network results are stored in the result file for each trial. Mean, standard deviation (SD) and the number of failures are recorded for each independent trial on Breast Cancer (Wisconsin) [18], IRIS [19], and Australian Credit Card Approval [20].

4.2 Breast Cancer (Wisconsin) Classification Problem

Breast Cancer Dataset was taken from UCI Machine Learning Repository databases. The dataset was created on the information gathered by Dr. William H. Wolberg [18] during the Microscopic study of breast tissue samples selected for the diagnosis of breast cancer. This problem deals with the classification of breast cancer as benign or malignant. The selected feed forward neural network architecture used for this classification problem is 9- 5-2. The target error is set to 0.01. The best momentum values for GDM and GDM-AG is 0.6 and 0.7 respectively. While for GDAM, the best momentum value is found in the interval $[0.3-0.8]$.

Table 1. Algorithm Performance for Breast Cancer Problem [19]

	Breast Cancer Problem, Target Error = 0.01		
	GDM	GDM-AG	GDAM
Accuracy (%)	92.84	94.0	94.71
SD	10.75	6.98	0.19
Failures	0	0	0

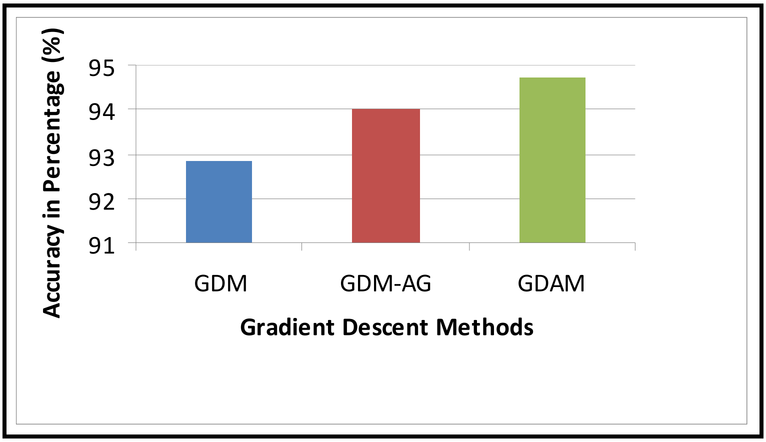


Fig. 1. Performance comparison of GDM, GDM-AG and GDAM for Breast Cancer Classification Problem

Table 1 show that the proposed algorithm (GDAM) shows far better performance in reaching the desired target error value of 0.01 than the previous improvements used for comparison. The proposed algorithm (GDAM) gives 94.71 percent accuracy when the network converges while GDM and GDM-AG is a left behind with 92.84 and 94.0 percentile accuracies respectively. Figure 1 clearly shows that GDAM outperforms GDM with an accuracy improvement ratio of 1.02.

4.3 IRIS Classification Problem

IRIS flower data set classification problem is one of the novel multivariate dataset created by Sir Ronald Aylmer Fisher [19] in 1936. IRIS dataset consists of 150 samples from Iris setosa, Iris virginica and Iris versicolor. Length and width of sepal and petals is measured from each sample of three selected species of Iris flower. The feed forward network is set to 4-5-2. The target error is set to 0.01. For Iris, the best momentum value for GDM and GDM-AG is 0.2. While for GDAM, the best momentum value is found in the interval $[0.6-0.8]$.

Table 2. Algorithm Performance for IRIS problem [20]

	IRIS Problem, Target Error = 0.01		
	GDM	GDM-AG	GDAM
Accuracy (%)	93.85	90.63	94.09
SD	0.23	3.46	1.09
Failures	1	3	0

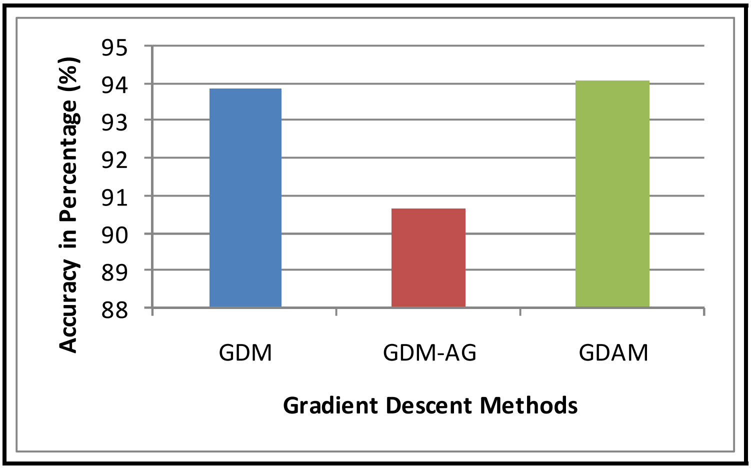


Fig. 2. Performance comparison of GDM, GDM-AG and GDAM for IRIS Classification Problem

From the Table 2, it is easily seen that the proposed algorithm (GDAM) is superior in performing convergence with a percentile accuracy of 94.09 which is better than the accuracies shown by GDM and GDM-AG. Also, it can be noted that GDM and GDM-AG have a failure rate of 1 and 3 respectively. In the Figure 2, GDAM can be seen outperforming GDM with an accuracy ratio of 1.0.

4.4 Australian Credit Card Approval Classification Problem

Australian Credit Approval dataset is taken from the UCI Machine learning repository. With a mix of 690 samples, 51 inputs and 2 outputs, this dataset contains all the details about credit card applications. Each sample in this data set represents a real credit card application and the output describes whether the bank (or similar institution) will grant the credit card or not. It was first submitted by Quinlan in 1987 [20]. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. The feed forward network architecture for this classification problem is set to 51-5-2. The target error is again set to 0.01. For Australian Credit Card Approval, GDM and GDM-AG both have the same best momentum value of 0.4. GDAM works best on the momentum value interval of $[0.7-0.8]$.

Table 3. Algorithm Performance for Australian Credit Card Approval [21]

	Echocardiogram Problem, Target Error = 0.01		
	GDM	GDM-AG	GDAM
Accuracy (%)	94.28	91.05	96.60
SD	1.15	11.87	0.53
Failures	1	1	0

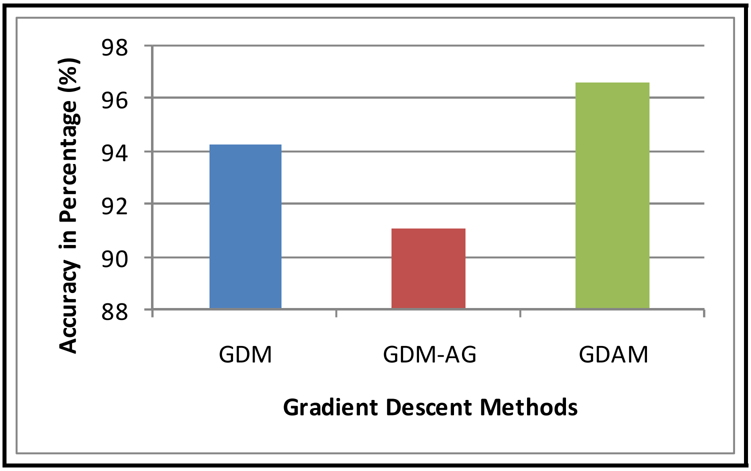


Fig. 3. Performance comparison of GDM, GDM-AG and GDAM for Australian Credit Card Approval Classification Problem

It is apparent from the Table 3, that GDAM is giving a percentile accuracy of 96.60 during convergence. Here also GDM and GDM-AG show 1 failed trial while there is no failure with GDAM even once on this classification problem. From the Figure 3, GDAM clearly shows an accuracy ratio of 1.02 on GDM.

5 Conclusions

The Back-propagation Neural Network (BPNN) is one of the most capable supervised learning algorithms deployed successfully in all engineering fields. Regardless of its high success rate at providing many practical solutions, it has a problem of slow convergence and network stagnancy which still needs to be answered. This paper proposed a further improvement on the current working algorithm by Nazri [17]. The proposed 'Gradient Descent Method with Adaptive Momentum (GDAM)' works by adaptively changing the momentum and keeping the 'gain' parameter fixed for all nodes in the neural network. The performance of the proposed GDAM is compared with 'Gradient Descent Method with Adaptive Gain (GDM-AG)' and 'Gradient Descent with Simple Momentum (GDM)'. The performance of GDAM is verified by means of simulation on the three classification problems like Breast Cancer (Wisconsin), IRIS, and Australian Credit-Card Approval respectively. The final results show that GDAM is far better than previous methods with an accuracy ratio of 1.0 for all classification problems.

Acknowledgments. The Authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for supporting this research under the Post Graduate Research Incentive Grant Vote No. 0737.

References

1. Kosko, B.: Neural Network and Fuzzy Systems, 1st edn. Prentice Hall of India, Englewood Cliffs (1994)
2. Krasnopolsky, V.M., Chevallier, F.: Some Neural Network application in environ-mental sciences. Part II: Advancing Computational Efficiency of environmental numerical models. Neural Networks 16(3-4), 335–348 (2003)
3. Coppin, B.: Artificial Intelligence Illuminated, USA. Jones and Bartlet Illuminated Series, ch.11, pp. 291–324 (2004)
4. Basheer, I.A., Hajmeer, M.: Artificial neural networks: fundamentals, computing, design, and application. J. of Microbiological Methods 43(1), 3–31 (2000)
5. Zheng, H., Meng, W., Gong, B.: Neural Network and its Application on Machine fault Diagnosis. In: ICSYSE 1992, September 17-19, pp. 576–579 (1992)
6. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Internal Representations by error Propagation. J. Parallel Distributed Processing: Explorations in the Microstructure of Cognition (1986)
7. Lee, K., Booth, D., Alam, P.A.: Comparison of Supervised and Unsupervised Neural Networks in Predicting Bankruptcy of Korean Firms. J. Expert Systems with Applications 16, 1–16 (2005)
8. Zweiri, Y.H., Seneviratne, L.D., Althoefer, K.: Stability Analysis of a Three-term Backpropagation Algorithm. J. Neural Networks 18, 1341–1347 (2005)
9. Fkirin, M.A., Badwai, S.M., Mohamed, S.A.: Change Detection Using Neural Network in Toshka Area. In: NSRC, 2009, Cairo, Egypt, pp. 1–10 (2009)
10. Sun, Y.J., Zhang, S., Miao, C.X., Li, J.M.: Improved BP Neural Network for Trans-former Fault Diagnosis. J. China University of Mining Technology. 17, 138–142 (2007)

11. Hamreeza, N., Nawi, N.M., Ghazali, R.: The effect of Adaptive Gain and adaptive Momentum in improving Training Time of Gradient Descent Back Propagation Algorithm on Classification problems. In: 2nd International Conference on Science Engineering and Technology, pp. 178–184 (2011)
12. Shao, H., Zheng, H.: A New BP Algorithm with Adaptive Momentum for FNNs Training. In: GCIS 2009, Xiamen, China, pp. 16–20 (2009)
13. Rehman, M.Z., Nawi, N.M., Ghazali, M.I.: Noise-Induced Hearing Loss (NIHL) Prediction in Humans Using a Modified Back Propagation Neural Network. In: 2nd International Conference on Science Engineering and Technology, pp. 185–189 (2011)
14. Swanston, D.J., Bishop, J.M., Mitchell, R.J.: Simple adaptive momentum: New algorithm for training multilayer Perceptrons. J. Electronic Letters 30, 1498–1500 (1994)
15. Mitchell, R.J.: On Simple Adaptive Momentum. In: CIS 2008, London, United Kingdom, pp. 1–6 (2008)
16. Nawi, N.M., Ransing, M.R., Ransing, R.S.: An improved Conjugate Gradient based learning algorithm for back propagation neural networks. J. Computational Intelligence. 4, 46–55 (2007)
17. Nawi, N. M.: Computational Issues in Process Optimization using historical data: PhD Eng. Thesis.Swansea University, United Kingdom (2007)
18. Wolberg, W.H., Mangasarian, O.L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. National Academy of Sciences 87, 9193–9196 (1990)
19. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annual Eugenics 7, 179–188 (1936)
20. Quinlan, J.R.: Simplifying Decision Trees. J. Man-Machine Studies 27, 221–234 (1987)

Fuzzy Models for Complex Social Systems Using Distributed Agencies in Poverty Studies

Bogart Yail Márquez[1], Manuel Castanon-Puga[1], Juan R. Castro, E. Dante Suarez[2], and Sergio Magdaleno-Palencia[1]

[1] Baja California Autonomous University, Chemistry and Engineering Faculty, Calzada Universidad 14418, Tijuana, Baja California, México. 22390
bogart@uabc.mx, puga@uabc.edu.mx, jrcastror@uabc.edu.mx,
http://fcqi.tij.uabc.mx

[2] Trinity University, Department of Business Administration,
One Trinity Place, San Antonio, TX, USA. 78212
esuarez@trinity.edu
http://www.trinity.edu

Abstract. There are several ways to model a complex social system, as is the poverty of an entity, the object of this paper is to present a methodology consisting of several techniques that offers to solve complex social problems with soft computing.

Keywords: Complex Social Systems, Data Mining, Neuro-Fuzzy, Distributed Agencies, Poverty.

1 Introduction

Philosophy in science refers to aspects such as how scientific research should be conducted. It tries to explore the relation between the ontology, epistemology, and methodology. The methodology refers to the forms in which reality and knowledge can be studied; it does not question knowledge that has been accepted as true by the scientific community but instead concentrates on strategies to expand the knowledge.

In the beginning, poverty was defined as a static and unidimensional concept, studied primarily in a static fashion with a clear economic nuance, attending to what is considered acceptable lower income levels in society. But in recent decades, it has passed to a concept of dynamic and multidimensional nature. Recent research shows that not only the income levels determine poverty but also the absence of resources and opportunities such as education, housing and other factors. Thus we see the existence of multidimensional poverty studies with various aspects and multidimensional analysis[1].

As in other developing countries, the primary cause of the existing inequality in Mexico is poverty, which is reflected in the results of economic development.

The purpose of this paper is not to discuss the various determinants factors for the conceptualization of poverty, but instead propose a methodology with different

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 391–400, 2011.

computational techniques that aid complex social simulations and provide an option for the analysis of social problems.

Social systems contain many components which depend on many relationships; this makes it difficult to construct models closer to reality [2]. To analyze these systems with a dynamic and multidimensional perspective, we will consider data mining theory, fuzzy logic and distributed agencies[3].

There are several problems in analyzing a complex system, like the interactions between individuals who, because of these relationships, may be forced to change their environments. A new mathematical theory to study this evolution is complex systems, an emerging field in the dynamics of adaptation. This system could explain the long-term effects of interactions between the evolutionary processes[4]. A complex system is a conceptualization of the complexity and the related systems. Complex Adaptive Systems (CAS) established by physicists and economists since two decades ago[5], classification systems, ecosystems and simulation systems, helped lead the creation of evolutionary computation and artificial life. The framework is derived from many natural and artificial complex systems that have collaborated in such diverse fields such as psychology, anthropology, genetics evolution, ecology, and business management theory. The methodologies using simulation CAS on models in the field of biological computing have contributed in this field; CAS is the result of the artificial systems[6].

In this work a fuzzy system model is used as a poverty model as CAS. The model is composed by the combination from two theories: fuzzy logic originally introduced by Zaded [7] to model vagueness and uncertainty in the real world, and ditributed agencies [3, 8].

For clarity, before going into details about the methodology, some important distinctions between distributed agencies in relation to agents, multi-agent systems, and holons need to be made.

1.1 Agents

There is no widely accepted definition for the term agent by the community that works with Multi-Agent Systems (MAS). Wooldridge and Jennings present two definitions in [Wooldridge & Jennings, 1995] for what is considered an agent. According to their definition, an agent is any computer system that presents the following characteristics: autonomy, social capacity, reactivity, and pro-activity. This definition does not contemplate important characteristics such as the ability to acquire and use knowledge, cooperate with other agents, or the existence of beliefs, intentions and obligations. For Genesereth and Ketchpel [citation needed], an agent is a software component that communicates with other agents by exchanging messages using an Agent Communication Language (ACL). The main focus of this definition lies in the consideration of an ACL. Petrie accepts the definition given by Genesereth and Ketchpel and proposes the explicit addition of the motive for message passing with an ACL is for the purpose of performing tasks.

1.2 Multi-Agents Systems

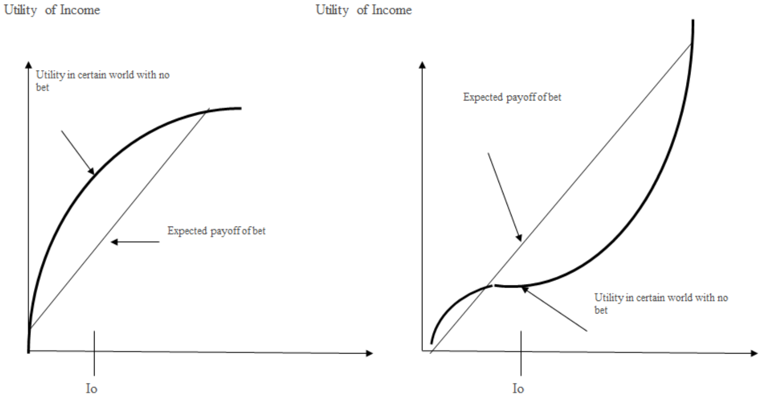
Multi-agent systems consist of autonomous agents that work together to solve problems, characterized by the fact that no one agent has all the information or all the capabilities needed to solve the problem, there is no global control system, information is decentralized, and agent interactions are asynchronous; agents dynamically decide what tasks to perform. On the other hand, MAS setups can be understood as complex entities where a multitude of agents interact within a structured environment designed for a global purpose.

1.3 Holons

Holons refer to complex social communities that cannot be directly controlled and have a fractal like autonomous structure. Speaking in system's terms, this means a community is both complex and autonomous, they normally posses a dynamic behavior in limited stable conditions and operate thru self organization.

1.4 Distributed Agencies

This new approach treats agents as something that can be agent-like, unlike a traditional approach where something either is or is not an agent. In distributed agencies (DA), it is possible to consider degrees of agency, and an agent can be anything from the most traditional view of agents as a representation of humans to societies, countries, religions, political parties, coalitions, species, emotions, or families. Fuzzy agents are defined in many levels and in potentially infinite dimensions. The scope and multiple levels of social reality are saturated with uncertainty, a characteristic that we propose be handled in distributed agencies by fuzzy logic and this degree of uncertainty is what creates "emergence".



A. For normal amounts the preference is for a certain amount of income over the expected result of the bet.

B. For the possibility of an extraordinary amount of income, the individual prefers the expected payoff of the lottery bet over the certain income of not buying such a ticket.

Fig. 1. Utility of income.(Alternative Perspective DA)[8]

2 Methodology

Real world simulations such as poverty must include some means of validation [9]. In econometrics, studies performed on populations and economy have abundant data, the main problem is finding data sets that are adapted to the desired architecture [10].

With the steady increase in availability of information, from existing projects such as databases needed for social simulation, it becomes essential to use data mining. In our case, for the vast amount of data, we obtain the necessary quantitative information on the most important subject of the social and economic system from government databases. Data mining extract implicit information such as social patterns in order to discover knowledge[11]. The use of these techniques has been wide spread in this field in recent years, most research efforts are dedicated to developing effective and efficient algorithms that can extract knowledge from data[3, 12].

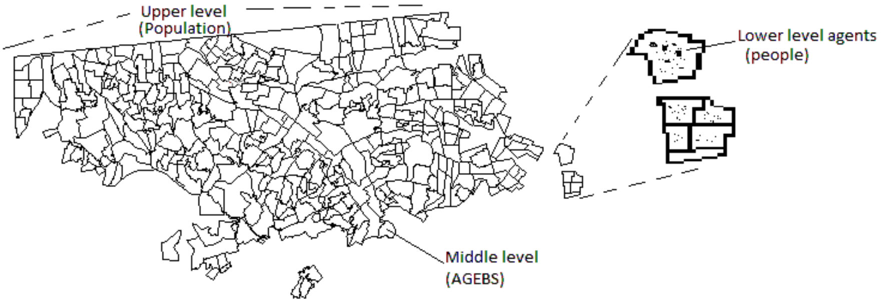


Fig. 2. Levels agents represented on the social system

For the particular case of the City of Tijuana, in the state of Baja California Mexico we use the geographical, demographic and economic indicators provided by the databases of the National Institute of Statistics and Geography (INEGI).

The information for this city is fragmented in 363 “AGEBs”. “The AGEB delimitates urban areas, a whole locality of 2,500 inhabitants or more, or a municipal seat, regardless of the number of people, in groups that typically range from 25 to 50 blocks. Rural AGEB’s frame an area whose land use is predominantly agricultural and these are distributed communities of less than 2,500 inhabitants that for operational purposes, are denominated rural locations[13]. For each AGEB we determine the extent of poverty, taking into account 10 variables of income and employment rate, 23 variables on education, and 15 variables on the resources available in a home, such as TV, telephone, refrigerator and more. This gives 48 by 363 matrix containing the information necessary to analyze.

The databases for the model were compiled in a geographic information system that helped in the creation, classification and format of the data layers required. This saves us the issue of having different thematic maps of information;each map, quantifies the spatial structure to display and interpret the different areas and spatial patterns of Tijuana.

3 Implementations

Using NetLogo platform; we are able to simulate social phenomena, model complex systems and give instructions to hundreds or thousands of independent agents all operating holistically [14]. It also allows us to filter information from geographical information system with spatial and statistical data. This makes it possible to explore the relationship and behavior of agents and the patterns that emerge from the interactions within a geographical space.

Each AGEB contains quantitative information about employment, education, income, articles and infrastructure that a home has.

Using neuro-fuzzy system to automatically generating the necessary rules, this phase of data mining with a fuzzy system becomes complicated as there is no clear way to determine which variables should be taken into account[15]. The Nelder-Mead (NM) search method is used; even though being more efficient on other methods such as genetic algorithms, as shown in other studies[16],since the NM method seems to produce more accurate models with fewer rules. This optimization algorithm is widely used and is a numerical method that minimizes an objective function in a multidimensional space, finding the approximate local optimal solution to a problem with N variables [17].

Using this grouping algorithm we obtain the rules, which are assigned to each agent which represents an AGEB. The agent receives inputs from its geographical environment and also must choose an action in an autonomous and flexible way to fulfill its function [18].

Considering the use of agent-based models, is the fact that they are very similar to artificial societies, following the same techniques. Their main differences focus on system simulations, and the research program designs[9]. Thus considering all the properties of the agents [19]is suitable for the purpose of our research. Each agent has a distinct function depending on the rule that it is assigned and the geographic environment where it will decide the tasks dynamically, so there is no global control system.

The purpose is to provide agents with the least possible rules and observe the operation of the system by the interactions, the system itself generates intelligent behavior was not necessarily planned or defined within the agents themselves; thus achieving an emergent behavior.

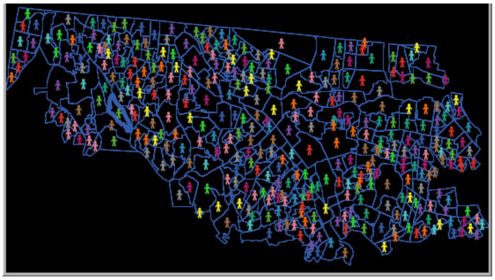


Fig. 3. Tijuana divided into AGEBS and Assignment of each AGEB to an agent, using NetLogo

The rules obtained from the clustering algorithm can tell us which agents have more income; which are at a higher or lower educational level, and what resources are available.

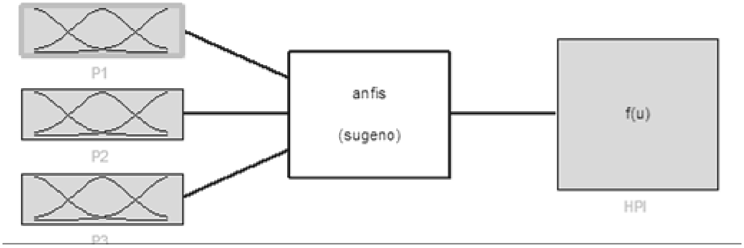


Fig. 4. Network architecture[16]

Taking the compound poverty index which measures poverty as reference on the three basic dimensions (health, education and an acceptable quality of life) we calculate this factor with:

$$HPI = \left[\frac{1}{3}(P_1^{\alpha} + P_2^{\alpha} + P_3^{\alpha})\right]^{\frac{1}{\alpha}} \quad (1)$$

P1: Population that has no access to medical services
P2: Adult illiteracy rate
P3: Population with no access to basic services such as running water, pluming or electricity.

These variables correspond to the 363 AGEBs that the municipality of Tijuana contains. We establish the objective functions of all levels of agency that are considered, as well as the interactions that are prevalent in the corresponding networks, as shown in figure 5.

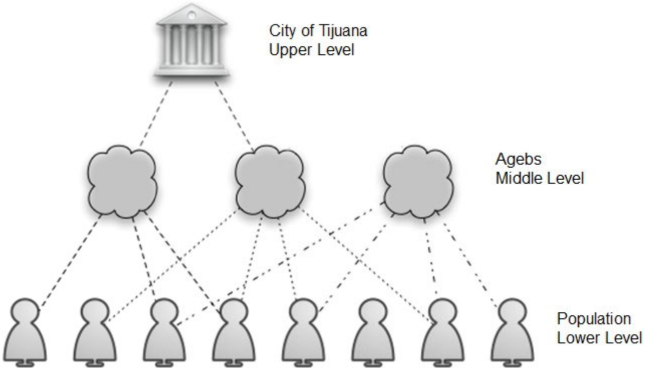


Fig. 5. Representation of multiple levels

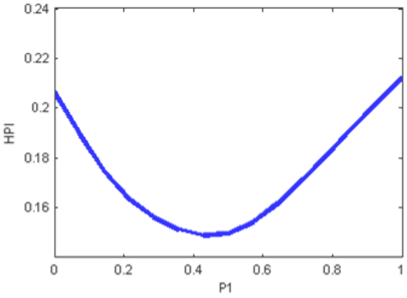


Fig. 6. Human Poverty Index generated by model using population that has no access to medical services

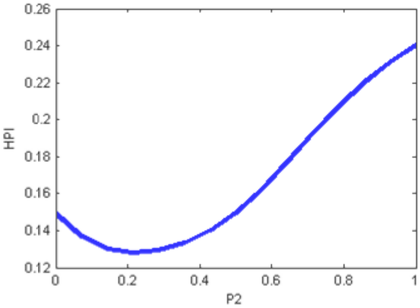


Fig. 7. Human Poverty Index generated by model using Adult illiteracy rate

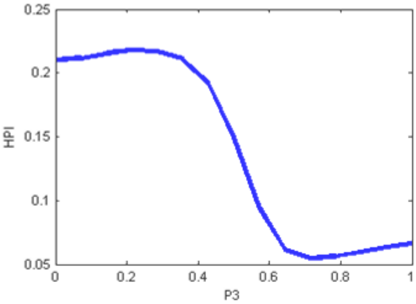


Fig. 8. Human Poverty Index generated by model using Population with no access to basic services such as running water, pluming or electricity

Agent-based technology has been considered appropriate for the social development of distributed systems [20]. Distributed agents are a promising strategy that can correct an undesirable centralized architecture[21]. Distributed Agents do not define independent agents. The idea behind the distributed agency modeling language stems from a worldview that is ubiquitous in appearance, in which we find groups that are irreducible to their parts, and exist in different dimensions where different rules apply [22].

Therefore, the next step is for the agencies to try and solve problems distributed among a group of agents, finding the solution as the result of cooperative interaction. Communication facilitates the processes of cooperation; the degree of cooperation between agents can range from complete cooperation to a hostile [23].

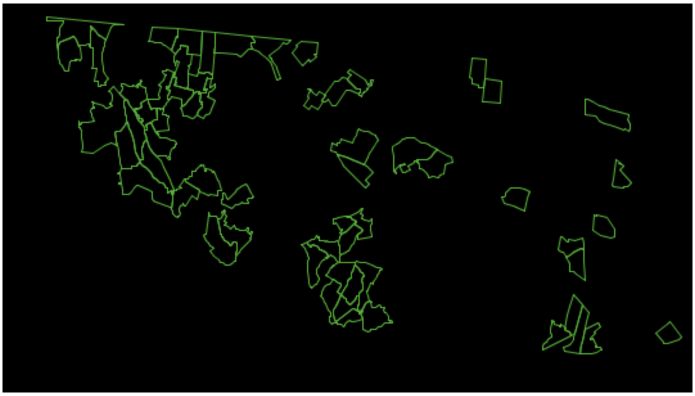


Fig. 9. Examples of areas where : Population that has no access to medical services

An example of cooperation might be agents of a given area sharing certain resources for the mutual good. A hostile situation may occur by agents blocking the objectives of others fear of sharing resources and by a lack of resources in the area. For cooperation and coordination mechanisms to succeed in a system of agents there must be an additional mechanism: negotiation, by which, members of a system can reach an agreement when each agent defends its own interests, leading to a situation that benefits all, taking into account all points of view [23].

4 Conclusions

Our intention is to create a system that is composed of agents where each agent represents an AGEB 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 poverty model using different techniques can be a powerful tool for any planning process. The use of distributed agencies allows us to unearth new theoretical models and furthers our understanding the relationships found within distinct levels of reality; and helped us to see very concrete cases. And fuzzy logic system integrated into the poverty model represents a methodological advance for distributed agencies for modeling the uncertain.

The field is inherently interdisciplinary, linked to the science of complexity, systems theory, data mining, neuro fuzzy and distributed agencies. The results of such simulations provide a powerful alternative shown to supplement, replace and expand the traditional approaches in the field of social sciences.

References

1. Akindola, R.B.: Towards a Definition of Poverty Poor People's Perspectives and Implications for Poverty Reduction. Journal of Developing Societies 25, 121–150 (2009)
2. Ding, Y.-C., Zhang, Y.-K.: The Simulation of Urban Growth Applying Sleuth Ca Model to the Yilan Delta in Taiwan. Journal Alam Bina, Jilid 09(01) (2007)
3. 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)
4. Sigmund, K.: Complex Adaptive Systems and the Evolution of Reciprocation, vol. 16 (1998)
5. Brownlee, J.: Complex Adaptive Systems Technical Report 070302A (2007)
6. Miler, J., Page, S.: Complex Adptative Systems. An introduction to computational models of social life, 284 (2007)
7. Zaded, L.A.: Fuzzy sets. Information and Control 8 (1965)
8. Suarez, E.D., Rodríguez-Díaz, A., Castañón-Puga, M.: Fuzzy Agents, in Soft Computing for Hybrid Intelligent Systems. In: Castillo, O., et al. (eds.), pp. 269–293. Springer, Berlin (2007)
9. Drennan, M.: The Human Science of Simulation: a Robust Hermeneutics for Artificial Societies. Journal of Artificial Societies and Social Simulation 8(1) (2005)
10. Marquez, B.Y., et al.: On the Modeling of a Sustainable System for Urban Development Simulation Using Data Mining and Distributed Agencies. In: 2nd International Conference on Software Engineering and Data Mining 2010. IEEE, Chengdu (2010)
11. Dubey, P., Chen, Z., Shi, Y.: Using Branch-Grafted R-trees for Spatial Data Mining. In: International Conference on Computational Science (ICCS), pp. 657–660. Springer, Heidelberg (2004)
12. Peng, Y., et al.: A Descriptive Framework for the Field Of Data Mining and Knowledge Discovery. International Journal of Information Technology & Decision Making 7, 639–682 (2008)
13. INEGI, II Conteo de Población y Vivienda 2005. Instituto Nacional de Estadística Geografía e Informática (2006)
14. Wilensky, U.: NetLogo Software (1999), `http://ccl.northwestern.edu/netlogo`
15. Castro, J.R., Castillo, O., Martínez, L.G.: Interval Type-2 Fuzzy Logic Toolbox. Engineering Letters 15 (2007)
16. Rantala, J., Koivisto, H.: Optimised Subtractive Clustering for Neuro-Fuzzy Models. In: 3rd WSEAS International Conference on Fuzzy Sets and Fuzzy Systems, Interlaken, Switzerland (2002)
17. Stefanescu, S.: Applying Nelder Mead's Optimization Algorithm for Multiple Global Minima. Romanian Journal of Economic Forecasting, 97–103 (2007)
18. Gilbert, N.: Agent-Based Models. Sage Publications Inc., Los Angeles (2007)
19. Wooldridge, M., Jennings, N.: Intelligent Agents: Theory and Practice. Knowledge Engineering Review (1995)
20. Botti, V., Julián, V.: Estudio de métodos de desarrollo de sistemas multiagente. In: Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial (2003)

21. Russell, S., Norvig, P.: Inteligencia Artificial. Un Enfoque Moderno. 2o ed, ed. Pearson Prentice Hall, M. Pearson Educacion. S.A (2004)
22. Suarez, E.D., Rodríguez-Díaz, A., Castañón-Puga, M.: Fuzzy Agents. In: Castillo, O., et al. (eds.) Soft Computing for Hybrid Intelligent, Springer, Heidelberg (2007)
23. Gilbert, N.: Computational social science: Agent-based social simulation. In: Phan, D., Amblard, F. (eds.) Agent-Based Modelling and Simulation, pp. 115–134. Bardwell, Oxford (2007)

Grid Jobs Scheduling Improvement Using Priority Rules and Backfilling

Zafril Rizal M. Azmi[1], Kamalrulnizam Abu Bakar[2], Abdul Hanan Abdullah[2], Mohd Shahir Shamsir[3], Rahiwan Nazar Romli[1], and Syahrizal Azmir MD Sharif[1]

[1] Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang
[2] Faculty of Computer Science and Information System, Universiti Teknologi Malaysia
[3] Faculty of Biosciences and Bioengineering, Universiti Teknologi Malaysia
zafril@ump.edu.my, kamarul@fsksm.utm.my,
hanan@fsksm.utm.my, shahir@fbb.utm.my, rahiwan@ump.edu.my,
syazmir@ump.edu.my

Abstract. Over the past decade, scheduling in grid computing system has been an active research. However, it is still difficult to find an optimal scheduling algorithm to achieve load balancing. Most of the researchers have focus on schedule-based algorithms such as genetic algorithm and particle swarm optimization to solve this problem and use priority rules algorithms as initial schedule in those algorithms. The main reason this paper was produced is that most of these researchers failed to justify why they use a specific priority rules scheduler as initial schedule in their work. This paper addresses this issue by presenting a comparison results on several priority rules algorithms based on several performance metrics. To add novelty to this paper, we have proposed several schedule-based algorithms that basically based on the combination of backfilling technique and priority rules algorithms. Our results show the significant improvements compared to the original priority rules algorithms.

Keywords: Grid scheduling, priority rules, backfilling, performance.

1 Introduction

In the beginning, computing problems basically were solved using single computers. Later on, the multiprocessor systems has been developed which lead problems to be divided and solved in parallel. The growth of multi processor and parallel computing is one of the reasons why the distributed system is needed to stand as one. Distributed computing is a method of computer processing in which different parts of a program run simultaneously on two or more computers that are communicating with each other over a network. Distributed computing is a type of segmented or parallel computing, but the latter is the most commonly used term to refer to processing in which different parts of a program run simultaneously on two or more processors that are part of the same computer. While both types of processing require that a program be segmented, distributed computing also requires that the division of the program takes into account the different environments on which the different sections of the program will be

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 401–415, 2011.

running. For example, two computers are likely to have different file systems and different hardware components.

There are many different types of distributed computing systems and many challenges to overcome in successfully designing one. The main goal of a distributed computing system is to connect users and resources in a transparent, open, and scalable way. Ideally this arrangement is drastically more fault tolerant and more powerful than many combinations of stand-alone computer systems. Examples of technique that uses distributed system are computer cluster and grid computing. A cluster consists of multiple stand-alone machines acting in parallel across a local high speed network while grid uses the resources of many separate computers, connected by a network (usually the Internet), to solve large-scale computation problems. This paper will focus on the second technique or architecture which is grid computing.

Grid computing can have variety of heterogeneous resources connected with each other. However, the available resources such as the network and computational power are continually changing with respect to every available node. The constantly changing characteristic of the heterogeneous resources is known as dynamic resources. To make it more critical, the jobs that need to be process by these resources are arriving from different length of time and not knowing by the system until the particular jobs arrive to the system.

In order to utilize these dynamic resources and jobs optimally, a scheduling strategy should be able to continually adapt to the changes and properly distribute the workload and data amounts scheduled to each node. Unfortunately, several researchers have stated that serious difficulty in concurrent programming of a grid computing has occurred in terms of dealing with scheduling and load balancing of such a system, which may consist of heterogeneous computers [2],[5]. The current popular techniques in grid scheduling are schedule-based algorithms that adapting optimization technique such as Genetic Algorithm, Ant Colony Optimization and Tabu Search. The corresponding authors of these techniques have proof that their techniques are much better compared to the priority rules algorithms. Priority rules scheduling has been widely used in the popular scheduling production systems such as Condor and PBS. Although schedule-based algorithms were proven to be more effective (based on objective the researchers want to achieve), priority rules techniques are still very important because most of the schedule-based algorithms will have to apply priority rules algorithms for initial schedule in their schedule-based algorithms.

Most of the researchers failed to justify why they are using a specific priority rules technique instead of other existing priority rules technique as initial schedule. In order to justify which priority rules algorithms are better than the other, this work will compare some of the popular priority rules algorithms. It is meant to be a guideline for the selection of initial algorithms for the schedule-based algorithms based on which priority of metrics the researchers want to achieve.

Furthermore, we have proposed schedule-based algorithms which are basically designed by the combination a backfilling technique and priority rules algorithms. The proposed algorithms were design in order to improve the existing priority rules scheduler in term of performance based on five performance metrics.

The rest of the paper is structured as follows. The next section will explain the problem description. Section 3 define major objective function in grid scheduling which also the performance metrics to compare all the algorithms. Section 4 will generally describe exactly six priority rules algorithms that will be covered throughout the paper, namely First Come First Serve (FCFS), Shortest Job First (SJF), Longest Job First (LJF), Earliest Release Date (ERD), Earliest Deadline First (EDF) and Minimum Time To Due Date (MTTD). In Section 5 the EG-EDF algorithm and backfilling which are the basic theories for our proposed schedule-based algorithms will be presented. Section 6 explained the proposed algorithm. The experimental setup will be explained in Section 7 while Section 8 will present the results and discussion. Finally Section 9 draws the paper's conclusion.

2 Problem Description

The grid scheduling problem consists of scheduling j jobs with given processing time on m resources. In this study, we assume that the system have fixed sized of m machines. However the number of jobs j will be changing dynamically based on the arrival time a_j of each jobs. Since the grids environment is dynamic, the characteristics of jobs are not known in advance and they will keep entering the system although others are still running. A set of jobs consists of several jobs with different characteristics, $j = \{j_1, j_2, j_3,\ldots j_i\}$ Each job j_i is an independent job. A set of grid resources consists of a set of heterogeneous machines with various specifications, $m = \{m_1, m_2, m_3,\ldots,m_i\}$ and each machines can have more than 1 processing unit, PU.

Generally, this mapping is done by utilizing the resources effectively. The proposed scheduler tries to improve the performance of five metrics explained in the following subsection.

3 Performance Metrics

This section will explain the objectives in scheduling which also the metrics parameters evaluated in almost all research related to grid scheduling. However, we have focused on 5 objectives which are minimizing total tardiness, minimizing makespan time, minimizing response time, minimizing number of delayed jobs and maximizing resource utilization.

3.1 Minimizing Total Tardiness

One of the main objectives of the scheduling procedure is the completion of all jobs before their agreed due dates. Failure to keep that promise has negative effects on the credibility of the service provider. If we define lateness of job i as the difference between its completion time C_i and the corresponding due date d_i, then the tardiness of the job is calculated from the following expression:

$$\mathrm{T_i}=max(0, C_i- d_i) \qquad (1)$$

In other words, tardiness represents the positive lateness of a job. In a single machine environment, we define the total tardiness problem as follows:

A number of jobs $J_1, J_2 ...., J_n$ are to be processed in a single facility. Each job is available for processing at time zero, and is completely identified by its processing time p, and its due date d_i. We seek to find the processing sequence that minimizes the sum of tardiness of all jobs:

$$\sum_{i=1}^{n} max(0, Ci - di) \tag{2}$$

Where C_i is the completion time of job i. The total tardiness problem is a special case of the weighted total tardiness problem. Both problems are not easy to solve, especially for large values of n.

3.2 Minimizing Makespan Time

Makespan is a standard performance metric to evaluate scheduling algorithms. Small values of makespan mean that the scheduler is providing good and efficient planning of jobs to resources. The makespan of a schedule can be defined as the time it takes from the instant the first task begins execution to the instant at which the last task completes execution [18]. In a simplest words, makespan is the time when finishes the latest job.

3.3 Minimizing Response Time

Response time also known as flow time. Response time is the sum of finalization times of all the tasks [18]. Response time and makespan are the two major objectives to be minimized in research involving scheduling. However, minimization of makespan always results in the maximization of response time [1].

3.4 Minimizing Number of Delayed Jobs

Delayed Jobs means jobs that fail to meet their deadline [10]. Deadline or due date is a period of time a job must be completed [3]. The goal typically in such problems is to complete the maximum number of jobs by their deadlines. According to Klusacek and Rudova in [10] a higher machine usage fulfills resource owner’s expectations, while a higher number of non delayed jobs guarantees a higher Quality of Service (QoS) provided for the users. By reducing the number of delayed jobs, QoS for the system that using the proposed scheduling technique also will be improve.

3.5 Maximizing Resource Utilization

Maximizing machine usage or resource utilization of the grid system is another important performance metrics. Utilization is the percentage of time that a resource is actually occupied, as compared with the total time that the resource is available for use. Low utilization means a resource is idle and wasted.

4 Priority Rules Scheduling Algorithms

Priority rules also referred as Queue-based [9]. Instead of guaranteeing optimal solution, these techniques aim at finding reasonably good solutions in a relatively short time. Although it is a suboptimal algorithms, it is yet the most frequently used for solving scheduling problem in real world because of the easiness to implement and their low time complexity. The most basic priority rules scheduling is First Come First Serve (FCFS). This section will explain six priority rules algorithms used in this work.

4.1 First Come First Serve (FCFS)

FCFS or also known as First In First Out (FIFO) is the simplest and the most basic of scheduling. It is known to be used worldwide in many fields that involve client-server interaction. In grid scheduling, FCFS policy manage the jobs based on their arriving time which mean first job will be process first without other biases or preferences. This concept has been used by several well known enterprise scheduler such as MAUI [8] and PBS [7].

4.2 Earliest Deadline First (EDF)

EDF is a policy that checks all the incoming jobs due date or deadline. Jobs will be process or put in the queue based on the time indicate by the dateline. First job to meet the deadline will be put first in the queue.

4.3 Shortest Job First (SJF)

SJF also known as Shortest Job Next (SJN) or Shortest Process Next (SPN)) is a scheduling technique that selects the job with the smallest execution time to execute next. It also means that job with the longest execution time will receive the lowest priority and will be put last in the queue.

Abraham in [1] said that the theoretical rule to minimize the response time is to schedule the shortest job on the fastest resource. Since this policy gives preference to some groups of jobs over other group of jobs, this policy is not fair compared to FCFS policy. In extreme cases, when jobs with shorter execution time continue arriving, the jobs with longer execution period may not get a chance to execute and may have to wait forever. This situation is known as starvation and could be a serious problem and shows the low degree of fairness for this policy [6]. In addition SJF also is believed to have the maximum makespan time compared to other algorithms in this paper because of its characteristics mentioned.

4.4 Longest Job First (LJF)

LJF have the contradiction behavior of SJF. While shortest job is believed to reduce the response time, processing longest job first on the fastest resource according to Abraham in [1] will minimize the makespan time. However, LJF will be suffering due to slightly increase in the response time.

4.5 Earliest Release Date (ERD)

ERD put the first priority to the job that has the least release date in the queue. Release date is the start time of each and every job and it can be different or same. If there exist two or more jobs that have the same release date, FCFS rule will be applied. Furthermore, Rasooli mentioned in [13] that if the number of jobs is few in the queue, ERD performance will be almost similar to FCFS but when the number of jobs increases, the results will defer.

4.6 Minimum Time to Due Date (MTTD)

MTTD is a scheduling algorithm which put the priority on the jobs according to the time that can be considered for the job to be executed without tardiness [13]. In the simplest words MTTD aims for producing a scheduler that has minimum tardiness. To achieve this objective, MTTD define the time as follow:

$$\textit{(Deadline-Release Date)} \tag{3}$$

5 Related Work

In this paper, new schedule based scheduling algorithms are introduced. The proposed algorithms are basically adapted from the backfilling and Earliest-gap Earliest Deadline First (EG-EDF) algorithms introduced in [10]. In this work we have produce 5 new algorithms namely Backfill First Come First Serve (BF-FCFS), Backfill-Shortest Job First (BF-SJF), Backfill Longest Job First (BF-LJF), Backfill Earliest Release Date (BF-ERD) and Backfill Minimum Time To Due Date (BF-MTTD). The next sub section describes how the gap utilization is done in the EG-EDF algorithm.

5.1 Earliest Gap – Earliest Deadline First (EG-EDF)

Klusacek and Rudova [10] introduced this method by considering a gap as a period of idle CPU time. A gap appears each and every time the number of available CPUs of a machine is greater than the number of CPUs requested by the specific jobs in the given time period. EG-EDF works by adding newly arriving job into the existing schedule.

There are five important steps in the EG-EDF algorithm which are:

- Step 1: Determines which particular machine is suitable to execute the job.
- Step 2: Places new job to the particular suitable machines schedule.
- Step 3: Check the current machines schedule whether a suitable gap for the new job exists.
- Step 4: If more than one suitable gaps, use the earliest one (EG policy).
- Step 5: Evaluate the new schedule (acceptance criterion). If the new schedule better than current schedule, accept.

5.2 Backfilling

Gap or Hole filling technique was originally developed from the backfilling algorithm introduced in EASY [12]. The purpose of backfilling is to improve system utilization of scheduler that used FCFS. Although FCFS is a simple policy and have been widely used, it suffers from the low system utilization [15]. This happens because there exist a gap between two jobs that make the resource idle [17]. Backfilling improves resource utilization by allowing small job to fill in those gaps. Many backfilling variants have been suggested, but most of them consider jobs candidate both for execution and for backfilling according to a FCFS strategy [16].

6 Proposed Algorithm

The proposed algorithms introduced in this paper which are BF-FCFS, BF-SJF, BF-LJF, BF-ERD and BF-MTTD were designed by combining backfilling like procedure with priority rules algorithm. Figure 1 shows general architecture for the proposed scheduling algorithms.

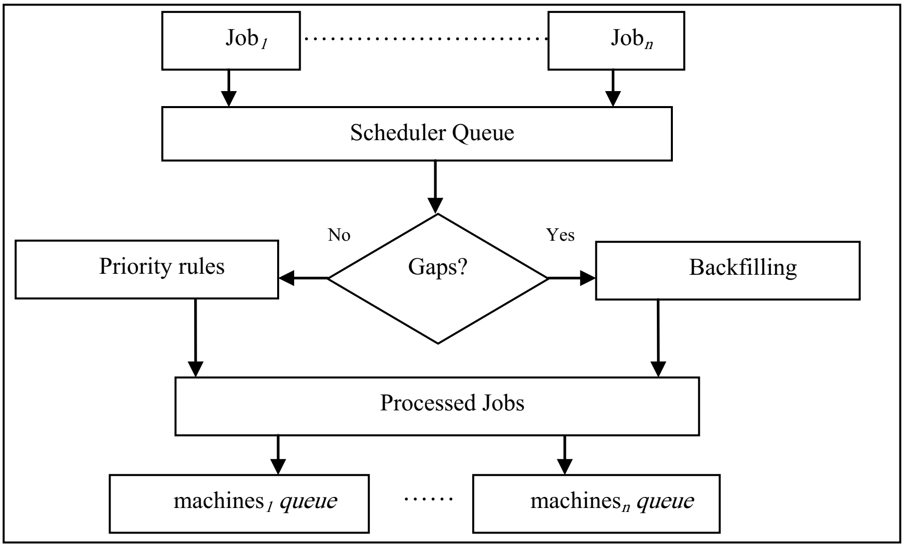


Fig. 1. General architecture of Priority rules-Backfilling

Priority rules-Backfilling scheduler is designed to manage newly-arrived jobs submitted by the grids user to the grid system. The incoming new jobs are sorted using First Come First Serve (FCFS) in the *Scheduler Queue*. This *raw* queue then is checked whether the first job in the queue can fit in into the gap found in the *machine$_n$ queue*. Since the gap only appear in the queue when more than one job exist at the same cycle time, the priority rules will always be the initial scheduler used to allocate the job to selected *machine$_n$ queue*. However, if there exist even one gap that can be fit in by a new job, the backfilling approach will be used. The backfilling algorithm considers not only small jobs but apply to all new job that entering the system.

The mechanism used to evaluate each decision is based on the makespan and prediction of job that meet the deadline objective functions [10]. To achieve this, the weight for both objectives is calculated and if the total is more then *0*, the current *job-machine* mapping is considered the best schedule so far, compared to the previous

Algorithm 1. Priority rules-Backfilling

```
resourceSize = Get total number of resource in the grids
for i=0 to resourceSize do
    if number of PU < number of PU requested by $job_m$ then
        break operation
    else $machine_i$ is suitable to perform $job_m$
    if there exist suitable gap in $machine_i$ schedule then
        insert $job_m$ into $machine_i$ schedule
    end if
    else if no suitable gap in $machine_i$ schedule then
        insert $job_m$ into $machine_i$ schedule based on selected rule
        based algorithms (FCFS/SJF/LJF/ERD/MTTD)
    end if
    total weight=weight for makespan + weight for jobs that meet
    deadline
    if total weight < 0.0 then
        remove $job_m$ from $machine_i$ schedule
    else
        put $job_m$ into $machine_i$ schedule
    end if
end for
```

match. These steps will continue until all the machines were tested with the objective functions. The formal algorithm for the proposed scheduler is given in Algorithm 1.

7 Experiment Setup

We have evaluated the performance using five parameters which are makespan time, resource utilization, total tardiness, response time and total number of delayed jobs. The experiments have been conducted on AMD Turion 2.0 GHz machine with 1024MB RAM.

The experiment were also conducted by using simulation of 150 machines with different CPU number and speed and total number of 3000 jobs with 1 second inter-arrival time. The smaller this time is, the greater the system workload is [16][10][11].

This work uses the Alea [9] the extended version of GridSim to simulate the scheduling process in a grid computing environment.

8 Experimental Results and Discussions

In this section, the experimental results are reported. The first sub section will present the performance comparison of priority rules algorithms while the following sub section mainly presenting the schedule-based algorithms performance. Those schedule-based algorithms are based on priority rules combined with backfilling algorithms mentioned before. Performance metrics used for comparison are delayed jobs, makespan, response time, total tardiness and machine usage.

8.1 Results for Priority Rules Algorithms

In this section, we compare the experimental results for FCFS, EDF, SJF, LJF, ERD and MTTD. These algorithms are responsible for selecting jobs to be put in the queue. One more important thing in priority rules scheduling is the selection of respective resources. Note that for this experiment we have considered of using the fastest resource available first as mentioned in [4],[1]. The fastest available resource will be selected to process the job first.

Table 1. Total number of delayed jobs

Scheduler	delayed jobs
FCFS	1967
EDF (Earliest Deadline First)	1951
SJF (Shortest Job First)	1697
LJF (Longest Job First)	1953
ERD (Earliest Release Date)	1967
MTTD (Minimum Time To Due Date)	1698

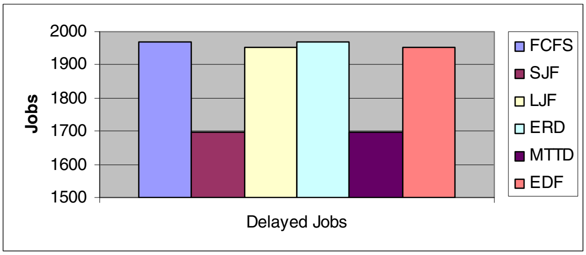


Fig. 2. Total number of delayed jobs

Table 1 and Figure 2 show the total number of delayed jobs. It can be seen that SJF outperform other scheduling algorithm. Total numbers of delayed jobs are 1697 for SJF but followed closely by MTTD which is 1698 numbers of jobs. Other priority rules schedulers are way too far behind from those numbers.

Table 2. Makespan time

Scheduler	makespan
FCFS	24486
EDF (Earliest Deadline First)	24316
SJF (Shortest Job First)	25106
LJF (Longest Job First)	19942
ERD (Earliest Release Date)	24262
MTTD (Minimum Time To Due Date)	24303

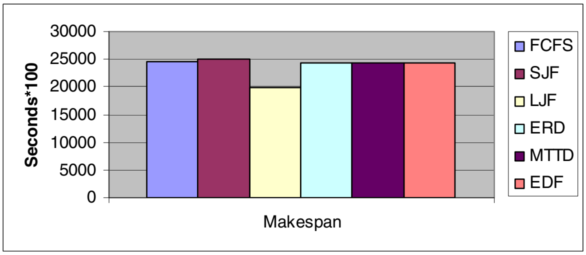


Fig. 3. Makespan time

While reducing the number of delayed jobs contribute to the QoS of the system, makespan time determine the overall time for the system. Table 2 and Figure 3 proved the Longest Job Fastest Resource (LJFR) theory [1]. It also means that more jobs are able to be completed in a shorter time by LJF compared to others. Another theory proven from these results is SJF shows the most makespan.

Table 3. Total Tardiness

Scheduler	tardiness
FCFS	15172016
EDF (Earliest Deadline First)	8203722
SJF (Shortest Job First)	10607690
LJF (Longest Job First)	20324174
ERD (Earliest Release Date)	15164512
MTTD (Minimum Time To Due Date)	7523838

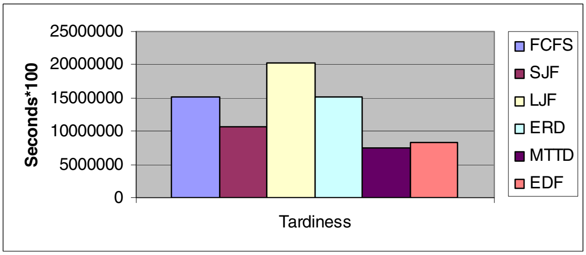


Fig. 4. Total Tardiness

Table 4. Percentage of Machine Usage

Scheduler	Machine usage
FCFS	88.896
EDF (Earliest Deadline First)	89.0425
SJF (Shortest Job First)	90.0245
LJF (Longest Job First)	86.373
ERD (Earliest Release Date)	88.8405
MTTD (Minimum Time To Due Date)	88.8515

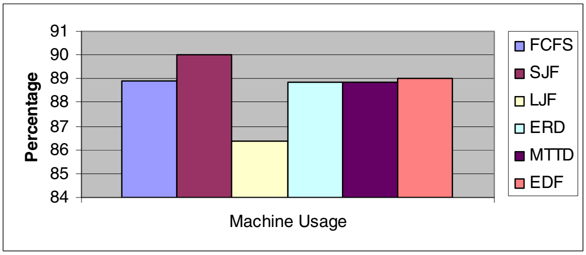


Fig. 5. Percentage of Machine Usage

Table 5. Average Response Time

Scheduler	Response time
FCFS	9024.624
EDF (Earliest Deadline First)	8358.4395
SJF (Shortest Job First)	6818.5905
LJF (Longest Job First)	11456.981
ERD (Earliest Release Date)	9019.339
MTTD (Minimum Time To Due Date)	8109.2295

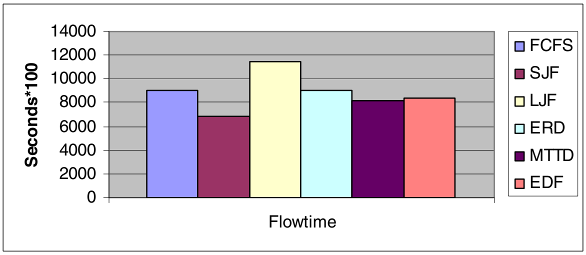


Fig. 6. Average Response Time

However, despite the poor makespan time, SJF do have advantage in other performance criteria. For example, SJF have the second least tardiness (Figure 4), the highest machine usage (Figure 5) and also the least response time (Figure 6) compared to other algorithms.

8.2 Results for Proposed Algorithms

Figure 7-11 shows the comparison between priority rules algorithms and proposed algorithms. The results obtained from the experiment is remarkable where almost all the results shows the improvement when backfilling is combined with priority-rules algorithms and some of them are better compared to the EG-EDF presented in [10]. For example, Figure 7 and Figure 10 show that both delayed jobs and percentage of machine usage performance using BF-SJF are way better then EG-EDF. The reason why there are significant improvement shows by these algorithms is simply because the scheduler will try to avoid as much as possible any idle machines during the jobs execution. By filling the gap between two jobs in the queue with newly arriving job, the schedulers have successfully optimized the system.

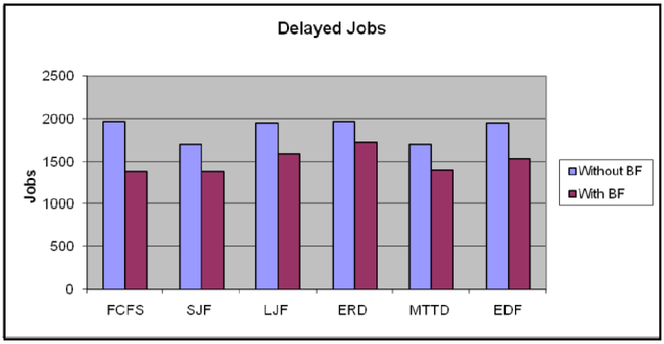


Fig. 7. Total number of delayed jobs

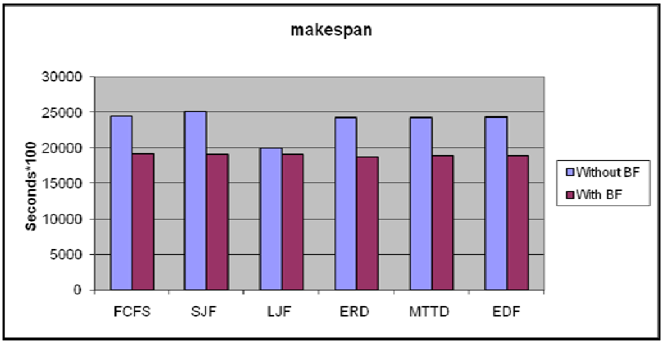


Fig. 8. Makespan time

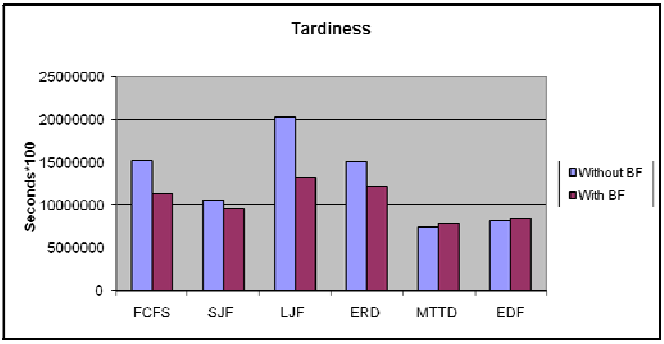


Fig. 9. Total Tardiness

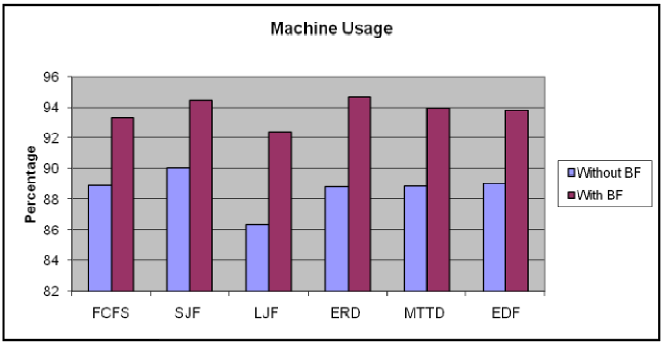


Fig. 10. Percentage of Machine Usage

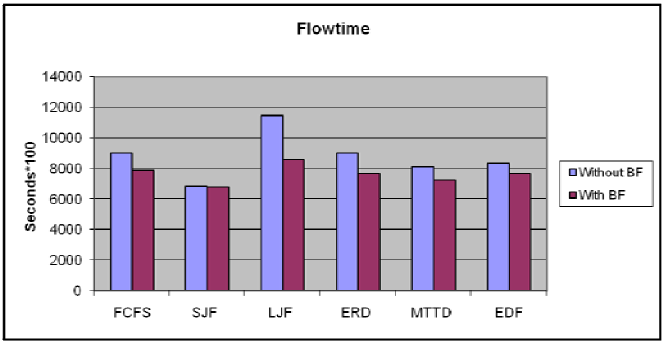


Fig. 11. Response Time

Another important achievement from this work is the fact that BF-SJF not only reduced the makespan of SJF but also make it fairly equal with BF-LJF and at the same time totally solve the problems Shortest Job Fastest Resource (SJFR) and Longest Job Fastest Resource (LJFR) in [1] which have said that any attempt to minimize makespan will result in maximization of response time. Based on the results shows in Figure 8 and Figure 11, we have proved that both makespan and response time can be reduced at the same time. We have successfully reduced both the LJF makespan and response time by using BF-LJF (4.3% for makespan and 24.9% for response time). For SJF, although BF-SJF only reduced 0.47% of the response time, it has also reduced the SJF makespan down by 23.8% which is the highest percentage compared to other algorithms tested. To recall, SJF have the highest makespan time compared to the other priority rules algorithms.

9 Conclusion

In this paper, we have presented a performance comparison of priority rules scheduling algorithms that schedule jobs in grid computing system. The results are very important as a reference for researchers that wish to use this technique as a seed or initial schedule for their schedule-based algorithms. To add novelty to this paper, we have proposed scheduling algorithms based on backfilling and EG-EDF technique proposed by Klusacek in [10]. We have proved that the backfilling technique can also be applied to other priority rules algorithms other then EDF and the results shows significant improvement compared to EG-EDF.

References

1. Abraham, A., Buyya, R., Nath, B.: Nature's heuristics for scheduling jobs on computational Grids. In: Proceedings of the 8th International Conference on Advanced Computing and Communications, pp. 45–52. Tata McGraw-Hill, India (2000)
2. Boeres, C., Lima, A., Rebello, V.E.: Hybrid Task Scheduling: Integrating Static and Dynamic Heuristics. In: Proceedings of the 15th Symposium on Computer Architecture and High Performance Computing. IEEE Computer Society, Washington, DC (2003)
3. Brucker, P.: Scheduling Algorithms. Springer, Berlin (2007)
4. Carretero, J., Xhafa, F.: Use of Genetic Algorithms for Scheduling Jobs in Large Scale Grid Applications. Journal of Technological and Economic Development, A Research Journal of Vilnius Gediminas Technical University 12(1), 11–17 (2006), ISSN 1392-8619
5. Chronopoulos, A.T., Benche, M., Grosu, D., Andonie, R.: A Class of Loop Self-Scheduling for Heterogeneous Clusters. In: Proceedings of the 3rd IEEE international Conference on Cluster Computing. IEEE Computer Society, Washington, DC (2001)
6. Garrido, J.M.: Performance modeling of operating systems using object-oriented simulation: a practical introduction. Kluwer Academic Publishers, Norwell (2000)
7. Henderson, R.: Job scheduling under the portable batch system. In: Job Scheduling Strategies for Parallel Processing, pp. 337–360. Springer, Berlin (1995)
8. Jackson, D., Snell, Q., Clement, M.: Core Algorithms of the Maui Scheduler. In: Feitelson, D.G., Rudolph, L. (eds.) JSSPP 2001. LNCS, vol. 2221, pp. 87–102. Springer, Heidelberg (2001)
9. Klusáček, D., Matyska, L., Rudová, H.: Alea – grid scheduling simulation environment. In: Wyrzykowski, R., Dongarra, J., Karczewski, K., Wasniewski, J. (eds.) PPAM 2007. LNCS, vol. 4967, pp. 1029–1038. Springer, Heidelberg (2008)
10. Klusacek, D., Rudova, H.: Improving QoS in Computational Grids through Schedule-based Approach. In: Scheduling and Planning Applications Workshop (SPARK) at the International Conference on Automated Planning and Scheduling (ICAPS 2008), Sydney (2008)

11. Klusacek, D., Rudová, H., Baraglia, R., Pasquali, M., Capannini, G.: Comparison of Multi Criteria Scheduling Techniques. In: CoreGRID Integration Workshop 2008. Integrated Research in Grid Computing. CoreGRID series. Springer, Heidelberg (2008)
12. Lifka, D.: The ANL/IBM SP Scheduling System. In: Job Scheduling Strategies for Parallel Processing (JSSPP), pp. 295–303 (1995)
13. Rasooli, A., Mirza-Aghatabar, M., Khorsandi, S.: Introduction of Novel Rule Based Algorithms for Scheduling in Grid Computing Systems. In: Second Asia International Conference on Modeling & Simulation (2008)
14. Singh, S.K.: Efficient Grid Scheduling Algorithm based on Priority Queues. Master of Engineering in Software Engineering Thapar University, Patiala (2008)
15. Srinivasan, S., Kettimuthu, R., Subramani, V., Sadayappan, S.: Characterization of backfilling strategies for parallel job scheduling. In: Proceedings of the International Conference on Parallel Processing Workshops, pp. 514–519. IEEE Computer Society Press, Los Alamitos (2002)
16. Techiouba, A.D., Capannini, G., Baraglia, R., Puppin, D., Pasquali, L., Ricci, M.: Backfilling Strategies for Scheduling Streams of Jobs on Computational Farms. Making Grids Work, CoreGRID series. Springer, Heidelberg (2008)
17. Tsafrir, D., Etsion, Y., Feitelson, D.G.: Backfilling Using System-Generated Predictions Rather than User Runtime Estimates. IEEE Transactions on Parallel and Distributed Systems 18(6), 789–803 (2007)
18. Xhafa, F., Abraham, A.: Meta-heuristics for Grid Scheduling Problems. In: Metaheuristics for Scheduling in Distributed Computing Environments. Series Studies in Computational Intelligence, vol. 146, pp. 1–37. Springer, Heidelberg (2008)

Dynamic Load Balancing Policy in Grid Computing with Multi-Agent System Integration

Bakri Yahaya, Rohaya Latip, Mohamed Othman, and Azizol Abdullah

Department of Communication Technology and Network,
Faculty of Computer Science & Information Technology,
Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
bakriy@gmail.com, {rohaya, mothman, azizol}@fsktm.upm.edu.my

Abstract. The policy in dynamic load balancing, classification and function are variety based on the focus study for each research. They are different but employing the same strategy to obtain the load balancing. The communication processes between policies are explored within the dynamic load balancing and decentralized approaches. Multi-agent system characteristics and capabilities are explored too. The unique capabilities offered by multi-agent systems can be integrated or combined with the structure of dynamic load balancing to produce a better strategy to produce a better dynamic load balancing algorithm with multi-agent systems.

Keywords: Dynamic load balancing, policy, multi-agent system, grid.

1 Introduction

There are many load balancing techniques proposed in grid computing environment such as randomized load balancing, round robin load balancing, dynamic load balancing, hybrid load balancing, agent based load balancing and multi-agent load balancing. Round robin and randomized load balancing are considered as basic, simple and easy to implement but not for dynamic, hybrid, agent and multi-agent load balancing. These load balancing methods has undergone an improvement or new ones introduced in the grid load balancing solution.

The load balancing goal is to fully utilize the computing power from multiple hosts without the disturbing the user, regardless of the number of hosts available in the background and aims to improve the overall performance. Besides, load balancing aims to ensure that the workload is fairly distributed among the nodes and that none of the nodes are overloaded or under loaded. Basically, there are two load balancing strategies to consider off, which are called static load balancing and dynamic load balancing,.

Static load balancing makes the balancing decision at compile time and it will remain constant. Meanwhile the dynamic load balancing makes more informative decisions in sharing the system load based on runtime state. Comparatively, dynamic

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 416–424, 2011.

load balancing have the potential to provide better performance than static load balancing. Dynamic load balancing which are based on runtime state needs to process the collected information with firm procedures. The balancing procedures are placed in the dynamic load balancing policy. It contains of a set of rule referred by the system to run and to employ dynamic load balancing for better performance.

System and network performance issues have been explored previously by many researchers. Some look into the resource management, scheduling strategy and load balancing strategy which aim to improve the performance of grid computing [7,8,9,11]. This paper will discuss about the dynamic load balancing in grid computing and multi-agent system (MAS). The paper is organized as follows. In Section 2, we discuss the dynamic load balancing policy. Section 3, we carry on the discussion with multi-agent system. The implementation strategy in section 4 and conclusion in the research is discussed in section 5.

2 Related Work

Dynamic load balancing can be classified into as the centralized approach and the decentralized approach. The Centralized approach is managed by central controller that has a global view of load information in the system which is used to decide how to allocate jobs to each node. In the decentralized approach all joint nodes are involved in making the load balance decision [1]. Dynamic load balancing are based on redistribution of tasks among the available processors during execution time [2]. It transfers the tasks from heavily loaded processors to the lightly loaded ones [4]. Therefore, none of the nodes are heavily loaded.

2.1 Dynamic Load Balancing Policy

The decisions to balance the workload are based on the setup policy. Dynamic load balancing considers and involves four policies [2,3] which consists of transfer policy, selection policy, location policy and information policy. The dynamic load balancing algorithm proposed in [4] takes into consideration the load estimation policy, process transfer policy, state information exchange policy, priority assignment policy and migration limiting policy. Table 1 describes the Dynamic Load Balancing Policy.

Although the policy structure used are diverse but they apply the same strategy to implement the proposed dynamic load balancing solution. It is still involved with information, selection, location and transfer policy which act as the basic policy. These policies work closely pertaining to each unique role and share the decision made to the related or needed policy for the subsequent process. Policy processes or communications will be discussed in detail in 2.2.

Table 1. Dynamic Load Balancing Policy

No	Policy	Function
1	Transfer Policy	- Need for load balance to initiate. - Determine the condition under which a task should be transferred.
2	Selection Policy	- Select job to transfer. - Select a task for transfer.
3	Location Policy	- Determine under loaded node. - To find a suitable transfer partner. - To check the availability of the service(s) required for proper execution of migration.
4	Information Policy	- Containing all needed information. - To decide the time when the information about the state of other hosts in the system is to be collected

2.2 Dynamic Load Balancing Policy Communication

Policies in the dynamic load balancing need to communicate with each other to determine the processes involved. They need to share the information or decision made to ensure that the subsequent process is able to start. Figure 1 portrays the interaction or communication path among the policy in the dynamic load balancing algorithm.

The incoming jobs are directed to the *transfer policy* to determine whether it should be transferred or not and it is based on various criteria such as workload value and computing power. In other word, it determines the need for the load balancing. If load balance is needed, so the decision will be sent to the selection policy. If not, the jobs will process locally. Receiving the information from transfer policy indicate the *selection policy* to commence the job selection for transference or migration. The decisions made by selection policy are then directed to the location policy for further process.

The *location policy* is responsible to determine the under loaded node, to find a suitable transfer partner and to check the service availability. If location policy managed to fulfill all the requirement then the particular jobs will be migrated. Otherwise, the local processor will be appointed to process the jobs.

In brief, the information policy plays a big role or possess high responsibility in dynamic load balancing. Information policy provides the transfer policy and location policy with the necessary information in order for them to build their decision.

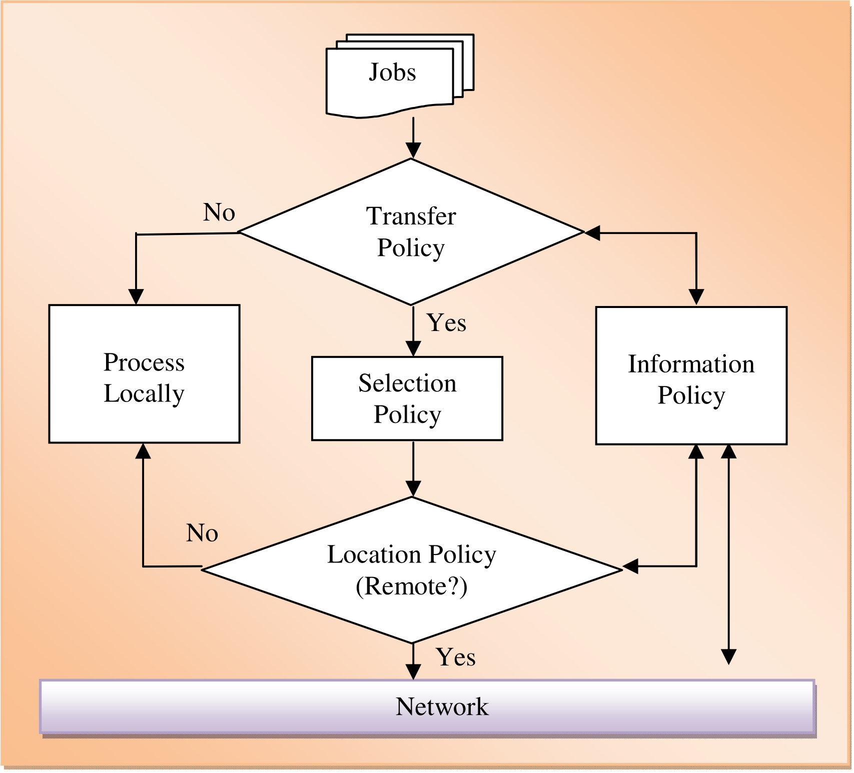


Fig. 1. Policy Interaction in Dynamic Load Balancing

3 Multi-Agent System

An agent is a computer system that is capable of independent action on behalf of its user or owner. The Agent can figure out for itself what it needs to do in order to satisfy its design objectives. Basically agent is developed to provide services to a particular system. The communication is done by exchanging messages through a computer network. Some of the agents are developed to interact cooperatively and some interact competitively between the agents.

The agents in multi-agent system hold several characteristics such as autonomy, local views, cooperation, social ability, reactivity, proactivity, goal oriented and decentralized. Agents are developed with layer structure and it consists of [11] communication layer, coordination layer and local management layer. The communication layer provides an agent with interfaces to heterogeneous networks and operating systems. It will receive the request and then explain and submit to the coordination layer to decide the suitable action according to its own knowledge. The local management layer performs functions of an agent for local grid load balancing.

A Multi-agent system is composed of multiple intelligent agents that have the ability to interact or communicate, collaborate and negotiate among them. The cooperation between agents permits them to accomplish a common goal. For management and scheduling to be effective, such system must develop intelligent and autonomous decision making techniques [11]. Concordantly the agent itself should also poses intelligence on their own role. In connection with that, agent can make decisions without direct orders or interference. This social ability enables the agent to get assistance from other agents for tasks that are difficult to handle independently. This allows a multi-agent system to have the capability to solve problems which are difficult or impossible for individual agents or monolithic system to solve. The structure of generic multi-agent system is illustrated in Figure 2.

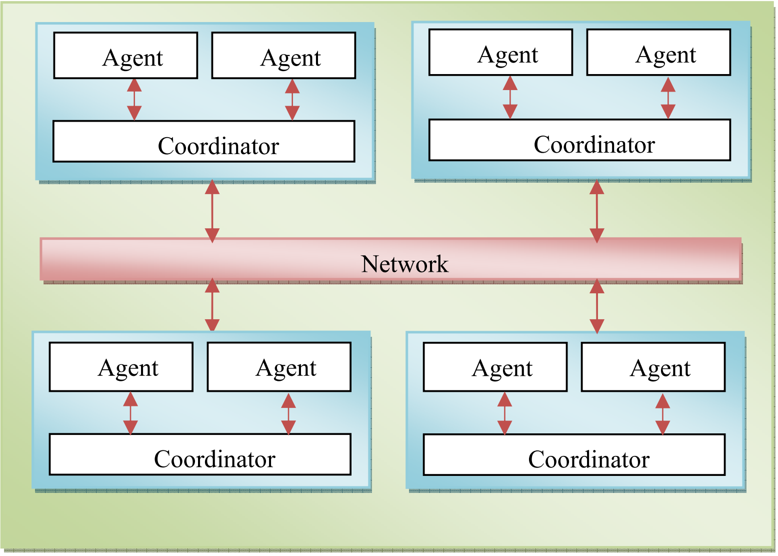


Fig. 2. Structure of Multi-Agent System

Recently, many load balancing approaches in grid have been suggested, using Multi-Agents. By using Multi-agent system [5], brings to the fore an efficient dynamic load balancing scheme to retrieve and provide the agent-based services. The proposed dynamic load balancing is implemented using new definitions of models and policies on load data collection and agent migration. They employed a credit-based index model to decide which agent needs to be migrated using the credit value. They utilized the load data collection policy, agent selection policy and destination selection policy to enable the dynamic load balancing. The structure proposed complies with FIPA and Figure 3 shows the FIPA agent management reference model.

In [6] prediction-based dynamic load balancing has been proposed, using multi-agent system that predicts the load of the agent based on the predicted data and measured data. They also employ the data collection policy, agent selection policy and migration policy to enable the dynamic load balancing. The proposed scheme succeeds in avoiding unnecessary agent migration and reduced the communications overhead. Figure 4 elucidate the agent platform structure use in the paper.

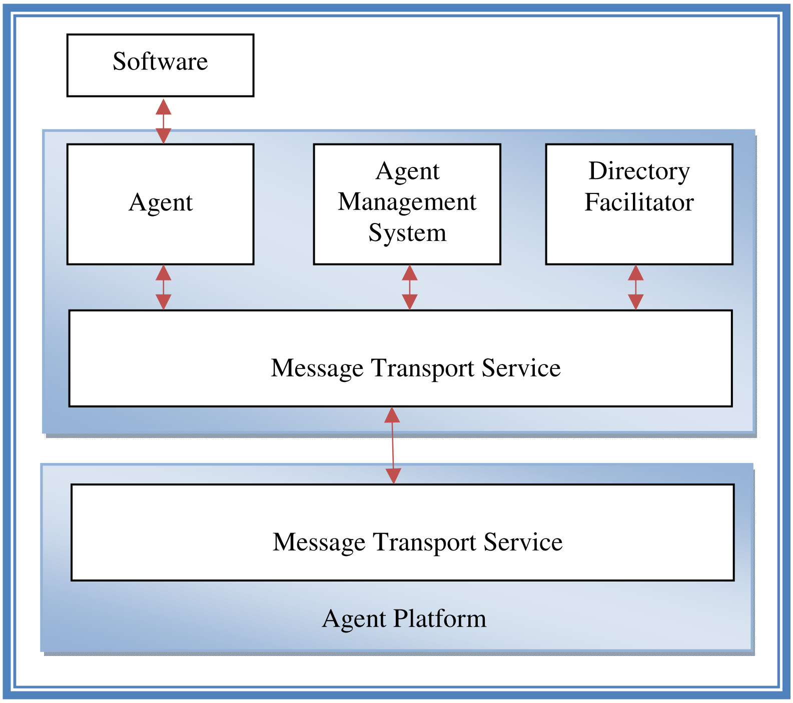


Fig. 3. The FIPA Agent Management Reference Model

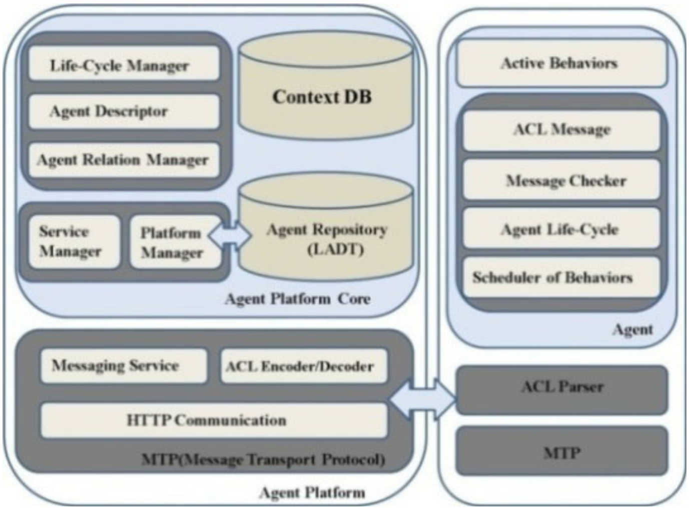


Fig. 4. The Structure of Developed Agent Platform

4 Implementation Strategy

The information policy contributes a lot to the decision making process in the dynamic load balancing and hold the authority in that sense. Besides, we can conclude that information policy has a big implication on performance in grid computing through accurate, suitable and efficient parameter use. In fact, the information policy are connected directly to the information directory called index load, profiling or task profiling that does the workload management too. Index load issues [5,6] explored the weightage strategy using the credit-based index model in considering the load balancing factor. But this paper will explore a different method based on agent using the computing power information with an adapted method.

Agents will be developed with multifunction capabilities due to the role embedded into them. They will be in 2 (two) statuses which are as leader of the computing element or as worker of the computing element. The agent will determine what they are and automatically turn themselves into the determined status or role. If the agent is a leader, it will auto-notify the workload system manager. The agent itself has the capability to communicate among the agent and performs the information exchange.

In this paper load balancing function will be implemented at the global and local grids. In the global grid the load balancing decision will be made by workload system manager which sits at the top of the grid described in Figure 5. It makes the decision based on computing element power or index. This is to allocate the correct load value to the correct computing elements which are the leaders in the local grid. This is based on information provided by the computing element leaders to the workload system manager. Then, the computing element leader will decide how to distribute the load according to the worker node computing power available. The worker node will auto-notify the computing element leader on its computing power information if there are any changes to its states since its last update. This will also reduce the communication overhead compared to the polling method.

This implementation will utilize all the discussed policies for the dynamic load balancing which are transfer policy, selection policy, location policy and information policy. The transfer policy is combined with the selection policy, to be known as migration policy as depicted in Figure 6. This will reduce the internal communication between the policies in the agent. The migration policy will be the main door way for receiving data. As it is already holding data, it will analyze the load and decide whether to process locally or remotely. The decision made by migration policy will submit to the location policy to look for a processing partner. The information policy will play a role as information collector to supply information for the agent to make a decision. The proposed strategy is planned to comply with FIPA model.

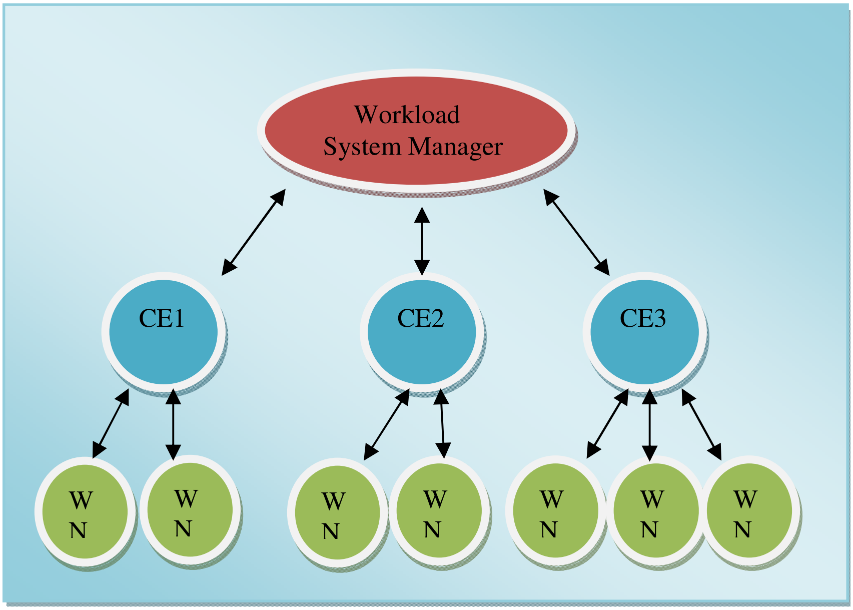


Fig. 5. Grid Structure Environment

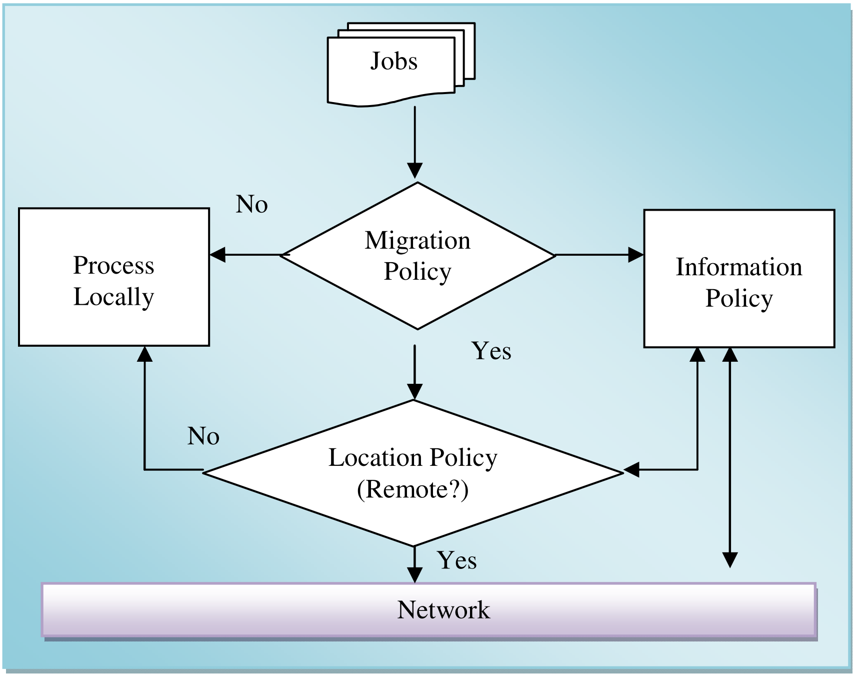


Fig. 6. Proposed Policy Interaction in Dynamic Load Balancing

5 Conclusion

Multi-agent system capabilities are embedded broadly in various discipline of study. The successes of integration triggered more researchers to explore these opportunities. The unique capabilities offered by multi-agent system can be integrated into or combined with the structure of dynamic load balancing to produce a better strategy in producing a better dynamic load balancing algorithm with multi-agent system. In the future, this paper will explore the algorithm development that will be embedded into the agent and to implement the testing.

References

1. Lu, K., Subrata, R., Zomaya, A.: An Efficient Load Balancing Algo-rithm for Heterogeneous Grid Systems Considering Desirability of Grids Sites, pp. 311–319. IEEE, Los Alamitos (2006)
2. Mukhopadhyay, R., Ghosh, D., Mukherjee, N.: A Study On The Application Of Existing Load Balancing Algorithms For Large. Dynamic. Heterogeneous Distributed Systems, 238–243 (2010)
3. Janhavi, B., Surve, S., Prabhu, S.: Comparison of Load Balancing Algorithms in a Grid, pp. 20–23. IEEE, Los Alamitos (2010)
4. Chhabra, A., Singh, G., Waraich, S.S., Sidhu, B., Kumar, G.: Qualitative Parametric Comparison of Load Balancing Al-gorithms in Parallel and Distributed Computing Environment. In: WASET, vol. 16, pp. 39–42 (2006)
5. Kim, Y.H., Han, S., Lyu, C.H., Youn, H.Y.: An Ef-ficient Dynamic Load Balancing Scheme for Multi-Agent System Reflecting Agent Workload, pp. 216–222. IEEE, Los Alamitos (2009)
6. Son, B.H., Lee, S.W., Youn, H.Y.: Prediction-Based Dy-namic Load Balancing Using Agent Migration for Multi-Agent System, pp. 485–490. IEEE, Los Alamitos (2010)
7. Miyashita, M., Haque, E., Matsumoto, N., Yoshida, N.: Dynamic Load Distribution in Grid Using Mobile Threads, pp. 629–634. IEEE, Los Alamitos (2010)
8. Bohlouli, M., Analoui, M.: Grid-HPA: Predicting Resource Requirements Of A Job In The Grid Computing Environment. In: WASET, vol. 45, pp. 747–751 (2008)
9. Lee, J., Keleher, P., Sussman, A.: Decentralized Dynamic Sched-uling Across Heterogeneous Multi-core Desktop Grids. IEEE, Los Alamitos (2010)
10. Foundation for Intelligent Physical Agents, `http://www.fipa.org/`
11. Cao, J.W., Spooner, D.P., Jarvis, S.A., Nudd, G.R.: Grid Load Balancing Using Intelligent Agents. Future Generation Computer System 21, 135–149 (2005)

Dynamic Multilevel Dual Queue Scheduling Algorithms for Grid Computing

Syed Nasir Mehmood Shah, Ahmad Kamil Bin Mahmood, and Alan Oxley

Department of Computer and Information Sciences
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, Tronoh,
31750, Perak, Malaysia
nasirsyed.utp@gmail.com, {kamilmh, alanoxley}@petronas.com.my

Abstract. Grid computing is the enabling technology for high performance computing in scientific and large scale applications. Grid computing introduces a number of fascinating issues to resource management. Grid scheduling is a vital component of a Grid infrastructure. Reliability, efficiency (in terms of time consumption) and effectiveness in resource utilization are the desired quality attributes of Grid scheduling systems. Many algorithms have been developed for Grid scheduling. In our previous work, we proposed two scheduling algorithms (the Multilevel Hybrid Scheduling Algorithm and the Multilevel Dual Queue Scheduling Algorithm) for optimum utilization of CPUs in a Grid computing environment. In this paper, we propose two more flavours of Multilevel Dual Queue scheduling algorithms, i.e. the Dynamic Multilevel Dual Queue Scheduling Algorithm using Median and the Dynamic Multilevel Dual Queue Scheduling Algorithm using Square root. We evaluate our proposed Grid scheduling, in comparison to other well known scheduling algorithms, on an SGI super computer using parts of the 'AuverGrid' workload trace.

The main purpose of scheduling algorithms is to execute jobs optimally, i.e. with minimum average waiting, turnaround and response times. An extensive performance comparison is presented using real workload traces to evaluate the efficiency of the scheduling algorithms. To facilitate the research, a software tool has been developed which produces a comprehensive simulation of a number of Grid scheduling algorithms. The tool's output is in the form of scheduling performance metrics. The experimental results, based on performance metrics, demonstrate that our proposed scheduling algorithms yield improvements in terms of performance and efficiency.

Our proposed scheduling algorithms also support true scalability, that is, they maintain an efficient approach when increasing the number of CPUs or nodes. This paper also includes a statistical analysis of the 'AuverGrid' real workload traces to show the nature and behavior of jobs.

Keywords: Distributed systems; Cluster; Grid computing; Grid scheduling; Workload modeling; Performance evaluation; Simulation; Load balancing; Task synchronization; Parallel processing.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 425–440, 2011.

1 Introduction

Grid computing is a mainstream technology for large-scale resource sharing. Grid computing increases a system's computing capability and the tendency has been to use them to solve complex and large-scale scientific problems using geographically dispersed computing resources. A large number of complex and large-scale scientific issues cannot be solved by using a traditional network; that is why research and development into an alternative, Grid computing, has been taking place. A Grid computing system connects available computing resources, such as computers, applications, and storages devices, to networks for high performance computing and the reduction of system execution time [1, 2].

Grid scheduling is a process of ordering tasks on compute resources and ordering communication between them. It is also known as the allocation of computation and communication over time [3]. Grid scheduling can be divided into three stages. Stage one is resource discovery, which provides a list of available resources. The second stage is the selection of appropriate resources for job allocation, and the third stage is the placement of jobs onto the selected resource queue for execution [4]. In the second stage, the selection of the best match of jobs to resources is an NP-complete problem.

Grid scheduling becomes more challenging when performance and scalability issues are considered. Scalability can mean two things. First, it is a measure of how the efficiency of a Grid is affected by using more processing power given to it in the form of additional CPUs or cores. Grid vendors often refer to scalability as a measure of parallelizing an application across different nodes. The second meaning for scalability is given in [5]. The authors defined the concept of performance and scalability as "The terms 'performance' and 'scalability' are commonly used interchangeably, but the two are distinct: performance measures the speed with which a single request can be executed, while scalability measures the ability of a request to maintain its performance under increasing load."

Two fundamental issues have to be considered for the performance evaluation of new Grid scheduling algorithms. First, "… representative workload traces are needed to produce dependable results." [6] Second, "… a good testing environment should be set up, most commonly used through simulations."

The motivation of this paper is to develop Grid scheduling algorithms that can perform efficiently and effectively in terms of overall Grid efficiency, performance and scalability. In our previous work, we proposed two Grid scheduling algorithms (the Multilevel Hybrid Scheduling Algorithm and the Multilevel Dual Queue Scheduling Algorithm (MDQ)) [7]. We evaluated the performance of these scheduling algorithms, in comparison to four other scheduling algorithms, in processing the 'LCG1' real workload trace. The performance of MDQ is dependent on the value of a fixed time quantum. If the value is too small then MDQ results in too many context switches. If the value is too large then MDQ loses its efficiency and behaves like the First Come First Served scheduling algorithm (FCFS). In this paper we present a solution to the fixed time quantum problem by proposing two more flavours of MDQ (namely the Dynamic Multilevel Dual Queue scheduling algorithm using Median (MDQM) and the Dynamic Multilevel Dual Queue scheduling algorithm using Square Root (MDQR)). In this paper, we evaluate the performance of our proposed scheduling algorithms in comparison to other well known algorithms in using an SGI super computer to process the 'AuverGrid' workload trace.

The structure of the paper will now be described. Section 2 is a literature review of Grid scheduling methodologies. Section 3 presents the proposed scheduling algorithms and section 4 is about the statistical analysis of real workload traces. Section 5 shows the homogenous implementation of the new scheduling algorithms. In section 6, the scheduling simulator's design and development are discussed. Section 7 describes the experimental setup and section 8 shows the simulation data. Section 9 shows the performance analysis of the Grid scheduling algorithms and section 10 concludes the paper.

2 Related Research

A 'Grid' is an infrastructure for large scale resource sharing. It consists of a large number of distributed and heterogeneous resources [8]. Grid computing is an emerging technology for solving large-scale problems in science, engineering, and commerce. It uses resources of many computers connected by a network. Scientific applications usually consist of numerous jobs that process and generate large datasets [1, 9]. Grid computing allocates these jobs to appropriate processors for efficient and effective execution. Therefore, a Grid needs scheduling policies that assign various jobs to processors. Scheduling policies also decide on the order of processing of allocated jobs and manage computing resources in such a way as to optimize system computing capabilities. Scheduling policies are directly related to optimizing Grid performance [10, 11, 12, 13].

Grid job scheduling policies can generally be categorized into space-sharing and time-sharing policies. In space-sharing policies, however, each job is allocated to only one processor until its completion. The well known space-sharing policies are FCFS, Job Rotate Scheduling Policy (JR), Multilevel Opportunistic Feedback (MQF), Shortest Job First (SJF), Shortest Remaining Time First (SRTF), Longest Job First (LJF), Priority (P) and Non Preemptive Priority. In time-sharing policies, processors are temporally shared by jobs; and the famous time-sharing scheduling policies are Round Robin (RR) and Proportional Local Round Robin Scheduling [14, 15, 16].

In [16] the author performed a comparative performance analysis of three space-sharing policies (i.e. FCFS, JR and MQF) and two time-sharing policies (i.e. Global Round Robin and Proportional Local Round Robin scheduling), all of which have been designed and developed for Grid computing. [16] concludes that time-sharing scheduling policies perform better than space-sharing scheduling policies.

In [17] the authors proposed the idea of 'backfilling'. Backfilling is a space sharing policy that allows a scheduler to make better utilization of available resources by running jobs in prioritized order. Smaller jobs are assigned a higher priority than larger queued jobs. It also requires that all job service times must be known before a scheduling decision is taken. Proposed method has been evaluated using the IBM SP2 system. [17] also states that the service time can be estimated by the users at the time of job submission, or predicted by the system based on historical data.

In [18] the authors give a performance analysis of CPU scheduling algorithms using a simulation of a Grid computing environment. Three space-sharing scheduling algorithms (i.e. FCFS, SJF and P) are used.

[19] proposes a Grid level resource scheduling technique with a Job Grouping strategy so as to maximize resource utilization and minimize the processing time of jobs. A combination of Best Fit and RR is applied at the local level to achieve better performance. In RR, a fixed time quantum is given to each process present in the circular queue for fair distribution of CPU times. (In Grid computing, RR is most often chosen for job scheduling [16, 19, 20].)

[21] proposes a Self-Adjustment-Round-Robin (SARR) scheduling approach based on a dynamic-time-quantum algorithm. For this algorithm the ready queue is managed as a FIFO queue. A process can execute up to the value of a computed time quantum for each round. This approach computes the time quantum, which is repeatedly adjusted according to the CPU burst time of the now-running processes. The time quantum is calculated by taking the median of the remaining CPU times of all processes in the ready queue. The minimum value for the time quantum used in the algorithm is 25 units. In our paper we also evaluate SARR . As far as we are aware, this is the first time that this has been done in a Grid environment.

In [22] the author introduced a dynamic scheduling model for parallel machines from an implementation perspective. The proposed model of a parallel job is based on a penalty factor. This paper also lists open issues for researchers. First, the theoretical and experimental analysis of idle regulation is needed with more variations of job scheduling strategies (LJF, back filling, etc.) and optimization criteria, from both a user and a system perspective. Second, what is needed is the analysis of the system in a practical scheduling environment that supports dependent jobs, and jobs that can arrive at any moment.

To evaluate job scheduling policies effectively we require realistic workloads that can be used in an experiment to measure the scheduling performance. The result of a performance evaluation is strongly dependent on the workload used [23]. The use of a workload that does not correctly represent the real situation may lead to inaccurate performance evaluation.

Most of the scheduling algorithms highlighted in the literature have not been evaluated using real workload traces. The aim of this paper is to evaluate the performance and scalability of proposed scheduling algorithms in comparison to other well known scheduling algorithms, using an SGI super computer with real workload traces. Our scheduling performance metrics are average waiting time, average turnaround time and average response time.

3 Proposed Scheduling Algorithms

In [7, 24, 25] we proposed two scheduling algorithms - MH and MDQ. They are based on a fixed time quantum. In this paper we propose two new Dynamic Multilevel Dual Queue scheduling algorithms namely MDQM and MDQR. MDQ, MDQM and MDQR will now be described.

3.1 Multilevel Dual Queue Scheduling Algorithm (MDQ)

MDQ is based on a master/ slave architecture. MDQ employs a round robin allocation strategy for job distribution among slave processors; the Dual Queue Scheduling algorithm (DQ) is used on each slave processor for computation. Once a computation is completed at the slave processor, then notification is sent to the master processor. A block diagram of MDQ is shown in Fig.1.

A process state diagram of DQ is shown in Fig. 2. For the DQ algorithm, the ready queue comprises two queues – the waiting queue and the execution queue. The waiting queue is maintained as an FIFO queue. A new process submitted to the slave is linked to the tail of the waiting queue. Whenever the execution queue is empty, all processes in the waiting queue are moved to the execution queue, leaving the waiting queue empty. The execution queue is maintained in order of CPU burst length, with the shortest burst length at the head of the queue. Two numbers are maintained. The first number, t_{large} , shows the burst length of the largest process in the ready queue (waiting queue and execution queue combined) while the second one, t_{exec} , represents a running total of the execution time of all processes (since a reset was made). The algorithm dispatches processes from the head of the execution queue for execution by the CPU. Processes being executed are preempted on expiry of a time quantum, which is a system-defined variable. Following preemption, t_{exec} is updated as follows:

$$t_{exec} = t_{exec} + quantum$$

The numbers are then compared. If $t_{exec} < t_{l\arg e}$ then the preempted process is linked to the tail of the execution queue. The next process is then dispatched from the head of the execution queue. If $t_{exec} \geq t_{l\arg e}$ then the process with the largest CPU burst length is given a turn for execution. Upon preemption, all processes in the waiting queue are moved to the execution queue, leaving the waiting queue empty. The execution queue is then sorted on the basis of SJF. The value of t_{large} is reset to the burst length of the largest PCB and t_{exec} is reset to 0. The next process is then dispatched from the head of the execution queue. When a process has completed its task it terminates and is deleted from the system. t_{exec} is updated as follows:

$$t_{exec} = t_{exec} + \text{ time to complete}$$

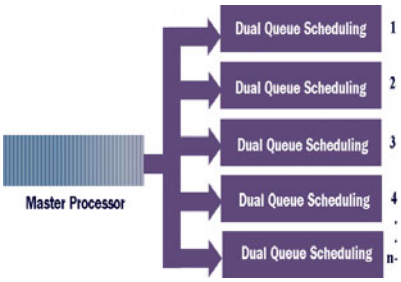


Fig. 1. Block Diagram of MDQ

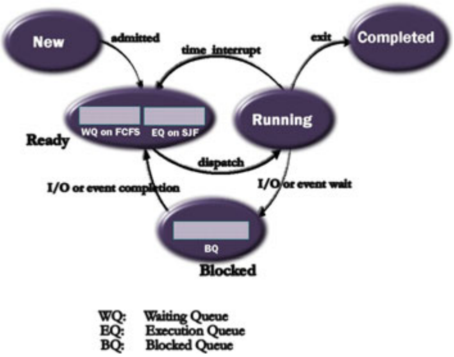


Fig. 2. Process State Diagram of DQ

3.2 Dynamic Multilevel Dual Queue Scheduling Algorithm Using Median (MDQM)

Our proposed MDQM algorithm is a variant of MDQ. MDQM works in the same manner as the MDQ but uses a dynamic time quantum approach instead of fixed time quantum. MDQM computes the dynamic time quantum by taking the median of CPU times of processes in the ready queue. We used the dynamic time quantum approach as detailed in [21].

$$TimeQuantum = median(C_1, C_2, C_3, ... C_n)$$

where C_i is the estimated CPU time of process *i,* and *i* ranges from 1 to *n*.

3.3 Dynamic Multilevel Dual Queue Scheduling Algorithm Using Square Root (MDQR)

Our proposed MDQR algorithm is another flavour of MDQ. MDQR calculates the dynamic time quantum using the square root of the average of CPU times of processes in the ready queue. MDQR computes the time quantum for each round and executes processes for the computed dynamic time quantum value. This approach also reduces the number of context switches in the system.

$$TimeQuantum = sqrt(avg(C_1, C_2, C_3, ... C_n))$$

where C_i is the estimated CPU time of process *i,* and *i* ranges from 1 to *n*.

Our proposed dynamic scheduling algorithms (MDQM and MDQR) solve the fixed time quantum problem encountered with MDQ.

4 Statistical Analysis of Real Workload Traces

In [26], a comprehensive statistical analysis has been carried out for a variety of workload traces on clusters and Grids. We reproduced the graphs of [27] to study the behaviour of the dynamic nature of workload ‘AuverGrid’, using our developed software. The total numbers of jobs in AuverGrid are 404176. We looked at the number of jobs arriving in each 1024 second period (i.e. the interval size). The number of jobs arriving in a particular period is its ‘job count’. In our experiments, we drop trace job entries that have a negative runtime or a negative number of allocated processors. The left side of figure 3 shows the number of jobs submitted by each user, whilst the right of figure 3 shows the top 20 users. User ‘U3034S2’ is the top most user of AuverGrid, who submitted ‘18021’ jobs to the system for execution; it is highlighted by a tooltip in Figure 3. Users belong to groups. The left side of figure 4 shows the number of jobs submitted by each group, whilst the right of figure 4 shows the top 10 groups. Group ‘G3’ had submitted the largest number of jobs. Figure 5 shows the distribution of job counts and run time CPU demands for the whole trace. Next we performed an autocorrelation of the job counts at different lags. The left hand graph of figure 6 shows the autocorrelation plot. Then we performed a Fourier analysis by applying the FFT on the values of the autocorrelation output. This is shown in the right hand graph of figure 6.

Figures 3 to 6 depict that job arrivals show a diversity of correlation structures, including short range dependencies, pseudo periodicity, and long range dependence. Long range dependencies can cause significant performance degradation, whose effects should be taken into consideration in evaluating of scheduling algorithms.

Figures 5 and 6 show that 'AuverGrid' has a rich correlation and scaling behaviour, which is different from conventional parallel workloads and cannot be captured by simple models such as Poisson or other distribution based methods. 'AuverGrid' played a key role in the performance evaluation of our proposed scheduling algorithms.

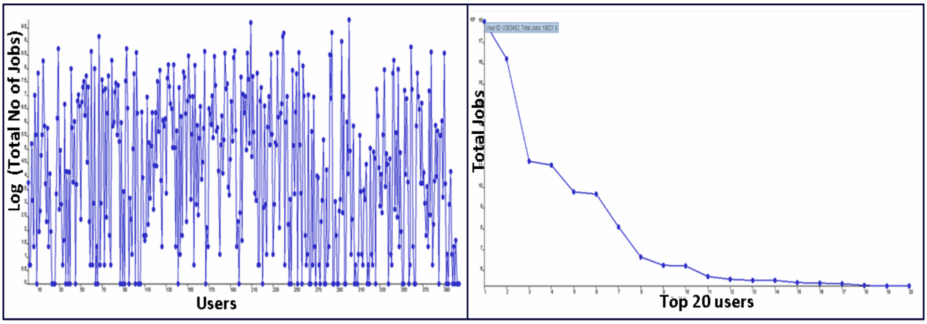


Fig. 3. The user jobs and top 20 users for AuverGrid

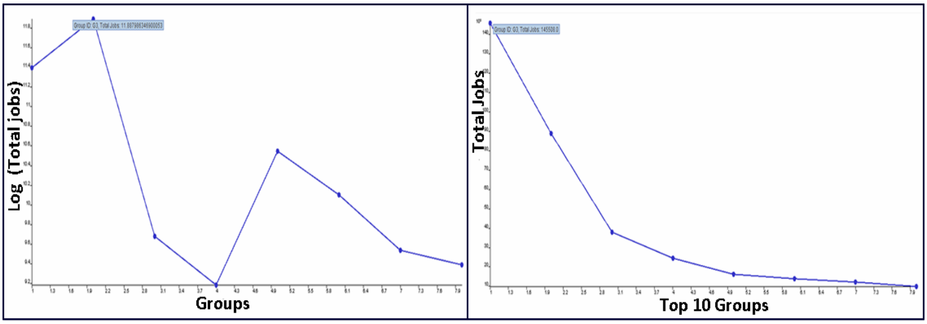


Fig. 4. The Group jobs and top 10 Groups for AuverGrid

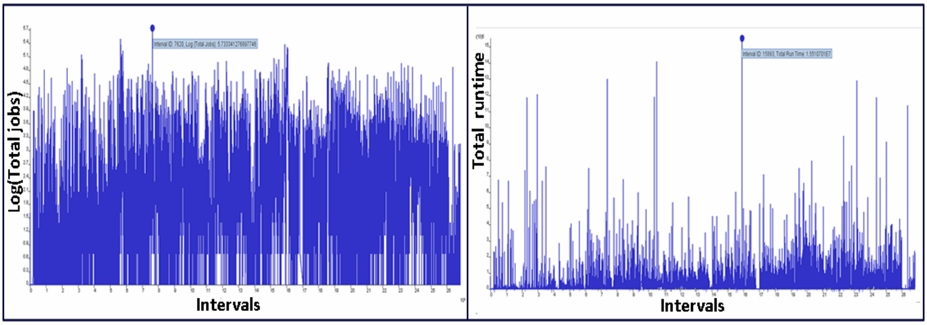


Fig. 5. The sequence plot and run time demand for the count process of AuverGrid

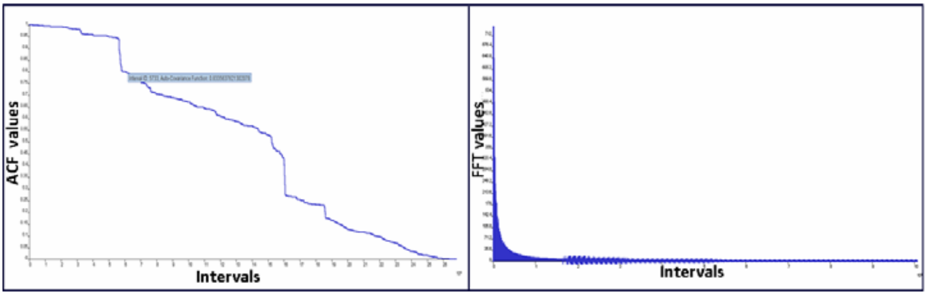


Fig. 6. The autocorrelation function(ACF) and Fast Fourier transformation(FFT) for the count process of AuverGrid

5 Homogeneous Implementation of Proposed Scheduling Algorithms

In this paper we used the same implementation strategy as we detailed in [7]. For comprehensiveness, the implementation strategy will now be explained.

We used a master/slave architecture for implementation of our proposed algorithms, as shown in Fig.7. One of the cluster node processors is dedicated as the master processor. The master processor is responsible for distribution of the workload among the slave processors using round-robin scheduling (i.e. 1, 2, 3…. n, 1).

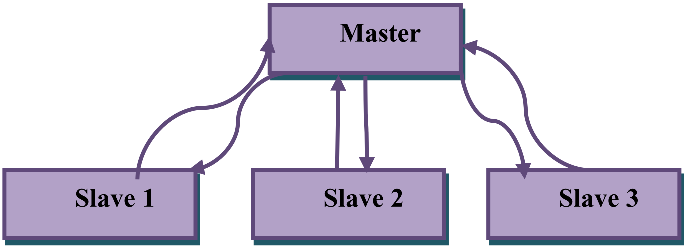


Fig. 7. Block diagram of master/slave architecture

The same algorithm, either MDQ or MDQM, is used on each slave processor. Once a computation is completed, notification is sent to the master processor.

6 Scheduling Simulator Design and Development

In this paper we used the same development strategy as we explained in [7]. For comprehensiveness, the development strategy will now be explained.

We developed a Java based simulator to evaluate the efficiency of our dynamic scheduling algorithms. The MPJ Express Java messaging system is an open source and free implementation of the mpiJava 1.2 API. We used MPJ Express for development of our simulator. The simulation software has two main programs. One program runs on the master node (SimM). The other program runs on each slave processor (SimS). The metadata for each process includes its ID, its arrival time, its CPU burst time, its priority and the number of slaves that the job is to be divided between. SimM uses a workload and distributes among slave processors in a round robin fashion. SimM receives notification from each slave processor for each job (or part of a job) that has finished. Each slave runs SimS and computes the average waiting time; the average turnaround time and the average response times. SimS processes the metadata for the list of processes that have been assigned to it. Upon completion of a process, SimM is informed. SimS keeps a detailed record of the processes being run on the slave - process ID; CPU burst length; arrival time; time quantum.

All slaves use the same CPU scheduling algorithm, the choice being input by the user of SimM. The user can select one of a range of algorithms including the newly developed ones, MH, MDQ, MDQM and MDQR, as well as established ones, FCFS, SJF, SRTF, RR, and SARR. The purpose of the simulator is to produce a comparative performance analysis of Grid scheduling algorithms using real workload traces.

7 Experimental Setup

AuverGrid has been processed using all of the scheduling algorithms listed above. The results have been evaluated. The experiments made use of a HPC facility in the HPC Service Center at Universiti Teknologi PETRONAS. We ran our experiment using a cluster of 128 processors. hpc.local was used as the default execution site for job submission. A detailed experimental setup is shown in Table 1.

Table 1. Experimental Setup

Name	Type	Location	Configuration
gillani	Shell terminal	Lab Workstation	Intel Core 2 Duo CPU 2.0GHZ 2 GB Memory
hpc.local	Execution site	HPC facility	SGI Altix 4700 128 Core Intel(R) Itanium2(R) Processor 9030 arch : IA-64 CPU MHz : 1.6 GHz

8 Simulation Data

We used two traces of (the first 5000 and the first10000 processes of AuverGrid) in our experiment for evaluation of the Grid scheduling algorithms. Details of the input simulation data is given in Table 2.

Table 2. Simulation Data

Trace id	Data Type	Total number of Jobs	Number of CPUs
1	Real workload trace	5000 processes of AuverGrid	{16, 32, 64}
2	Real workload trace	10000 processes of AuverGrid	{16, 32, 64}

9 Performance Analysis of Grid Scheduling Algorithms

Performance metrics for the Grid scheduling algorithms are based on three factors - Average Waiting Time, Average Turnaround Time, and Average Response Time. We performed experiments for different scheduling algorithms using 'AuverGrid'. Our experiments include the scalability test of scheduling algorithms under an increasing real workload. The 'runtime' attribute is given for each process in 'AuverGrid'. 'Runtime' is taken to mean CPU time in our experiment. We used 50 units as the fixed time quantum for our experiment. In this section, we describe a comparative performance analysis of our proposed algorithms, i.e. MDQM and MDQR, with seven other Grid scheduling algorithms, i.e. FCFS, SJF, SRTF, RR, MH, MDQ and SARR.

9.1 Average Waiting Times Analysis

Figure 8 shows that the average waiting times computed by each scheduling algorithm for each real workload trace. Figure 8 illustrates that the SRTF, MH and MDQM scheduling algorithms produce the least average waiting times as compared to the other scheduling algorithms. The average waiting time computed for SRTF is slightly less than the value computed for the MH and MDQR scheduling algorithms. By increasing the number of CPUs, each algorithm shows the relative improvement in performance. MDQM and MDQR show the improved performance in average waiting times as compared to MDQ. MDQM shows better results as compared to the MDQ, MDQR, RR and FCFS algorithms. All scheduling algorithms show that the relative performance is independent of the nature of the workload, the workload size and the number of CPUs used for computation.

9.2 Average Turnaround Times Analysis

Figure 9 presents the pictorial view of the average turnaround times computed for the scheduling algorithms using real workload traces. Figure 9 illustrates that the average turnaround time computed by the SRTF, MH and MDQM scheduling algorithms are less than the other Grid scheduling algorithms. By increasing the number of CPUs, each algorithm has an improved average turnaround time.

Experimental results show that SRTF and MH are at the same performance level as regards average turnaround time. Figure 9 also shows that the average turnaround times computed for MDQM are slightly longer than those for the MH and SRTF scheduling algorithms but better than the values computed for the MDQ, MDQR, RR and FCFS scheduling algorithms. Moreover, all scheduling algorithms show that relative performance is independent of the nature of the workload, the workload size and the number of CPUs used in the experiment.

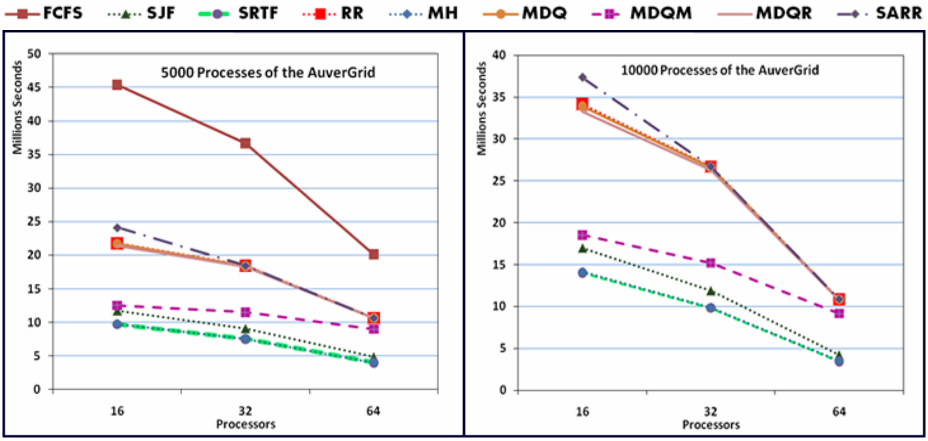


Fig. 8. Average Waiting Time Analysis for 5000 and 10000 Processes of AuverGrid

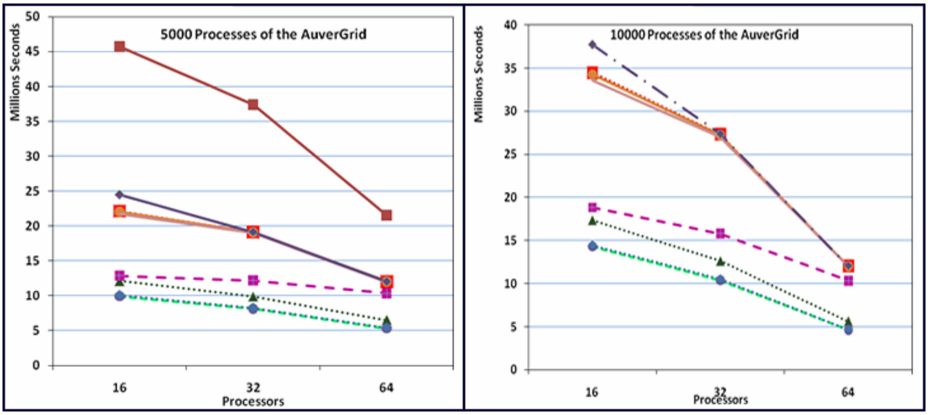


Fig. 9. Average Turnaround Time Analysis for 5000 and 10000 Processes of AuverGrid

9.3 Average Response Times Analysis

Figure 10 shows that MDQ and MDQR produce the least average response times as compared to the other scheduling algorithms. The average response times computed for MH are slightly longer than those for MDQR and slightly less than those for RR. The FCFS, SJF and SRTF scheduling algorithms result in poor response times as compared to the other scheduling algorithms. All scheduling algorithms show that relative performance is independent of the nature of the workload, the workload size

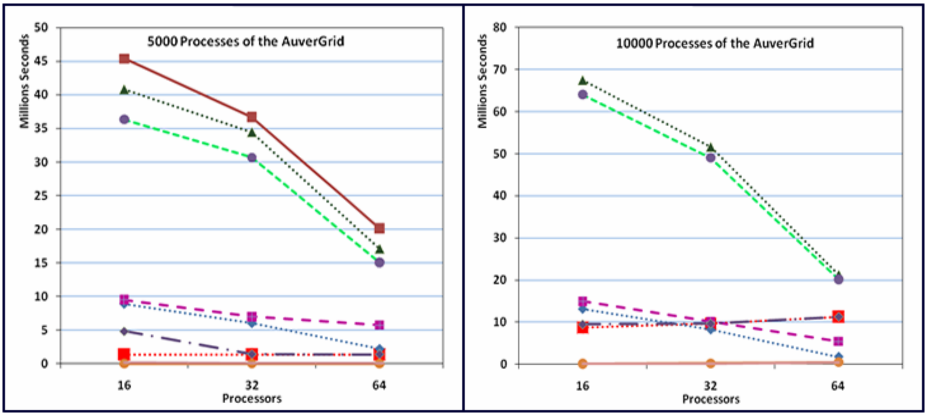


Fig. 10. Average Response Time Analysis for 5000 and 10000 Processes of AuverGrid

and the number of CPUs. MDQR gives consistently good results for different workloads and numbers of CPUs.

It is apparent from the charts that our proposed scheduling algorithms namely MH, MDQM and MDQR clearly outperform other Grid scheduling algorithms and achieve significant improvement for all performance parameters.

Table 3 shows the average performance measure for each algorithm, running 10000 processes of AuverGrid on an SGI Super Computer using 64 CPUs.

Table 3. Performance Analysis of Scheduling Algorithms for 10000 Processes of AuverGrid on SGI Super Computer using 64 Processors

Scheduling Algorithm	Average Waiting Times (seconds)	Average Turnaround Times (seconds)	Average Response Times (seconds)
FCFS	24972466.3090	26160492.2507	24972466.3090
SJF	4195654.4856	5621285.6156	21226596.3627
SRTF	3465610.6051	4637560.6329	20165266.5445
RR	10848690.6347	12036716.5764	11292464.6064
MH	3496378.7380	4684404.6797	1752911.2333
MDQ	10833965.7698	12021991.7115	477993.0796
MDQM	9154961.9476	10342987.8893	5417198.0584
MDQR	10809824.7974	11997850.7391	482089.1758
SARR	10858779.2695	12046805.2112	11306950.7576

9.4 Performance Analysis of Scheduling Algorithms by Changing Time Quantum

The RR, MH and MDQ scheduling algorithms work on a fixed time quantum value. If the value of the time quantum is very small, then these scheduling algorithms result in better average response times but produce many context switches. If the value of the time quantum is very high, then their efficiency is also degraded. Our proposed

MDQM and MDQR algorithms use a dynamic time quantum strategy instead of a static one and maintain system performance. The value of the time quantum is computed at runtime by considering the size of each job and the total number of jobs in the system.

In the experiment we computed results for each algorithm using a workload of 10000 processes on 64 processors and varying the time quantum from 50 to 5000 as shown in Figure 11.

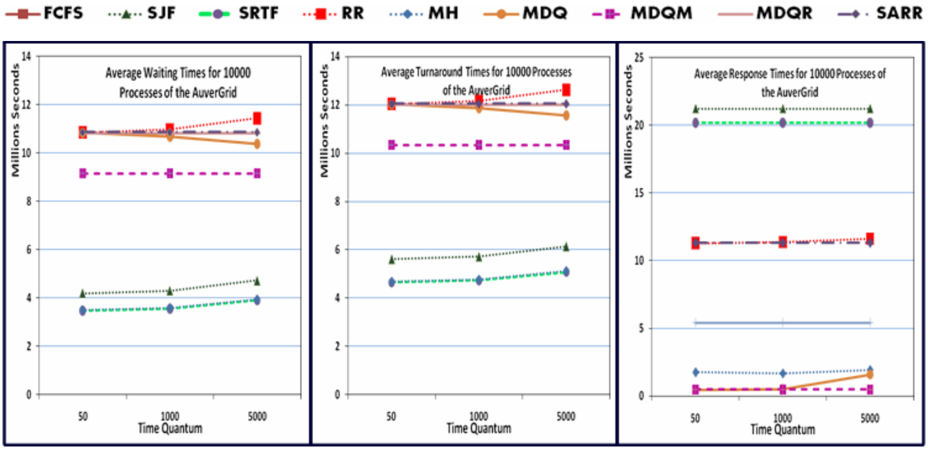


Fig. 11. Average Performance Analysis of Scheduling algorithms by changing the Time Quantum

Figure 11 shows that all scheduling algorithms show stable performance measures (average waiting time, average turnaround time and response time) on varying the time quantum, except the RR, MH and MDQ scheduling algorithms. It is apparent from the charts that our proposed scheduling algorithms (MDQM and MDQR) markedly outperform the other Grid scheduling algorithms. A significant improvement is achieved in all of the performance parameters.

The efficiency of the proposed Dynamic Multilevel Dual Queue Scheduling algorithms (MDQR and MDQM) doesn't degrade in this experiment because they work on a dynamic time quantum strategy. The efficiency of MH and MDQ are affected by the static time quantum value while MDQR and MDQM show stable performance. MDQ and MDQR have resulted in shorter response times as compared to all other Grid scheduling algorithms for AuverGrid workload traces.

In concluding the experimental remarks, MDQR and MDQM are better choices for fair and effective resource scheduling in a Grid environment. The relative performance of MDQR and MDQM are independent of the workload size and show good performance and support scalability.

10 Conclusion

In this paper we present two new dynamic multilevel scheduling algorithms, i.e. MDQM and MDQR. Our proposed algorithms compute the time quantum dynamically and execute the processes accordingly. We have evaluated these algorithms on a simulator running on a Super Computer using a wide range of CPUs. We compared the efficiency of our algorithms with seven other Grid scheduling algorithms using 'AuverGrid' workload traces. In this paper we also performed a statistical analysis of 'AuverGrid' workload traces to study the dynamic nature of jobs.

Experimental results show that the MH and SRTF scheduling algorithms are at the same performance level in producing the least average waiting times and average turnaround times for a variety of workloads. Experimental results also exhibit that MDQR produces the least average response times in comparison to all other scheduling algorithms. MDQM and MDQR show better performance than the other scheduling algorithms because their performance is not affected by the value of fixed time quantum.

We can say that MH and MDQM are scheduling policies from the system point of view; they satisfy the system requirements (i.e. less Average Waiting Time and less Turnaround Time). MDQ and MDQR work well from the user perspective due to their shorter Average Response Time). Moreover, MH, MDQR, MDQM and MDQ are scalable, i.e. the relationship between each performance measure (e.g. average waiting time) and the workload size is very nearly linear.

Acknowledgements

We are grateful to the HPC Service Centre at University Teknologi PETRONAS for providing the HPC facility to perform the experiments. We want to express our gratitude to Dr. M. Fadzil Hassan, Dr. M. Nordin B Zakaria, Dani Adhipta and Thayalan Sandran from Universiti Teknologi PETRONAS and Dr Aamir Shafi from Massachusetts Institute of Technology (MIT) for their help during the research. We are also thankful to the University of Reading (UK) for providing the MPJ Express API for our scheduling experiment on the HPC.

We thank Dr. Emmanuel Medernach and his team from the Université Blaise-Pascal (France) who provided the AuverGrid data. We also thank Hui Li, of the Parallel Workload Archive, and the Grid Workloads Archive for their contribution in making the data publicly available.

References

1. Foster, I., Kesselman, C.: The Grid 2: Blueprint for a New Computing Infrastructure, 2nd edn. Morgan-Kaufmann Publishers, San Francisco (2003)
2. Jang, S.H., Lee, J.-S.: Predictive Grid Process Scheduling Model in Computational Grid. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 525–533. Springer, Heidelberg (2006)
3. Grid Scheduling Use Cases, `http://www.ogf.org/documents/GFD.64.pdf`

4. Dong, F., Akl, S.G.: Scheduling Algorithms for Grid Computing: State of the Art and Open Problems, Technical Report No. 2006-504, Queens University, Canada (2006)
5. Haines, S.: Pro Java EE 5 Performance Management and Optimization. Apress (2006)
6. Li, H., Buyya, R.: Model-Driven Simulation of Grid Scheduling Strategies. In: Third IEEE International Conference on E-Science and Grid Computing (2007)
7. Shah, S.N.M., Mahmood, A.K.B., Oxley, A.: Development and Performance Analysis of Grid Scheduling Algorithms. Communications in Computer and Information Science 55, 170–181 (2009)
8. Lu, L., Yang, S.: DIRSS-G: An Intelligent Resource Scheduling System for Grid Environment Based on Dynamic Pricing. International Journal of Information Technology 12(4), 120–127 (2006)
9. Huang, P., Peng, H., Lin, P., Li, X.: Static Strategy and Dynamic Adjustment: An Effective Method for Grid Task Scheduling. Journal of Future Generation Computer Systems 25(8), 392–884 (2009)
10. Krauter, K., Buyya, R., Maheswaran, M.: A taxonomy and survey of grid resource management systems for distributed computing. Software Practice and Experience 32(2), 135–164 (2002)
11. Buyya, R., Abramson, D., Giddy, J.: Nimrod/G: an architecture for a resource management and scheduling system in a global computational grid. In: Proceedings, High Performance Computing in the Asia-Pacific Region, vol. 1, pp. 283–289 (2000)
12. Chunlin, L., Xiu, Z.J., Layuan, L.: Resource Scheduling with Conflicting Objectives in Grid Environments: Model and Evaluation. Journal of Network and Computer Applications 32(3), 760–769 (2009)
13. Shmueli, E., Feitelson, D.G.: Backfilling with look ahead to optimize the packing of parallel jobs. Journal of Parallel and Distributed Computing 65(9), 1090–1107 (2005)
14. Lawson, B., Smirni, E., Puiu, D.: Self-adaptive backfill scheduling for parallel systems. In: Proceedings of the International Conference on Parallel Processing (ICPP 2002), pp. 583–592 (2002)
15. Tsafrir, D., Etsion, Y., Feitelson, D.G.: Backfilling using system-generated predictions rather than user runtime estimates. IEEE Transactions on Parallel and Distributed Systems 18(6), 789–803 (2007)
16. Abawajy, J.H.: Job Scheduling Policy for High Throughput Grid Computing. In: Hobbs, M., Goscinski, A.M., Zhou, W. (eds.) ICA3PP 2005. LNCS, vol. 3719, pp. 184–192. Springer, Heidelberg (2005)
17. Mu'alem, A.W., Feitelson, D.G.: Utilization, Predictability, Workloads, and User Runtime Estimates in Scheduling the IBM SP2 with Backfilling. IEEE Transactions on Parallel and Distributed Systems 12(16), 529–543 (2001)
18. Rawat, S.S., Rajamani, L.: Experiments with CPU Scheduling Algorithm on a Computational Grid. In: IEEE International Advance Computing Conference, IACC 2009 (2009)
19. Sharma, R., Soni, V.K., Mishra, M.K.: An Improved Resource Scheduling Approach Using Job Grouping strategy in Grid Computing. In: 2010 International Conference on Educational and Network Technology (2010)
20. Laurence, T., Yang, M.G.: High-Performance Computing: Paradigm and Infrastructure. Wiley, Chichester (2005), ISBN: 978-0-471-65471-1
21. Matarneh, R.J.: Self-Adjustment Time Quantum in Round Robin Algorithm Depending on Burst Time of the Now Running Processes. American Journal of Applied Sciences 6(10), 1831–1837 (2009)
22. Tchernykh, A., Trystram, D., Brizuela, C., Scherson, I.: Idle regulation in non-clairvoyant scheduling of parallel jobs. Discrete Applied Mathematics 157(2), 364–376 (2009)

23. Feitelson, D.G.: Metric and workload effects on computer systems evaluation. IEEE Computer 36(9), 18–25 (2003)
24. Shah, S.N.M., Mahmood, A.K.B., Oxley, A.: Hybrid Scheduling and Dual Queue Scheduling. In: 2009 the 2nd IEEE International Conference on Computer Science and Information Technology, IEEE ICCSIT 2009 (2009)
25. Shah, S.N.M., Mahmood, A.K.B., Oxley, A.: Analysis and Evaluation of Grid Scheduling Algorithms using Real Workload Traces. In: The International ACM Conference on Management of Emergent Digital EcoSystems, MEDES 2010 (2010)
26. Li, H.: Workload dynamics on clusters and grids. The Journal of Supercomputing 47(1), 1–20 (2009)
27. Trace analysis report, `http://gwa.ewi.tudelft.nl/pmwiki/reports/gwa-t-4/trace_analysis_report.html`

Communications in Computer and Information Science: Performance Improvement and Interference Reduction through Complex Task Partitioning in a Self-organized Robotic Swarm

Mehrdad Jangjou[1], Alireza Bagheri[2], Mohammad Mansour Riahi Kashani[3],
and Koosha Sadeghi Oskooyee[3]

[1] Department of computer engineering,
Islamic Azad University, U.A.E. Branch, Dubai, Emirates
Mehrdadjangjou@live.com
[2] Department of computer engineering and IT,
Amirkabir University of Technology, Tehran, Iran
ar_bagheri@aut.ac.ir,
[3] Department of computer engineering,
Islamic Azad University, North Tehran Branch, Tehran, Iran
{M_Riahi_Kashani,K_Sadeghi_Oskooyee}@iau-tnb.ac.ir

Abstract. Interference has long been accepted as one of the critical problems in multi-robot co-operation. One of the most common kinds of interference is physical interference. A simple way of reducing this interference is to make robots remain in unique work areas and move the objects to the next robot as soon as they cross the borders of their areas. In this article, the problem of interference reduction is investigated through the complex task partitioning in self-organized robotic swarms. The presented method was simulated and the results show an improvement in the cost of operations.

Keywords: Interference, Self-organized task allocation, Robotic swarm, Foraging problem, Complex task.

1 Introduction

Interference is an important problem limiting the development of a swarm in the collective robotic science; when each robot performs the mere task in irrelevant behaviors, as the density of people increases, prevention of barrier increases too [1]. Operation of the task which is in trouble because of physical interferences can usually be improved by spatial (environmental) partitioning; for instance, by keeping each robot in its own work area [2].

The problem of foraging (or harvesting) the objects by a swarm of robots has been one of the conventional research areas for collective robots, which is because this problem can be easily modeled, has some variants in the nature and can be applied in a lot of scientific situations. The problem of foraging can provide a useful and effective system of measurement [3], [4].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 441–447, 2011.

In simple foraging, a swarm of robots has to collect objects of a single kind and they are usually stimulated by an internal motive or stimulus, such as artificial hunger, energy balance and so on [5].

There are considerably few studies on the self-organized task allocation. This area is at its initial stages because most of the studies try to solve simple problems without tasks being dependent on each other. The studies on the self-organized task allocation are mostly based on edge-based methods and are inspired by the division of labor in social insects [6]. There is one central robot in self-organized systems, which decides autonomously on the time to allocate a task to a robot.

Task partitioning is not a famous concept although it can be found in most insect societies. It is sometimes defined as a division of a single task among workers [7] and it is called so because several individuals divide a big task among themselves; it primarily allows tasks to be allocated not only to individuals but also to swarms. This concept is elaborated in the present paper as follows: Task partitioning is defined as the problem of dividing a general task to smaller (atomic) sub-tasks which can be solved by an individual or a group of individuals. The purpose is to find ways for the implementation of task partitioning (at least, minimally) in an automatic and self-organized manner.

Division of labor explains division of workforce for a large variety of tasks with which a swarm is confronted. It is essential for task processing and parallel functioning and it is a basis for training specialized individuals. Because of this specialization, division of labor may increase efficacy and facilitate specialist training for those who are not able to do the tasks and functions which are different from the ones in which they have been specialized. Therefore, division of labor may result in heterogeneous populations with respect to behaviors.

2 Interference in Multi-Robot Systems

Interference has long been accepted as one of the critical problems in multi-robot co-operation [8], [9]. A mathematical model has been created which makes it possible to determine the amount of interference and its effect on the efficacy of the swarm. Goldberg pointed out that one of the commonest types of interference is physical interference. A simple way to reduce this kind of interference is to make robots remain in unique work areas and move objects to the next robot as soon as they cross the borders of their areas [10].

2.1 The Recommended Method

In order to solve a complex task with the sequential dependence, foraging problem is used which is one of the conventional areas of research in the science of collective robots [11]. In this article, for implementing the problem, a swarm of robots should pick up target objects from the source area and move them to the nest area. With spatial partitioning of the environment, the general foraging task is divided into two sub-tasks:

1. Harvest target objects from an area (known as source).

2. Store them in an area (known as nest).

The robots which work on the first sub-task pick up target objects from the source and move them to the robots which work on the second sub-task and store the objects in the nest. These sub-tasks are sequentially dependent on each other in the sense that they should be done one after the other in order to immediately complete the general task: delivering a target object to the nest area.

2.2 Interference Reduction through Sequential Task Partitioning

Interference is an important problem limiting the development of a swarm in the science of collective robots: when each robot performs the task with irrelevant behaviors, as the density of people increases, prevention from barrier increases too [1]. Performance of the task which is in trouble because of physical interferences can be usually improved by spatial (environmental) partitioning; for example, by keeping each robot in its own work area.

The method which is presented here is in a way that robots deliver the objects to other robots which work in the next area in order to move the objects to their destinations. This method effectively limits the number of robots which have to be applied in the task. Robots apply a simple, absolute and edge-based model in order to decide on the time for changing the status of a task. When the time, t_w, is over an edge, a robot changes its sub-task. This strategy was compared with the strategy which did not include task partitioning and the way it helped in interference reduction was analyzed. In this study, these two strategies are called divided and undivided sequences.

3 Operation Environment

The sections to which the environment is divided are those which include the source and are situated on the left and those which include the nest and are situated on the right area. These two sides of the area are referred to as pick-up area and storage area, respectively. The exchange area is situated between these two areas. The robots working on the left are called harvesters and collect the target objects in the source and move them to the exchange area. In this area, objects are delivers to the robots working on the other side, referred to as storers, whose role is to move target objects to the nest and to store them over there [12], [13].

A graphic representation of the undivided strategy in problem operation is presented in Fig. 1 at time $t = 0$. All the robots are situated in the harvest area. After picking up each object, each robot enters the exchange area and, without hesitation (i.e. the value of edge is considered to be zero) enters the store area and puts the object in the nest.

Graphic representation of the divided strategy in the problem operation is presented in Fig. 2. After picking up each object, each robot enters the exchange area and waits there for the arrival of the other robot from the store area in 3 seconds in order to deliver the object to that robot; otherwise, the same robot enters the store area and performs the task of storer.

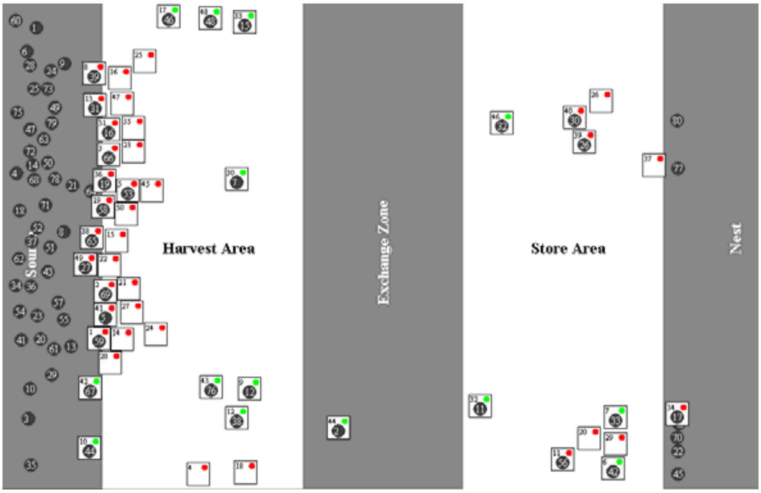


Fig. 1. A graphic representation of the undivided strategy

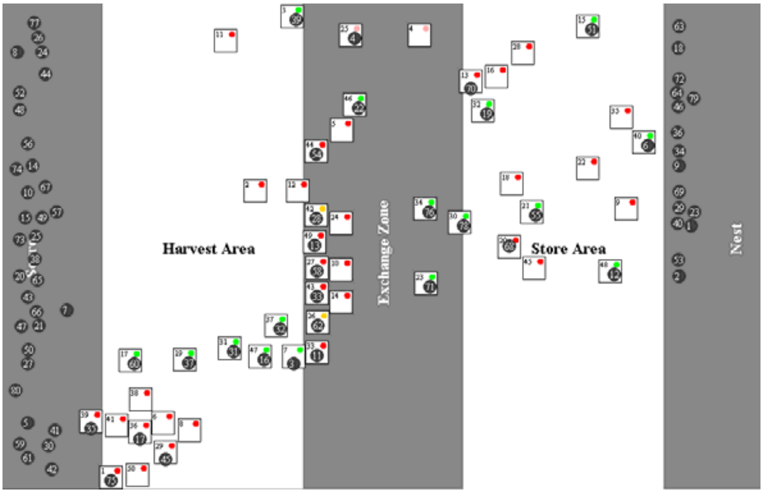


Fig. 2. A graphic representation of the divided strategy

As shown in Fig. 3, the experiment is performed by 10, 20, 30, 40 and 50 robots, respectively, with the threshold of zero. Because of the undivided complex tasks of the robots, increasing the number of robots raised the physical interference among robots and, as a result, the performance of the system decreased. As shown in Fig. 4, in this experiment, the performance of the system is measured by 10, 20, 30, 40 and 50 robots. Increase in the rate of inference among robots and lack of specialization decrease the performance of the system.

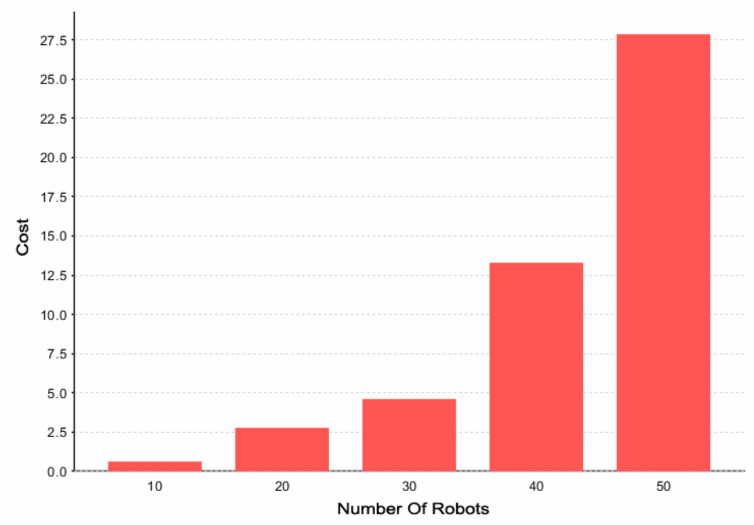


Fig. 3. A graphic representation of attained cost in the undivided strategy for various numbers of robots

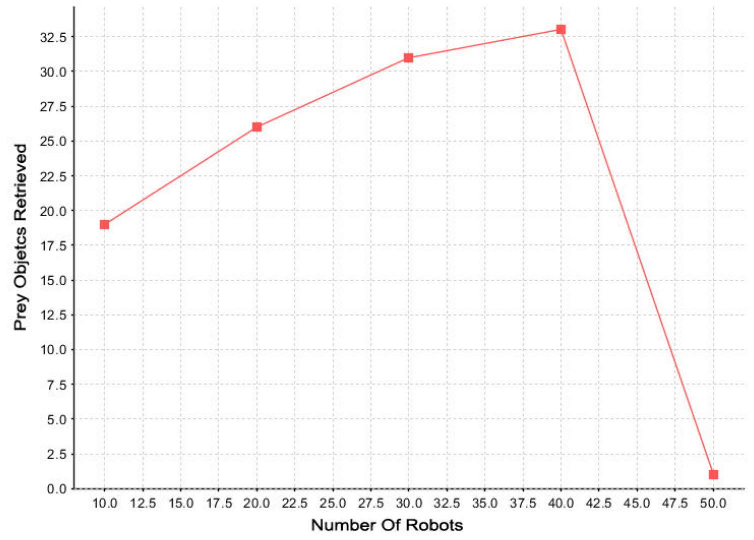


Fig. 4. A graphic representation of the undivided strategy for various numbers of robots

As can be seen in Fig. 5, the experiment is performed by 10, 20, 30, 40 and 50 robots with the threshold of 3 (i.e., each robot remained in the exchange zone for 3 seconds until the robot on the other side reached the zone to deliver the prey and take it to the nest). Unlike the undivided strategy, the cost did not increase exponentially in the undivided strategy. Moreover, decreasing the inferences in the divided strategy increased the performance.

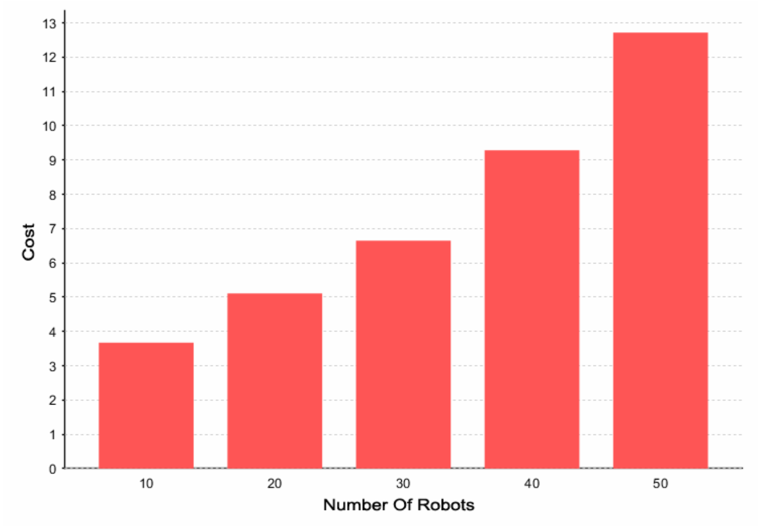


Fig. 5. A graphic representation of attained cost in the divided strategy for various numbers of robots

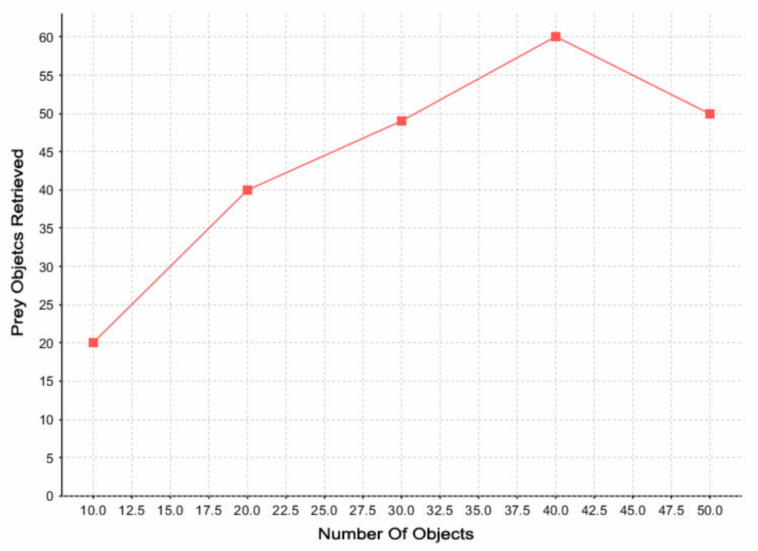


Fig. 6. A graphic representation of the divided strategy for various numbers of robots

In this experiment (Fig. 6), the performance of the system with 10, 20, 30, 40 and 50 robots was measured. As can be seen in Fig. 6, dividing the general task into two subtasks of harvester and storer increases the performance of the system. In this case, each robot was specialized in its subtask. Subsequently, the performance of the divided strategy increased in comparison with that of the undivided one.

4 Conclusion

The purpose of this article was to investigate whether task partitioning can reduce interference in critical task areas and examine the ways to allocate a robotic swarm to divisions. Interference was related to the number of individuals in the system. In addition, physical interference among robots was a function of the environment in which the robots worked.

The larger the size of the swarm, the higher was the density and rate of physical interference. According to the conducted experiment, it can be concluded that the amount of cost, which in reality is the physical interference or contacts of robots with each other, increases in the undivided strategy, which is in contrast with the divided one. Consequently, efficacy and the amount of performed work decreases.

References

1. Lerman, K., Jones, C., Galstyan, A., Mataric, M.J.: Analysis of Dynamic Task Allocation in Multi-robot Systems. International Journal of Robotics Research 25(3), 225–241 (2006)
2. Anderson, C., McShea, D.: Individual versus Social Complexity, with Particular Reference to Ant Colonies. Biological Reviews 76, 211–237 (2001)
3. Dall, S., Giraldeau, L.-A., Olsson, O., McNamara, J., Stephens, D.: Information and its Use by Animals in Evolutionary Ecology. Trends in Ecology and Evolution 20, 187–193 (2005)
4. Hirsh, A., Gordon, D.: Distributed Problem Solving in Social Insects. Annals of Mathematics and Artificial Intelligence 31, 199–221 (2001)
5. Mataric, M.J.: Learning Social Behavior. Robotics and Autonomous Systems 20(2), 191–204 (1997)
6. Krieger, M.J.B., Billeter, J.-B.: The Call of Duty: Self-organised Task Allocation in a Population of up to Twelve Mobile Robots. Journal of Robotics and Autonomous Systems 30, 65–84 (2000)
7. Ratnieks, F.L.W., Anderson, C.: Task Partitioning in Insect Societies. Insectes Sociaux 46, 95–108 (1999)
8. Goldberg, D., Mataric, M.J.: Maximizing Reward in a Non-stationary Mobile Robot Environment. Autonomous Agents and Multi-Agent Systems 6(3), 287–316 (2003)
9. Lerman, K., Galstyan, A.: Mathematical Model of Foraging in a Group of Robots: Effect of Interference. Auton. Robots 13(2), 127–141 (2002)
10. Mataric, M.J.: Learning Social Behavior. Robotics and Autonomous Systems 20(2), 191–204 (1997)
11. Labella, T.H., Dorigo, M., Deneubourg, J.-L.: Division of Labor in a Group of Robots Inspired by Ants' Foraging Behavior. ACM Transactions on Autonomous and Adaptive Systems 1(1), 4–25 (2006)
12. Groß, R., Dorigo, M.: Evolution of solitary and group transport behaviors for autonomous robots capable of self-assembling. Adaptive Behavior 16(5), 285–305 (2008)
13. Garnier, S., Gautrais, J., Theraulaz, G.: The biological principles of swarm intelligence. Swarm Intelligence 1, 3–31 (2007)

A Semantic Service Discovery Framework for Intergrid

Mahamat Issa Hassan and Azween Abdullah

Department of Computer and Information Sciences, Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, 31750 Tronoh, Peark, Malaysia
{mahamat.hassan,azweenabdullah}@petronas.com.my

Abstract. Resource/service discovery is a very critical issue in intergrid (grid of grids) system, which is the recent advancement in grid technology that uses multi-middleware. In this paper we present an intergrid service discovery framework that integrates semantic technology, peer-to-peer network and intelligent agents. The framework has two main components which are service description, and service registration and discovery models. The earlier consists of a set of ontologies that are used as a data model for service description and services to accomplish the description process. The service registration is also based on ontology, where nodes of the services (service providers) are classified to some classes according to the ontology concepts, which means each class represents a concept in the ontology. Each class will have an elected head that plays the role of a registry in its class and handles the interclass communication. We further introduce intelligent agents to automate the discovery process. The framework is evaluated via simulation experiments, and the result of which confirms the effectiveness of the framework in satisfying the required RD features (interoperability, scalability, decentralization and dynamism).

Keywords: Grid, semantic technology, intelligent agent, and peer-to-peer network.

1 Introduction

Intergrid/Global grids (in some literatures is also called multi-grid [1]) are known as a grid of grids, since they are a collection of small grids that cross organizational boundaries to create very large virtual systems that can be accessed from anywhere in the world. A very basic and first step in sharing resources over intergrid is the detection of suitable resource for a given task/application which is commonly known as *Resource Discovery* (RD). The RD process entails *description of the resource through its properties*, *registration/indexing of the described resource in common registry(s)*, and *discovering the registered resources that match with resource request specifications*. These steps correspond to the main components of the RD system, which are *Description*, *Registration* and *Discovery (*which is composed of *search* and *selection)*. RD is very important as resource reservation and task scheduling are based on it. Unfortunately, intergrids are normally associated with some complexities such as resources/services and users are distributed across different locations; resources are

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 448–462, 2011.

heterogeneous in their platforms; status of the resources is dynamic (resources can join or leave the system without any prior notice); and use of multi-middleware. These complexities pose a challenge to the development of an efficient RD system to discover the resources and services. In fact, these complexities also yield some requirements that should be fulfilled by any developed RD. These requirements include *high searchability* (interoperability) to retrieve the relevant and precise resources and services, and *high performance (scalability, decentralization, and dynamism)* to make the RD system sustainable with the scale of the intergrid.

Currently, there is a wealth of work on grid RD (e.g. Globus[1], Condor[2] , [2], [3], and [4]) which can be classified into two classes based on the description component, which are *keyword-based* RD systems and *semantic-based* RD systems. Keyword-based system uses syntactic information and data models such as directories [5] and special databases to describe and discover the resources and services. Unfortunately, syntactic information and data models are not efficient in describing resources at intergrid level. This is because resources and services are initially described by using multi information services that belong to different grid middlewares. As a matter of fact, much of the efforts in keyword-based RD systems have been focused on achieving the *high performance* requirement; staring from introducing centralized registration models such as Globus MDS-1 [6], R-GMA[3] [7] and Hawakeye [8]; then followed by hierarchical registration models [9], [10] and [11], and lastly peer-to-peer (P2P) registration models [12], [13], [4] and [14]. Keyword-based RD systems that are based on P2P registration models have achieved high performance compared to the centralized and hierarchical models, but we cannot go far as to say that they have achieved full scalability. Moreover, their use of syntactic description, especially at the intergrid level, prevents them from fulfilling the *high searchability* requirement. Semantic-based RD systems, on the other hand, use semantic information and data models (ontology and ontology languages) [15] to describe and discover the resources and services. Although, there is a considerable amount of work on semantic-based RD systems (e.g. [16], [17]), most of the existing approaches fail to achieve *high searchability.* This is due to the lack of a proper use of semantic description mechanism as the semantic technology is initially imported from the semantic web [18]. Actually, we have argued in an earlier study [19] that the main obstacle that leads to the continuous existence of this issue is the ad hoc research nature of these semantic-based RD studies (different research communities doing the same thing by different ways).

In this paper, we introduce a new intergrid RD framework that can overcome the shortcoming of the current studies and meeting the above mentioned requirements. The framework contains two main components which are *service description,* and *service registration and discovery* models. The earlier consists of a set of ontologies and services. Ontologies are used as a data model for service description, whereas the services are to accomplish the description process, we detail that in section 2.

The service registration is also based on ontology, where nodes of the services (service providers) are classified to some classes according to the ontology concepts, which means each class represents a concept in the ontology. Each class has an

[1] `http://www.globus.org/.`

[2] `http://www.cs.wisc.edu/condor/.`

[3] Relational Grid Monitoring Architecture: `http://www.r-gma.org/index.html`

elected head. Head plays the role of a registry in its class and communicates with the other heads of the classes in a peer to peer manner during the discovery process. We further introduce two intelligent agents to automate the discovery process which are *Request Agent* (RA) and *Description Agent* (DA). Each node is supposed to have both agents. DA describes the service capabilities based on the ontology, and RA carries the service requests based on the ontology as well. We design a service search algorithm for the RA that starts the service look up from the class of request origin first, then to the other classes, we detail that in section 3.

We finally evaluate the performance the framework with extensive simulation experiments, the result of which confirms the framework effectiveness in satisfying the required RD features (interoperability, scalability, decentralization and dynamism), we detail that in section 4.

In short, our main contributions are an interoperable semantic description RD component model for intergrid services metadata representation; a semantic distributed registry architecture for indexing service metadata; and an agent-based service search and selection algorithm.

2 Semantic-Based Resource Description Model

2.1 The Model Description

In order to have a description model that meets RD requirements, we refine the intergrid system in such a way that makes full use of the resources and services when the semantic technology is applied. A common ontology is used to formally represent the intergrid components. Subsequently, we merely rely on the latest grid system requirements that have been presented by the OGF[4] [20] in defining the grid level and intergrid level. As a result, we treat grid level system as a *service grid* that is provided by a *provider* to *consumers,* and this service grid is assumed to be among the grid types (e.g. computing, data, and application). Therefore, an intergrid level system is a collection of service grids that have agreed to work cooperatively as consumers and providers. Consequently, a service grid may be a consumer as well as a provider. It becomes a consumer when it uses other service grids without providing any service to them, whereas it functions as a consumer/provider when other services are added to its own and at the same time it also provides a complete service to the end user. The capabilities of service grids are described by aggregating their local metadata content and then integrating them into a common information model. Ontology ([21] and [15]) is used for the common information model. We call this ontology as *service grid domain ontology (SGDO).* SGDO defines all the service grid types, attributes that are needed for each service grid, relationships between all the services, structure of the values of each attributes and so on (see Fig. 1). To reduce the user interaction with programming details with RD system in specifying the service grid requests, we introduce a mechanism that is called Goal-based Service Grid Request Description (GSGR).

[4] Open Grid Forum: `http://www.ogf.org/`

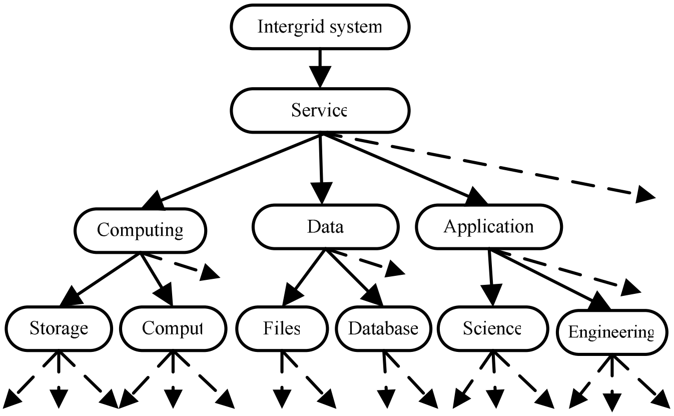


Fig. 1. Fragment of service grid domain ontology

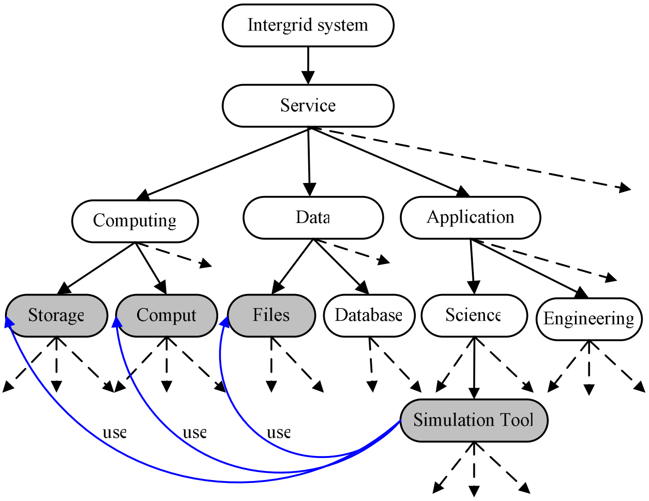


Fig. 2. The extraction of application goals from the service grid domain ontology

A goal refers to what a given consumer/end user wants to achieve by using the service grids. For example, if a user wants to simulate the weather condition of the earth so the simulation is his/her goal. Obviously, a goal requires a set of the grid services in order to be accomplished. For example, the simulation of weather condition of the earth requires computing service grid, satellite images data and temperature dataset which can be under the data service grid and so on. The SGDO, among other concepts of application service grid, includes all software applications that are available on the intergrid level system. In fact, these applications represent the goals that a user may want to achieve because application service grids are the only services that need one or more service grids to work on, as they cannot stand alone.

Thus, we can extract the goals from the SGDO. We introduce a relation so called "*use*" between the application service concepts and the other service grid concepts. The "use" relation is a binary relation between a particular application service concept and another service grid (e.g. data service). Indicating this application service requires the second service grid with which the relation is established (see Fig. 2).

2.2 The Model Building Block

Having described the fundamental components of the description model, in this section we illustrate the model building block. Fig. 3 shows the components of the model and their related subcomponents. The model is initially composed of Semantic Description Manager (SDM) and Service Grid Metadata Provider (SGMP). SDM generally is responsible of the global service grid description in the intergrid system, and a pool that accommodate the service grid metadata coming from SGMPs. Meanwhile, SGMP is responsible of managing local service grid metadata that belongs to a service grid provider.

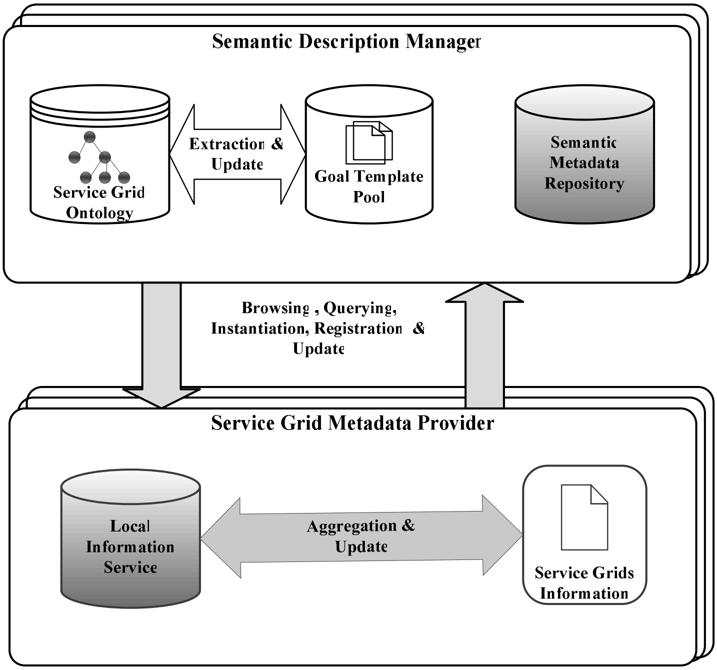

Semantic Description Manager
Service Grid Ontology
Extraction & Update
Goal Template Pool
Semantic Metadata Repository
Browsing , Querying, Instantiation, Registration & Update
Service Grid Metadata Provider
Local Information Service
Aggregation & Update
Service Grids Information

Fig. 3. The description of model building block

The reason for having SDM and SGMP in such architecture is that, SDM will provide all the needed information and data model management for a set of intergrid members. Therefore, interoperability can be ensured. In the meantime, SGMP provides autonomy to each service grid member as it handles the local information of the service grid.

2.3 The Description Process

The description process includes description of a service that will be advertised, and service request formulation. The steps of the first case are as follows:

- The user invokes the SGMP system.
- The user browses/queries the service grid ontology through which all the available content of the ontology can be manipulated (e.g. using add and drag menu), and selects the concept that is relevant to his/her service grids.
- The user gets an instance of the selected service grid concept using the service grid ontology tools.
- The user populates the instance with the actual service grid information, which is an aggregated summary of the overall service grid information.
- Finally, the user sends the service grids information to the respective SMR of the SDM node that is responsible of holding the metadata of the current service provider, and in turn, the SMR stores the semantic information about the service grid for the discovery process.

Meanwhile for the service request formulation is as follows:

- The user invokes the SDM system.
- The user browses/queries the available goals in the goal template manager, and selects the relevant goal.
- The user then gets an instance of the selected goal and adds it to his/her local information system.
- Since each goal requires one or a set of service grids, the user adds the concrete values of attributes of each service. For example, if among the required service is computing service, and one of the attributes is maximum number of computing nodes, the user may add a concrete number for such requirements.
- Finally, the user sends the service grid request to the respective SDM, and the SMR of that SDM will generate a proper query statement for each service among the required service grids.

2.4 Model Evaluation

From the building block, it is clear that the model has introduced the use of semantic information in way that does not require the use of local information service, which exists currently in grid middleware. For example, the intergrid participants (small grids) are able to use their local discovery system that would normally be possible through a keyword-based RD system. They just need to have one file that accommodates the summary of the overall capabilities of the service. As a result, this provides interoperability among the participants in the intergrid system. The model also reduces the cost of using semantic information in terms of processing time, as well as storage of the semantic information, since the semantic information is used at the intergrid level, and not at the grid level. Therefore, during the discovery we look up for a complete service grid, not components of it. Therefore, the model achieved the first RD requirement (high searchability).

3 Semantic Registration and Discovery Model

Registration and discovery components in any RD system are very much related, as the routing of request is subjected to the registration architecture. For this reason, we address the issues in registration and discovery jointly. We design a model for the two components that integrates super-peer architecture, ontology and intelligent agent. Super-peer is used to grant distribution of the registry where the service grid metadata is located.

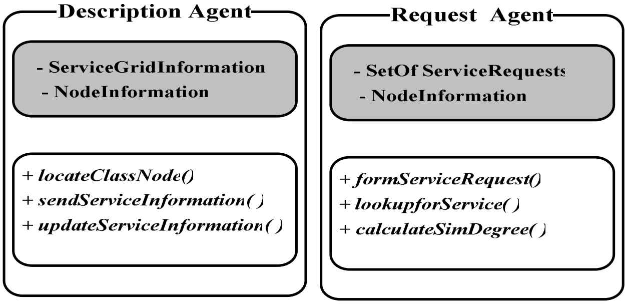


Fig. 4. The proposed DA and RA agents

3.1 The Model Components

The model consists of three components, domain ontology, an intelligent agent model, and super-peer architecture. Ontology concepts are normally arranged hierarchically, therefore, whenever, we visit the concepts from the root concept, as we go down deeper into the subconcepts, we will move from a more general class of concepts to a more specific class of concepts, and vice versa. We use this feature to classify nodes into several classes, which produce registry architecture to the RD system. The ontology that supplies service grid taxonomies is called Dictionary Ontology (DO). The DO may be the same as the service SGDO by omitting the relations that are out of the hierarchical relation such as the "use" relationship. Fig. 4 shows the proposed agent where *DA* is a static agent that carries some information; automatically performs some set of functions and belongs to a service grid provider node. RA is a mobile agent that carries some information; automatically performs some set of functions and belongs to a service grid node.

3.2 The Model Description

The registration and discovery model consists of three elements registry architecture, fault tolerance and load balancing strategy, and discovery algorithm. In this section we discuss all these elements.

a) Registry architecture

The registry architecture includes node class formulation, head appointment, node subscription. In class formulation, nodes are gathered together in a set of classes. This classification is based on the hierarchal relations among the service grids in the DO, which means their defined semantic relation on the DO. For example, nodes that provide service grids that belong to the computing concept in the DO can form a class of nodes called computing class. We design an algorithm to accomplish the class formulation. In head appointment, each class needs to have a head that will ease the communication between the different classes. In this process, we first need to define the headship features, for which a node needs to qualify to become a head. In the second step, a head appointment algorithm calculates the similarity between the nodes and the predefined headship features, and selects the class head based on the degrees of similarity. An algorithm to perform the head appointment is designed. Node subscription refers the procedure of assigning a new node to an existing class or set of classes that corresponds to its service concept. Subscription is done by the node subscription algorithm. In this algorithm, we assume that the new node has been given the information about the selected service grid concepts during the settings, and the new node sends a message that contains its service concept to any existing node (member/head). The algorithm takes the service concept of the new node, and calculates the similarity degree between the service concept and the related class heads. If the similarity degree attains the predefined threshold, the new node is added to the class of that head. Finally, the algorithm returns the list of heads, for which the node has been assigned to, if the new node has more than one service grids that belong to different concepts. More details about the three mentioned algorithms can be found in [22].

b) The fault tolerance and load balancing strategy

The fault tolerance and load balancing strategy address the issues of dynamicity of the service grid nodes status, and the management of the node of classes in terms of the number of classes and size of each class. We incorporated two approaches respectively. The first one is called class maintenance which deals with a situation of failure in a class head and failure of a class member. The approach replaces the respective failed node (head/member) with another node (detail about that can be found in [22]).

In load balancing strategy, the classes of the service provider nodes are supposed to be managed by their respective heads (e.g. hosting the service grid metadata of the class members). This management process involves a huge amount of messages due to the intraclass and interclass communications. As a matter of fact, if we do not have optimization strategy to manage this tremendous amount of traffic, we will eventually be in a situation of bottle neck in the head. We use the idea of having few hundreds of nodes to be managed by one head. Therefore, we can define a variable called max number of nodes (μ) in a class to control the number of nodes in the class. The number of classes in a given intergrid starts by selecting the most general concepts in the DO, then when the nodes under a particular class (concept) has reached μ, we split the concept by selecting a number of more specific subconcept. This will ensure that every class can grow smoothly with a balanced management in the heads.

c) The discovery algorithm

By assembling the above framework components, we will have an intergrid that has a set of nodes assigned to some classes with their heads. The collection of heads forms a head node layer, whereas the collection of classes and members forms the member node layer. Each node has two agents (DA and RA), and implements the SGMP element to describe their service grid information. Communication between the nodes will be through the exchange of messages between the agents. In addition to service SGMP, heads implement the SDM element to assist members to describe and register their services. Neighboring nodes in each class exchange information about their services so that each node will have local information. Each head will hold the entire information of its class service as they are sent by the member nodes. In addition to that head is supposed to have some information about the other heads which includes their classes' concept.

The discovery algorithm addresses the search of the intergrid services on the network based on the cached information and dynamic matching. The cached information is the presence of a particular service in a node, which is got through the information exchange. Dynamic matchmaking is the similarity calculation between agents that represent service provider and requester, using the similarity function. The algorithm works as follows:

a) Based on the goals that are stored in the service goal template of the semantic description manager or the head of that user' node, the user selects the preferred goal and obtains instance of the goal. The user then adds the actual values of the service capabilities, which enable the RA to form a service request vector(s), say 6 services.

b) If there is local information about some neighbouring nodes that has been given by their DAs, RA sends a request to any neighbouring node n_i that is associated with all or part of the 6 requested services and the threshold of the similarity degree.

c) Based on the description of services in RA and DA, the similarity degree of the two agents sim(RA, DA) concerning the service properties of the requested service and provided service is calculated.

d) If the similarity degree of sim(RA, DA) reaches a user defined threshold value, then the node n_i is selected; and the check is done whether there are still remaining requested services to be searched.

e) If there are remaining service request, steps c and d are repeated until none of the nodes in the class is any longer associated with the requested service.

f) If so, then the remaining requested services are sent to a class head c_i.

g) From the head information, head c_i sends the service request to another class head/heads c_j that may have the remaining requested service based on the concepts.

h) For each head, the steps (b), (c), (d) and (e) are performed until all the 6 requested services are found.

4 Results and Discussion

In this section we present a comprehensive quantitative evaluation with respect to the overall performance of the proposed RD framework. We have chosen one of the P2P simulators called PeerfactSim.KOM [23] to simulate the intergrid environment with the application of the proposed system. The evaluation of the system is based on some common performance metrics found in the literature [24] and [3]. This includes the percentage of the discovered services in a given goal request (Request/query hit), and the response time for the service request to be answered. These metrics are calculated in different settings of the nodes and service requests. Therefore, we start with a few numbers of nodes, and scale them gradually to simulate the increase of the services in the actual intergrid system. We also vary the rate of service requests from small number of requests to bigger number to simulate the increase of users in the intergrid system. We analyze the results of the different settings by highlighting the causes of the effects of the different setting to the results.

4.1 Experimental Setup

We build an intergrid system that consists of n nodes. The size of the nodes n is scaled from 100 to 1000 with scale of 100 and 200. Since the creation of service grid domain ontology and dictionary ontology are out of our scope, we simulate these ontologies by representing them numerically. Where the concepts of the ontology are simulated by positive integer values such as 1, 2, …k, and each concept has subconcepts/properties which are some predefined set of values. Based on that, the concepts are representing the services' concepts and the predefined values are representing the services themselves. Each of the nodes has a set of the services. The number of these services is varied between 1 and 6 services. The reason of having this range of numbers is that our RD system is based on aggregating the service grid resources and services metadata information, which obviously reduces the range of services in the intergrid system. The allocation of the size of services in each node is random, which is the same as the assignment of concepts to node. This is to simulate the fact that in intergrid system we may not be able to neither express precisely the number of the services in each node nor the type or the concept to which these services belong, but surely we can define the concepts in the first place. During the simulation of class formulation, each of these nodes will be joining a particular class based on the randomly assigned concept. The number of class is based on the number of concepts (if the selected concepts for the overall nodes are five, the corresponding number of the classes is also five). The selection of concepts is proportional to the size of the intergrid system. This is in correspondence to the super-peer architecture where the number of super peers is based on the size of the network. In [25] the number of super-peer node is implemented to be 5% of the nodes that have very high capacity to handle queries. We have adopted the same percentage so as to systemize the distribution of the node to classes in way that allows us to conveniently discuss the performance of the system. Therefore, an intergrid system that has a size of 1000 nodes will have 10 classes. For simplicity, during the simulation, nodes that have high capacity which are supposed to be the heads of classes will join the network first and declare themselves.

4.2 Performance of the New Framework

We conducted 16 (this number corresponds to the variation of service request generation and TTL values) independent experiments for different service request portions and intergrid sizes. In each experiment, the mechanisms and algorithms that we have designed mentioned above are simulated. We first start our evaluation with the first performance metric, which is the service request hit.

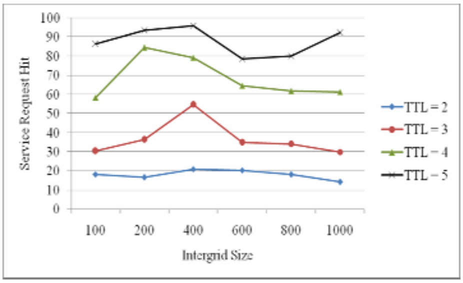


Fig. 5. Discovered Services for generated requests equivalent to 25% of the intergrid size with different TTL values

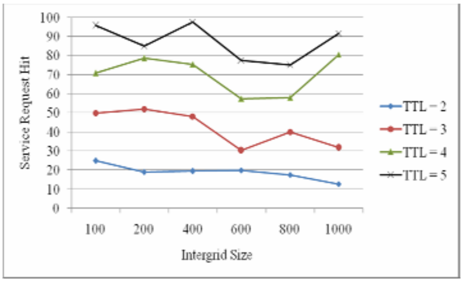


Fig. 6. Discovered Services for generated requests equivalent to 50% of the intergrid size with different TTL values

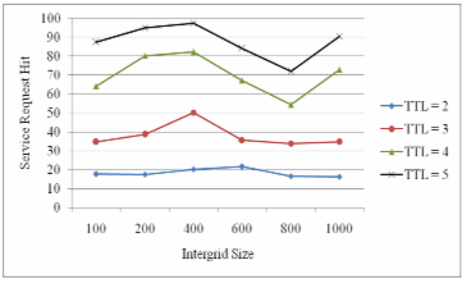


Fig. 7. Discovered Services for generated requests equivalent to 75% of the intergrid size with different TTL values

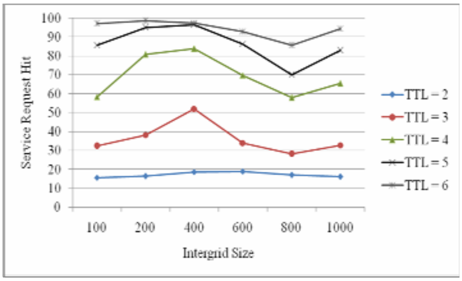


Fig. 8. Discovered Services for generated requests equivalent to 100% of the intergrid size with different TTL values

Fig. 5-8 show the simulation results for service requests generated by the nodes in percentages of 25%, 50%, 75% and 100% of the actual size of the intergrid. We control the forwarding of the request message from the requester node to the provider by the TTL values since we implement the super-peer architecture in our registry.

It appears from the Fig. 5-8 that the rate of discovered services is low when the TTL is equal to 2. This is because the scope of service request forwarding is limited to within the classes only, or between the heads if it happens that the head node itself generated the request. It is also very clear that the rate of discovered services becomes smaller with the increase of request rate and intergrid size. This can lead to an increase in the overall number of pages.

For example, in Fig. 5 the rate of the discovered services achieves 25% initially, and then drops gradually until 15.62% in Fig. 8. This is because as the service requests increase, the portion of the requests that is sent out of the requester node classes may be higher. This may also happen when the size of the intergrid system is scaled up. Also the four figures unambiguously indicate that the increase of TTL will allow the discovery of more services. For instance, the discovered service rate reaches its highest value 95.83% for an intergrid system consisting of 400 nodes. However, the cases of intergrid size 600 and 800 nodes appear to be different as the rate of discovered services decreases gradually until it reaches the lowest value at size of 800 nodes. The reason behind that is due to the implementation of the load balancing algorithm. In fact, the initial idea of the load balancing mechanism is to split the concept from general to a more specific concept so that we get more classes when a class reaches the maximum predefined size. However, this is hard to be simulated with the simulator as the creation of the nodes, services, and concepts is supposed to be before the intergrid join process starts. Therefore, we simulate the load balancing algorithm by creating new classes during the join process. In this case, if a head of class gets 100 hundred nodes in its class it will reject any new node that wants to join the system. When this happens, the rejected node will create a new class of the same concept and accept other nodes that want join the intergrid and have the same service concepts. Therefore, in the case of intergrid size 600, there are few classes that created, and there are more in the case of size 800 nodes. So these nodes cannot reach the services that are available beyond the TTL of value 4 for instance. As can be observed in all of the figures the rate of the discovered services starts increasing at TTL of value 5. To further investigate that observation, we increase the TTL value up to 6. As indicated in Fig. 8 the rate of the discovered services is slightly increased at the 800 intergrid size. Meanwhile, it achieves the highest rate with the small intergrid sizes such as 100 and 200 nodes. In addition to that, the rate of discovered services also increases with intergrid size of 1000 nodes. This could possibly be because the created new classes have more number of nodes, which influences the rate of the local discovered services to be higher.

All in all, it is observed that providing more TTL value causes the discovery of more services. However, one may argue that the increase of the TTL may inherit high traffic in the intergrid network. Nevertheless, in our case, the forwarding of service requests takes place only if the request has some semantic relation with the provider, if this not the case then the service request will be forwarded to all neighbors of the head node. Obviously, this will reduce the traffic in the intergrid system and the increase of the TTL value will not cause overhead on the network.

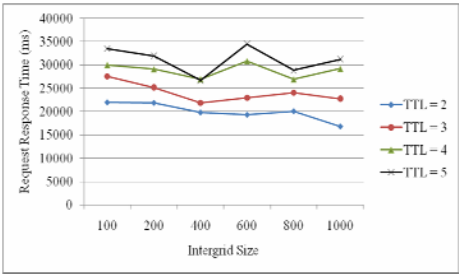


Fig. 9. Service Request Response Time for generated requests equivalent to 25% of the intergrid size with different TTL values

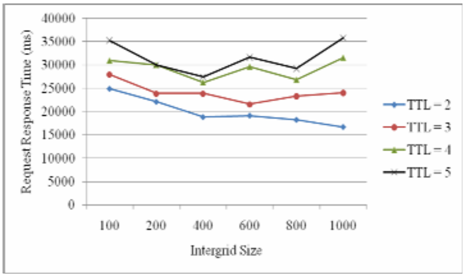


Fig. 10. Service Request Response Time for generated requests that equivalent to 50% of the intergrid size with different TTL values

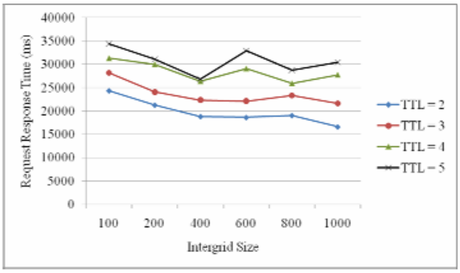


Fig. 11. Service Request Response Time for generated requests equivalent to 75% of the intergrid size with different TTL values

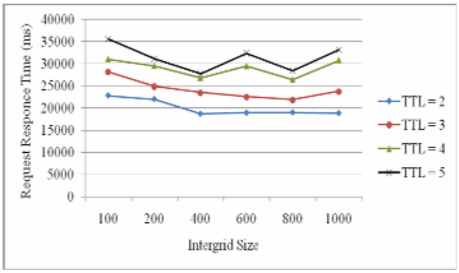


Fig. 12. Service Request Response Time for generated requests equivalent to 100% of the intergrid size with different TTL values

Our second point of discussion is on the service request response time of the proposed RD framework. In fact, we use the simulator timer to measure the time between the generation of service request by the requester node until when an answer is given to the requester node. For example, a node may generate a request at time 180000000 (simulation time) and a response may be given at the time of 180017503, therefore the response time is 17503 millisecond (ms). We calculated the average value of the response time in each set of generated service requests percentage. Fig. 9-12 illustrate these values.

It is apparent from all the figures that the increase of service request generation will increase the response time. This also happens when we increase TTL value. For example in Fig. 9, the average response time for generated service request equivalent to 25% of intergrid size of 100 nodes and TTL value 5 is 33486ms. The value becomes considerably higher (35569ms) in Fig. 12 when the service request rate is equivalent to 100% of intergrid nodes. However, the increase of intergrid size does not affect the request response time much, as the curve of the response time fluctuates in all four figures (9-12). Clearly, this indicates that the increase of the response time is not linearly related to the size of the intergrid nodes. This due to the decentralization of service requests processing as each head processes the service requests that are directed to it only. This ensures that the scale of the intergrid size will not cause performance degradation to the proposed RD system, which ensures sustainability of the system irrespective of the scale of the intergrid users as well as service grids. With this result, it is convincing that the proposed RD system is able to meet the performance requirements for the intergrid RD system. This includes scalability, decentralization and dynamism. The service request hit rates obtained from different intergrid sizes shows that the proposed RD system can scale with the intergrid system as well. The response time has no linear dependency on the scale of the intergrid size which proved the decentralization feature. Lastly, the dynamism feature has been achieved by the fault tolerance mechanism. It worth to mention that, the framework complexity is linear, which renders the system as capable of providing high performance.

5 Conclusions

In this paper we presented a new RD framework. The framework has a conceptual model for semantic description that treats the small grids of the intergrid system as services (service grids) and their semantic representation has been based on that; a semantic registry architecture that specifies semantically the distribution of the service grids metadata directories and their management with regard to scalability and dynamism of the service grids metadata; and an agent based discovery algorithm that exploits the description model and the registry architecture to search and select the service grids on behalf of the intergrid user. We have shown the effectiveness of the framework through some discussions and analysis, and an extensive simulation work which has confirmed the effectiveness of the framework.

References

1. Chao-Tung, Y., Wen-Jen, H., Kuan-Chou, L.: A Resource Broker with Cross Grid Information Services on Computational Multi-grid Environments. In: Proceedings of the 9th International Conference on Algorithms and Architectures for Parallel Processing. Springer, Taipei (2009)
2. Lamnitchi, A.L.: Resource Discovery in Large Resource-Sharing Environments. Department of Computer Science, PhD. The Univercity of Chicago, Illinois, 1-1 (2003)
3. Mastroianni, C., Talia, D., Verta, O.: A Super-Peer Model For Resource Discovery Services in Large-Scale Grids. Future Generation Computer Systems 21, 1235–1248 (2005)
4. Shen, H.: A P2P-based Intelligent Resource Discovery Mechanism in Internet-based Distributed Systems. Journal of Parallel and Distributed Computing 69, 197–209 (2009)
5. Tuttle, S., Ehlenberger, A., Gorthi, R., Leiserson, J., Macbeth, R., Owen, N., Ranahandola, S., Storrs, M., Yang, C.: Understanding LDAP Design and Implementation. IBM (2004)
6. Fitzgerald, S., Foster, I., Kesselman, C., von Laszewski, G., Smith, W., Tuecke, S.: A Directory Service for Configuring High-Performance Distributed Computations. In: 6th IEEE Symposium on High Performance Distributed Computing, pp. 365–375. IEEE Computer Society Press, Los Alamitos (1997)
7. Cooke, A., Gray, A.J.G., Ma, L., Nutt, W., Magowan, J., Oevers, M., Taylor, P., Byrom, R., Field, L., Hicks, S., Leake, J., Soni, M., Wilson, A., Cordenonsi, R., Cornwall, L., Djaoui, A., Fisher, S., Podhorszki, N., Coghlan, B.A., Kenny, S., O'Callaghan, D.: R-GMA: An information integration system for grid monitoring. In: Chung, S., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 462–481. Springer, Heidelberg (2003)
8. Zanikolas, S., Sakellariou, R.: A Taxonomy of Grid Monitoring Systems. Future Generation Computer Systems 21, 163–188 (2005)
9. Steven, F.: Grid Information Services for Distributed Resource Sharing. In: Proceedings of the 10th IEEE International Symposium on High Performance Distributed Computing. IEEE Computer Society, Los Alamitos (2001)
10. Schopf, J.M., Pearlman, L., Miller, N., Kesselman, C., Foster, I., D'Arcy, M., Chervenakand, A.: Monitoring the Grid with the Globus Toolkit MDS4. Journal of Physics: Conference Series 46, 521 (2006)

11. Ruay-Shiung, C., Min-Shuo, H.: A Resource Discovery Tree Using Bitmap for Grids. Future Generation Computer Systems 26, 29–37 (2010)
12. Trunfioa, P., Taliaa, D., Papadakisb, H., Fragopouloub, P., Mordacchinic, M., Pennanend, M., Popove, K., Vlassovf, V., Haridi, S.: Peer-to-Peer Resource Discovery in Grids: Models and Systems. Future Generation Computer Systems 23, 864–878 (2007)
13. Marzolla, M., Mordacchini, M., Orlando, S.: Peer-to-peer Systems for Discovering Resources in a Dynamic Grid. Parallel Computing 33, 339–358 (2007)
14. Brocco, A., Malatras, A., Hirsbrunner, B.: Enabling Efficient Information Discovery in a Self-Structured Grid. Future Generation Computer Systems 26, 838–846 (2010)
15. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What are Ontologies, and Why Do We Need Them? IEEE Intelligent Systems 14, 20–26 (1999)
16. Ludwig, S.A., Reyhani, S.M.S.: Introduction of Semantic Matchmaking to Grid Computing. Journal of Parallel and Distributed Computing 65, 1533–1541 (2005)
17. Said, M.P., Kojima, I.: S-MDS: Semantic Monitoring and Discovery System. Journal of Grid Computing 7, 205–224 (2009)
18. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284, 34–43 (2001)
19. Hassan, M.I., Abdullah, A.: Semantic-Based Grid Resource Discovery Systems A Literature Review and Taxonomy. In: International Symposium in Information Technology (ITSim), vol. 3, pp. 1286–1296. IEEE Xplore, Los Alamitos (2010)
20. Subramaniam, R., Nakata, T., Itoh, S., Oyanagi, Y., Takefusa, A., Anzaki, T., Mizoguchi, K., Tazaki, H., Mori, T., Suzuki, T., Hamada, M., Maeshiro, T., Takashima, H., Yoshioka, M.: Guidelines of Requirements for Grid Systems v1.0. GFD-I.145. Open Grid Forum (2009)
21. Gruber, T.R.: Toward Principles For The Design Of Ontologies Used For Knowledge Sharing. International Journal of Human-Computer Studies 43, 907–928 (1995)
22. Hassan, M.I., Abdullah, A.: A New Resource Discovery Framework. The International Arab Journal of Information Technology (IAJIT) 8, 20–28 (2011)
23. Kovacevic, A., Kaune, S., Mukherjee, P., Liebau, N., Steinmetz, R.: Benchmarking Platform for Peer-to-Peer Systems. In: it- Information Technology (Methods and Applications of Informatics and Information Technology), vol. 49, pp. 312–319 (2007)
24. Mastroianni, C., Talia, D., Verta, O.: Designing an information system for Grids: Comparing hierarchical, decentralized P2P and super-peer models. Parallel Computing 34, 593–611 (2008)
25. Yatin, C., Sylvia, R., Lee, B., Nick, L., Scott, S.: Making gnutella-like P2P systems scalable. In: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, Karlsruhe (2003)

Agent-Based Grid Resource Management

Ognian Nakov, Plamenka Borovska, Adelina Aleksieva-Petrova,
Anastasios Profitis, and Luka Bekiarov

Departmant of Computer Science, Technical University – Sofia, Kliment Ohridski 8,
1000 Sofia, Bulgaria
{nakov,pborovska,aaleksieva}@tu-sofia.bg, profit@sch.gr

Abstract. Grid technology and grid computing appear to be potentially the next generation platforms for solving large problems in science and engineering. Grid resource management is a main activity, which largely affects the productivity of the Grid environment. The main problem is how to manage millions of heterogeneous resources that are distributed across multiple organizations and administrative areas. In order to solve this problem and increase the efficiency of the available resources, a method is proposed for scheduling resources in a Grid. The method combines metaheuristics iterative local search forward algorithm for timetable and uses auction for conflicting resources.

Keywords: auction, agent-based technologies, Grid, resource management, timetable.

1 Introduction

In the area of grid resource management research aims to resolve a number of engineering and scientific problems. Grid resource management is a main activity, which largely affects the productivity of the Grid environment. A lot of research has been done in this area, which uses a variety of algorithms applied in the management of resources such as genetic algorithms [1, 2], ant colony optimization [3, 4] and other metaheuristics algorithms [5, 6].

There are two main methods for grid resource management: hieratical approach, which supports the centralized or decentralized scheduler organization and the market based paradigm. The first one is wide used and there are many solutions based on it but its main disadvantage is that it does not take into account resource's price during the scheduling of resources [7]. Markets have emerged as a new paradigm for managing and allocating resources in complex systems. Several research systems [8] have explored the use of different economic models for trading resources to manage resources in different application domains: CPU cycles, storage space, query processing, and distributed computing.

Agent-based technologies give new opportunity to solve the efficient in grid resource management and refers to a model in which the dynamic processes of agent interaction are simulated repeatedly over time, as in systems dynamics, and time-stepped, discrete-event, and other types of conventional simulation [9].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 463–470, 2011.

The main goal of this paper is to describe a method for grid resources management and agent-based middleware which implements it.

The next part of the paper proposes a method for resource management in grid, which provides effective scheduling of grid resources and satisfies the conflicting requests for shared resources. It applies the metaheuristic iterative local search forward algorithm to generate the timetable of grid resources and task. The problem of generating a timetable consists in distribution of tasks to resources corresponding to certain constraints and set time intervals. The conflicting grid resources are allocated through an auction. Auction is used for resource allocation in Grid because they provide a decentralized structure and are easier to implement than other economic models.

The third section describes the agent-based middleware for grid resource management and the using technology. The next section contains some experiments and the results of the implemented solution. Next is the conclusion and future work.

2 Method for Grid Resource Management

We propose the method for grid resource management using iterative algorithm with local search forward and auction for conflicting resources (Fig. 1).

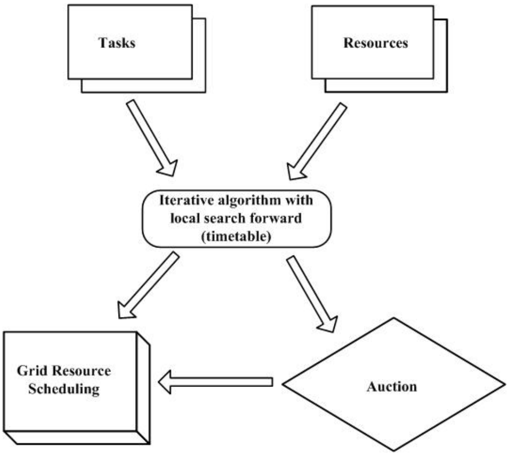


Fig. 1. Resource management using algorithm for timetable and auction.

We define a set of user tasks T = $\{T_1, T_2,\ldots,T_n\}$ and a set of grid resources R = $\{R_1, R_2,\ldots, R_m\}$. The initial data comprise the following items:

1. *Days, time unit and periods* - each day is divided into a fixed number of time intervals (duration 1 minute) within one day.

2. *Tasks (T)* - each task is described by the $T_i = (k_i; d'_i; w_i; b_i)$, where:

- k_i - the capacity of the task, for example CPU slots needed for the task *i*;
- d'_j - the final deadline for the task *i*;

- w_i - weight of task *i*, which presents the importance of the task;
- b_i - the maximum budget for the task *i.*

3. *Resources (R)* - every resource is described by the $R_j = (c_j; t_j; r_j; mp_j)$, where:

- c_j - capacity of the resource, for example available CPU slots;
- t_j - the time to perform the task;
- r_j - a minimum price of the resource;
- mp_j – a maximum price of the resource.

4. *Current timetable* - each schedule consists of a number of tasks that occupy the optimal number of resources.

The hard and soft constraints sets are defined. The hard constraints are as follows:

1. Tasks and resources must be of the same type, such as CPU slots or storage space.

2. The capacity of the resource must be greater than or equal to the capacity of the task.

3. Two tasks cannot use the same resource at the same time.

The soft constraints are:

1. First to perform tasks with greater weight, so the degree of importance is high.

2. If there is no idle resource, the task is assigned to resource, which is released first.

The algorithm proposed by Muller [10, 11] uses forward based search that extends a partial feasible solution towards a complete solution. The next solution is derived from the previous solution by allocating some event and removing conflicting events from the partial schedule. Every step is guided by a heuristic that aims at minimal violation of soft constraints requirements.

The algorithm progresses in iterations using two basic data structures: a list of tasks that have not been allocated yet and a timetable. At each step the algorithm tries to improve the current timetable. At the beginning of the algorithm the timetable is empty. Initially, all tasks are included in the list of not assigned tasks. In the next step the algorithm selects one unassigned task and calculates location (time and resources) according to given constraints which can be distributed. The algorithm continues from one solution to another until all hard constraints are met.

In order to find a better new solution, some tasks are removed from the schedule. These tasks are selected at random from all the tasks that have violated soft requirements. After re-allocating resources for these tasks (with possible removal of other tasks), the timetable is approved only if there is a smaller number of points soft constraints of requirements of the previous timetable, otherwise the schedule is rejected.

If there are some conflicting resources then the auction is started. The different auction mechanisms could be used like First-Price, Vickers and Double Auctions.

3 Agent-Based Grid Resource Management Middleware

The Java software program JADE (Java Agent Development Framework) [12] is used in creating the forthcoming application. It facilitates the introduction of a multi-agent system in an environment that abides by FIPA specifications [13] and supports the conscription of graphical instruments to maintain and eliminate errors. The agent platform is distributed across computers and the configuration is remotely controlled through a Graphical User Interface (GUI).

Fig. 2 shows the agent-based model for grid resource management which implements four main agents: broker agent, resource agent, trade agent and scheduler agent.

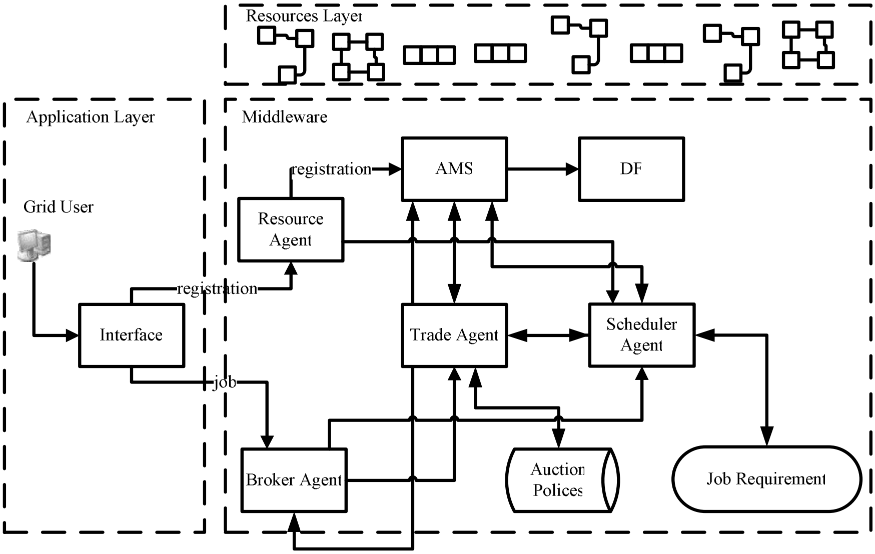


Fig. 2. Agent-based grid resource management model.

The broker agent represents the submitted user job. Its main function is to create and analyze the user job. The broker agent communicates with scheduler and trade agent in order to receive appropriated resources.

The main function of the resource agent is to publish all needed information about the resource into the agent management system (AMS). This information is used when the scheduler agent is looking for appropriated resources.

The scheduler agent can allocate resources for a task and partition the tasks to execute. Its main function is to reserve resources in advance, monitor job execution status, and reschedule events.

The trade agent is an agent that negotiates with resource users and sells access to resources. It aims to maximize the resource utility and consults pricing algorithms defined by the users during negotiation.

The agent-based middleware for resource management is shown in Fig.3, where every grid node has a container with resource agents representing the physical node resources. In the main container are located the trade and scheduler agent. All agents use the ACL (Agent Communication Language) message to migrate in JADE platform.

The main function of the agent management system (AMS) is to integrate all kinds of information in a grid system, static and dynamic and provide a unified information access interface for users.

Four primary packages and two additional packages are utilized and implemented:

- Broker – with two packages OfferRequestsServer and RequestsResponseServer, which records the behavior of the broker agent.
- Scheduler – with two packages ResourceRequestsServer and RequestsResponseServer, which records the behavior of the scheduler agent.
- Trade Agent – with one package (AuctionPerformer), which records the behavior of the auctioneer (the resources broker).
- Job – a supplementary package that is used by the Consumer in order to generate tasks.
- Resource– a supplementary package that records each given resource (type, price and owner).

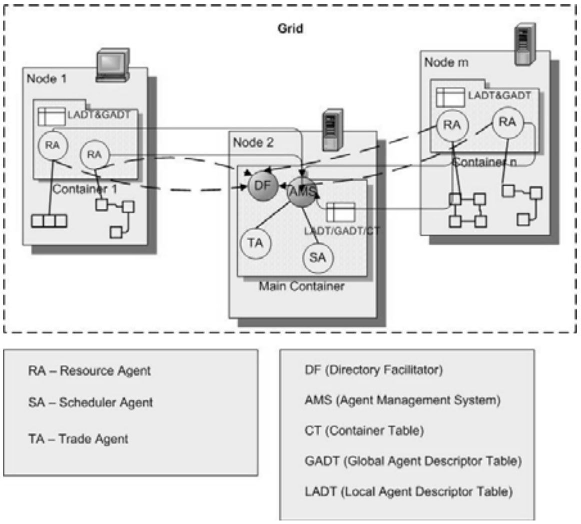


Fig. 3. Agent-based middleware using JADE framework.

All four packages are derived from the Agent class, which is a built-in JADE framework.

All the characteristics of the agent are initialized and the different behaviors that the agent will express are assigned. These behaviors are described via internal categories derived from the primary category Behavior.

Each agent maintains its own message trail. The primary parts of the message are the body (contents), sender, recipient(s), the type (according to the standards of FIPA), the topic of the "conversation", etc.

There are different types of messages. The agent, in keeping with its assigned behavior, waits for a CFP (Call For Proposals) message, which serves as an "invitational" message. Some of the agents await an ACCEPT_PROPOSAL message, which means "the offer has been accepted". The agent waits for an INFORM message, which means "the offer has been accepted". As long as there are no such messages on hand in the message trail the agent is blocked from using processor resources. When such a message arrives the agent is activated and the corresponding auctions are carried out.

For example Trade Agent has one primary behavior: AuctionPerform class – derived from Behavior.

In this case we have "round-specific behavior", whereby in each round the auctioneer carries out specific auctions, after which it proceeds.

Round 1: The auctioneer sends messages to all agents announcing the start of the auction, and then waiting for their offers;

Round 2: After the auctioneer has announced the start of the auction it waits for all agents to send their replies – requests from buyers and offers from sellers. Only when the auctioneer has received responses from all agents does it proceed.

Round 3: After information from all participants is received then it is possible for the auction to be carried out. The method auction() is called upon, which carries out the second-price auction.

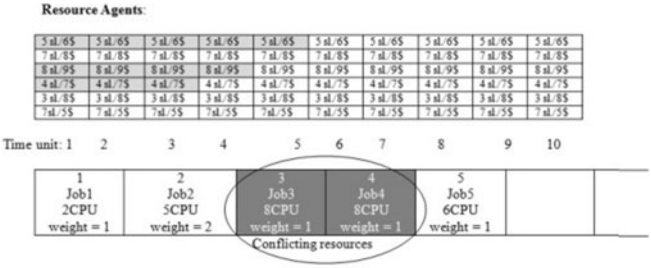
Round 4: After this auction is concluded and all resources have been distributed the corresponding messages are sent to all participants and the auctioneer moves on to the next auction.

The critical moment that needs to be mentioned is the registration of the agents into the so-called "yellow pages". Every agent is registered in them in order to easily be found in the future along with the type of services they offer as well as a brief additional description. Every agent has the ability both to register and to search for services. This registration is subsequently used by the auctioneer in order to restrict the relevant agents according to their type.

4 Experimental Results

First we describe our case study for grid resource management and present the results of auctions held.

The following parameters are given for the agents participating in the experimental formulation (Fig. 4).



Resource Agents:

5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$	5 sl/6$
7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$	7 sl/8$
8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$	8 sl/9$
4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$	4 sl/7$
3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$	3 sl/8$
7sl/5$	7sl/5$	7sl/5$	7sl/5$	7sl/5$	7 sl/5$	7 sl/5$	7 sl/5$	7 sl/5$	7 sl/5$

Fig. 4. Resource management case study.

The producers are suppliers of processor time (represented by six agents) with the following parameters:

- Agent 1 is CPU producer with 5 slots and the minimum price 6$/slot/time unit;
- Agent 2 is CPU producer with 7 slots and the minimum price 8$/slot/time unit;
- Agent 3 is CPU producer with 8 slots and the minimum price 9$/slot/time unit;
- Agent 4 is CPU producer with 4 slots and the minimum price 7$/slot/time unit;
- Agent 5 is CPU producer with 3 slots and the minimum price 8$/slot/time unit;
- Agent 6 is CPU producer with 7 slots and the minimum price 5$/slot/time unit.

The five jobs are submitted with the following parameters: the first job needs 2CPU slots for 3 time unit, the second - 5CPU slots for 2 time unit, the third job - 8CPU slots for 5 time unit, fourth job - 8CPU slots for 5 time unit and the fifth - 5CPU slots for 2 time unit.

The jobs and available resources are submitted to the Scheduler Agent which allocates resources for these jobs (with possible removal of other tasks) and timetable. With conflicting resources the second-price auction mechanisms are applied.

In auctions the suppliers submit their available resources to an auctioneer for processor time; i.e. with the average profit per slot on the basis of the value of the function of demand.

Customers submit their requests by means of jobs, as every job specifying the type, number and duration of resources needed by it. Every one of them has a defined available budget ($G) with which to pay for the resources needed for its tasks, which is current for a fixed period of time. They submit requests for resources as long as they have tasks and a current budget.

The auctioneer maintains a list of available resources, running through the resources in each round in the order in which they are listed.

Every round is conducted on the principle of the second-price auction in order to determine the winner. With this type of auction the buyer who wins is the one who has submitted the highest bid, thereby receiving the resource, and if there is a buyer who has submitted a second-highest bid that is above the minimum price set by the supplier, the final price for the resource is set at this second-highest bid.

5 Conclusion

In this paper, we presented the design, implementation, and evaluation of agent-based approach to grid resource management. We proposed a method for grid management which combines iterative local search forward algorithm to generate the timetable of grid resources and task and auction for conflicting resources.

Agent-based technologies are aimed at assisting in application development as they are based on a decentralized approach to intelligent agents. We have discussed the case study and the implementation of the agent-based simulation of auction for resource management. The advantage of this system is its agent-based architecture, taking into consideration their own requirements and problems.

Acknowledgments. Research is funded by the Bulgarian Ministry of Education and Science, project DTK 02-63/2009.

References

1. Cardon, A., Galinho, T., Vacher, J.: Genetic algorithms using multi-objectives in multi-agent system. Robotics and Autonomous Systems 33, 179–190 (2000)
2. Li, M., Yua, B., Qi, M.: PGGA: A predictable and grouped genetic algorithm for job scheduling. Future Generation Computer Systems 22, 588–599 (2006)
3. Leung, C.W., Wong, T.N., Mak, K.L., Fung, R.Y.K.: Integrated process planning and scheduling by an agent-based ant colony optimization. Computers & Industrial Engineerin (2009)
4. Xiang, W., Lee, H.P.: Ant colony intelligence in multi-agent dynamic manufacturing scheduling. Engineering Applications of Artificial Intelligence 21, 73–85 (2008)
5. Labat, J., Mynard, L.: Oscillation: Heuristic Ordering and Pruning in Neighborhood Search. In: Smolka, G. (ed.) CP 1997. LNCS, vol. 1330, pp. 506–518 (1997)
6. Buyya, R., Chapin, S., DiNucci, D.: Architectural models for resource management in the Grid. In: First IEEE/ACM International Workshop on Grid Computing. LNCS Series, Springer, Germany (2000)
7. Assuncao, M. D., Buyya, R.: An evaluation of communication demand of auction protocols in grid environments. Technical report, Computing and Distributed Systems Laboratory, The University of Melbourne, Australia (2006)
8. Li, C., Li, L.: The use of economic agents under price driven mechanism in grid resource management. Journal of Systems Architecture 50, 521–535 (2004)
9. Gomoluch, J., Schroeder, M.: Market-based resource allocation for grid computing. A Model and Simulation, 211–218 (2003)
10. Müller, T.: Interactive Heuristic Search Algorithm. In: Proceedings of the CP 2002 Conference - Doctoral Programme, Ithaca (2002)
11. Muller, T., Bartak, R.: Interactive timetabling: Concepts, techniques, and practical results. In: PATAT 2002—Proceedings of the 4th International Conference on the Practice And Theory of Automated Timetabling (2002)
12. Java Platform Development Framework (JADE), http://jade.tilab.com/
13. FIPA, http://www.fipa.org/

Current Practices of Programming Assessment at Higher Learning Institutions

Rohaida Romli[1], Shahida Sulaiman[2], and Kamal Zuhairi Zamli[3]

[1] College of Arts and Sciences (Applied Sciences), Universiti Utara Malaysia,
06010 UUM Sintok, Kedah, Malaysia
aida@uum.edu.my

[2] School of Computer Sciences, Universiti Sains Malaysia,
11800 USM, Penang, Malaysia
shahida@cs.usm.my

[3] School of Electrical and Electronic Engineering, Universiti Sains Malaysia,
Engineering Campus, 14300 Nibong Tebal, Penang, Malaysia
eekamal@eng.usm.my

Abstract. Assessing students' programming exercises has become a difficult activity that most educators encounter nowadays. The activity basically includes the tasks to construct questions and solution models in programming exercises as well as the method to evaluate students' solutions. Existing studies particularly in the area of programming assessment still have limited discussions on current practices in conducting the activity. This paper reports the preliminary study conducted among educators who have been teaching programming courses at higher learning institutions within the northern region in Malaysia. The study aims to gauge the current practices in the construction and evaluation of programming exercises item among educators at the associated institutions. The study used a questionnaire to gather the relevant data from the selected subjects. The results reveal that both the negative and positive testing criteria are essential in constructing and evaluating programming exercises. The findings of this study will be the input to identify the adequate criteria that should be included in developing a schema of test set for automatic programming assessment.

Keywords: programming assessment; positive testing; negative testing.

1 Introduction

Programming is one of the essential skills that computer science students should master. Thus, programming courses that are offered in any computer related discipline mainly aim to develop students' understanding of the programming principles. Students receive a lot of programming exercises as hands-on or take-home assignments in order to ensure they can practise consistently and effectively all principles and concepts of programming in their learning process.

Constructing questions in programming exercises is cumbersome. It requires adequate statements of specification that reflect the actual requirements that should be

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 471–485, 2011.

met. For an educational purpose, the requirements typically need to be justified by the teaching goals of the course [1] or specifically to be justified by the objectives to be achieved for each topic of a course syllabus. It is important to determine the extent to which students are able to acquire and practise all the programming concepts and principles that they learn during lectures. Besides, these requirements will contribute as measurement values in marking and grading students' programming exercises. The functional aspect of a program focuses on this testing aspect.

In addition to the difficulty in constructing programming exercises among educators, marking programming exercises is generally a tricky task too. Typically, the educators provide a solution model to guide them in the marking process for a particular programming exercise. The larger the number of students registered in computer programming courses, the longer the time required in the processes of marking and grading programming exercises. Indeed, it is more troublesome if it has to be carried out manually. Thus, an attempt to replace educators' effort with automated counterparts has gained a lot of attention. Several tools are available for automated assessment in different aspects of the program quality ranging from input to output testing such as GAME [2], PASS [3], CourseMaster [4], SAC [5], ELP [6], TRAKLA2 [7], Assyst [8], BOSS [9] and WeBWork [10].

Albeit the availability of diverse automated tools, researches with regard to automated assessment continue to mature. Existing studies employ different strategies in producing more accurate markers as well as to provide better functions. The main aim is to provide a more accurate assessment that is the key to assess students in achieving the learning outcomes in the programming course. To date, there is no definite approach to solve the problem of automated assessment. It may depend on the specific aims and objectives of the course that the educators teach or based on their own styles and preferences [11] or it may depend on the scheme of the program [12].

Diverse approaches for automatic programming assessment (APA) are available but there are limited studies in current practices among educators. Hence, this paper discusses the current practices in the construction and the evaluation of programming exercises item as the input to design a more objective schema of test set for APA.

The content of the remaining sections is organised as follows. Section 2 provides the details of the survey conducted for the preliminary study. Section 3 describes the results and discussion of the study. Section 4 highlights some of the existing approaches used in programming assessment. Finally, Section 5 concludes the paper.

2 The Survey

The respondents of the survey were educators who have been teaching programming courses at public higher learning institutions within the northern region in Malaysia. Four universities were selected in this five-month preliminary study: Universiti Utara Malaysia (UUM), Universiti Malaysia Perlis (UniMaP), Universiti Teknologi MARA (UiTM) Arau and Universiti Sains Malaysia (USM). One of the non-probability sampling technique which is known as judgement sampling was employed to choose a sample in the population. This technique was chosen because the respondents were selected on the basis of their expertise in the subject investigated [13]. The survey received a total of 27 responses. Table 1 tabulates the number of respondents based on each university in a descending order.

Table 1. Number of respondents from each participating university

University	Number of respondents
UUM	10
UiTM Arau	8
UniMaP	6
USM	3

The study aims to investigate the current practices of programming assessments in the construction and the evaluation of programming exercises item among educators who have been teaching programming courses at the selected institutions. The specific objectives include:

(1) To identify factors that might influence educators in constructing questions and solution model of programming exercises.
(2) To identify evaluation items that might influence educators in marking students' programming exercises.
(3) To measure how much educators prioritise the evaluation items in marking students' programming exercises.
(4) To gauge the current methods used in marking students' programming exercises.

We designed a questionnaire to collect the related information in the study. The main reason of choosing the questionnaire in the survey was due to the constraint of meeting all respondents who were scattered across different geographical areas. In this study, a mixture of close-ended and open-ended questions were designed to collect the investigative information. Open-ended questions aimed to derive additional information and suggestions from respondents. For the close-ended questions, the suggested factors that influence educators in the construction of questions and solution model of programming exercises as well as evaluation items were based on the information collected from literature surveys.

The questionnaire consisted of thirty-six questions that are divided into three parts: demography of respondents, construction of programming exercises item, and evaluation of programming exercises. There were both multiple choice and close-ended questions in the questionnaire. The questions that involved ratings used Likert Scale format. We used two types of estimations for the Likert Scale ranging from 1 to 5. The first type was frequency estimations which consisted of five values; *never*, *rarely*, *sometimes*, *often* and *always*. The second type was priority estimations which used the scale: *not a priority*, *low priority*, *medium priority*, *high priority*, and *essential*.

The result and discussion in Section 3 focuses on the following aspects:

(1) Brief background information of the respondents.
(2) The means of constructing questions and model solution of programming exercises.
(3) The evaluation items that might influence educators in marking students' programming exercises and how they are ranked.
(4) The current method undertaken in marking students' programming exercises.

3 Results and Discussion

The following sub-sections discuss the results of the study based on: demography of respondents, construction of programming exercise item and evaluation of programming exercise.

3.1 Demography of Respondents

The demography of respondents consisted of seven questions: gender, qualifications, status of appointment, teaching experience, background in teaching programming courses, programming courses that have been taught, and concept of programming applied in teaching. Fig. 1 shows the frequency of response for each question.

From the result tabulated in Fig. 1, it shows that most of the educators in the selected universities are master scholars (about 89%) and have sound experience since they have more than five years of general teaching experience and at least three years of teaching programming courses. This implies that senior educators (designation of senior lecturer and above) mostly teach core courses in computer science or information technology in this case programming course. In addition, a lot of efforts should be devoted to these courses to ensure students are able to understand very well all concepts and principles of programming which will be the basis for higher level courses.

As aforementioned in Section 1, the process of marking and grading programming exercises apparently increase the workload of educators. Thus, there is a significant tendency to most of the junior educators (with master scholar) to manage such courses as they have more enthusiasm in dealing with the workload. It is also shown that 67% of the respondents have experience in teaching introductory and advanced level of programming courses. Besides, about 74% of them are really familiar with object-oriented and/or procedural programming concepts. In term of programming concepts, most of the higher learning institutions applied object-oriented (OO, 33%) rather than procedural (19%) concepts.

3.2 Construction of Programming Exercise Item

This sub-section discusses the aspect of how educators construct questions and solution model of programming exercises. This solution model directly becomes the means to assess students' programming solutions. The questions of programming exercises typically comprise statements of specifications that comprise conditions or/and operations that a program can and cannot fulfil and invoke respectively. Besides, the source of references that educators use in preparing the mentioned aspects include topics in the syllabus, objectives of the corresponding topics, books, online resources, or self styles and preferences.

In term of constructing questions in programming exercises, educators construct the questions based on the criterion of *"the program does what it is supposed to do"*, or *"the program does what it is not supposed to do"*, or taking consideration of both of these two. Both of the criteria are categorised as positive testing [14]. The criterion of *"the program does what it is supposed to do"* is basically concerned with what operations or/and conditions that a program intends to accomplish and to satisfy respectively. While, the criterion of *"the program does what it is not supposed to do"* is commonly regarded with error handling operations or/and its conditions.

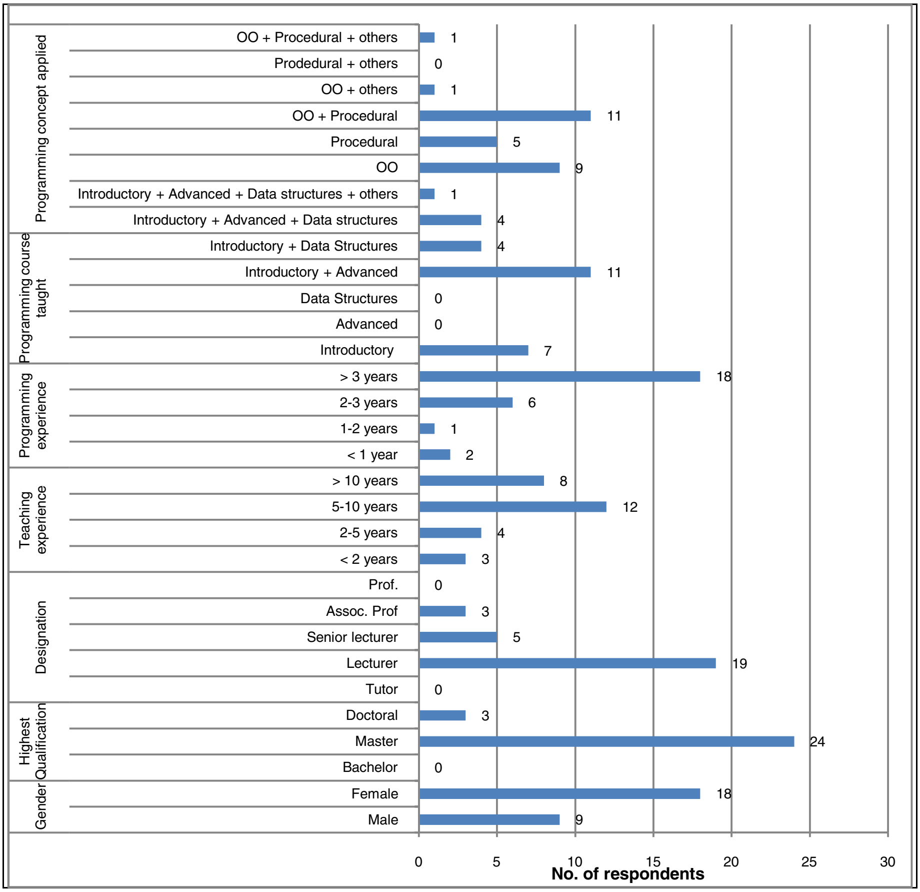


Fig. 1. Demography of respondents

Basically, for the lower level programming courses, educators intend to test students' programming ability in term of whether or not they are capable to solve some required operations by applying certain concepts of programming taught in a class. Thus, the specification statements will comprise a list of intended functions with certain specific conditions. They usually do not expect students to do something that is not specified to be done. As shown in Fig. 2, almost 75% of the respondents rated the frequency as *often* and *always*. This reflects that educators typically construct programming exercises specification based on *"the program does what it is supposed to do"* with regard to some basic operations to be solved. In conclusion, it implies that this criterion has essentially being considered as one of the main aspects in the construction of programming exercise items or questions.

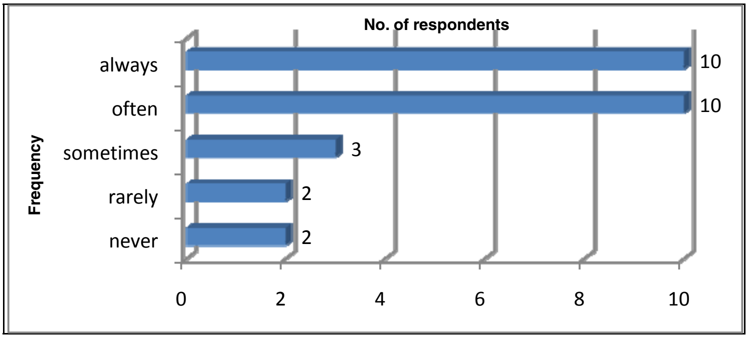


Fig. 2. Number of respondents who provide programming exercises specification based on "*the program does what it is supposed to do*"

Fig. 3 shows the frequency of providing programming exercises specification based on the criterion of *"the program does what it is not supposed to do"*. It differs from the criterion of *"the program does what it is supposed to"* because it considers the specification that caters error handling to equip more sufficient functions. The result shows the highest rating is narrowed down to the frequency of *sometimes*, that is 37% as the total percentage. The other 30% of respondents agreed to apply it as *often*. In conclusion, educators quite often consider this criterion when constructing programming exercise item.

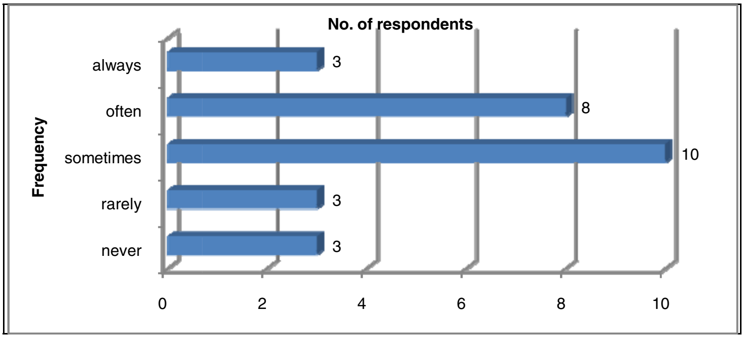


Fig. 3. Number of respondents who provide programming exercises specification based on "*the program does what it is not supposed to do*"

Fig. 4 shows the number of respondents who applied the concept of *"the program does what it is supposed to do"* in preparing solution model of programming exercises. The criterion is equivalent to what has been explained in constructing questions of programming exercises. As shown in Fig. 4, the highest frequency is given to *always* that is a total of 52%. The other 33% of respondents *often* used the concept. In conclusion, the criterion can be said as an essential aspect to be considered in preparing a solution model for programming exercises.

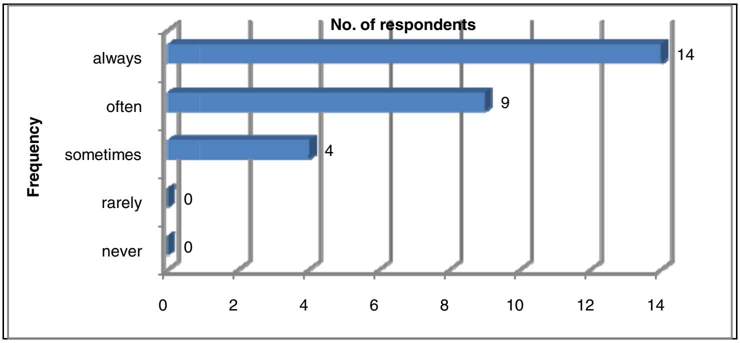


Fig. 4. Number of respondents who prepare a solution model of programming exercises based on "*the program does what it is supposed to do*"

However, for the aspect of constructing a solution model of programming exercises, there was an additional criterion included as part of the prior mentioned criteria. It was with regard to *"the program does not do anything that is not supposed to do"*. This criterion falls into negative testing [14]. It is concerned with condition or/and operations that purposely reveal unintended errors or faults that might break intended conditions or/and operations. Primarily, it depends on error guessing and relying upon the educators' experiences to anticipate the location of programming error.

The criterion of *"the program does what it is not supposed to do"* was also investigated in preparing a solution model of programming exercises. Fig. 5 depicts the result collected for this criterion. There are 41% and 26% of respondents stated as *sometimes* and *often* respectively. Thus, it depicts that more than 50% of the educators employ this criterion in preparing the solution model for programming exercises.

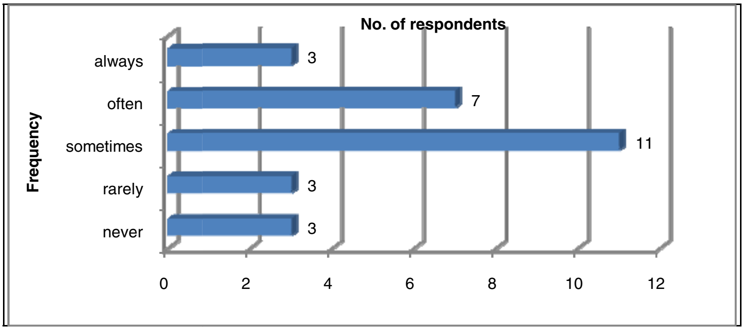


Fig. 5. Number of respondents who prepare model solution of programming exercises based on "*the program does what it is not supposed to do*"

The result of criterion *"the program does not do anything that it is not supposed to do"* (refer Fig. 6) seems to have quite a similar trend as shown in Fig. 5. The highest rating with the same percentage value was recorded as a similar frequency. Nevertheless, the higher rating was recorded for the frequency of *never* (19%) when compared to what is shown in Fig. 5. In total, about 70% of the respondents applied this criterion in preparing the solution model for programming exercises. Hence, it concludes that educators consider negative testing criterion as one of the important criteria in preparing a solution model.

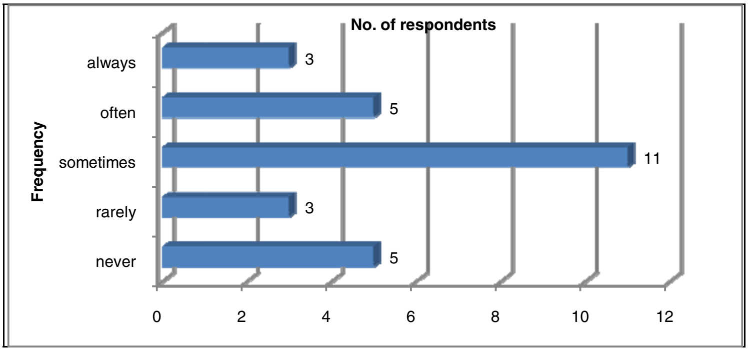


Fig. 6. Number of respondents who prepare a solution model of programming exercises based on "*the program does not do anything that it is not supposed to do*"

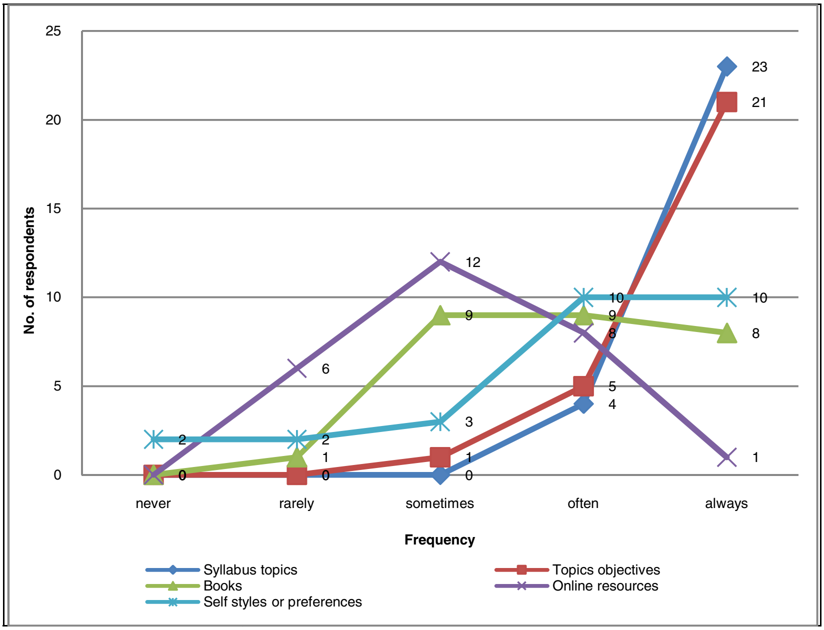


Fig. 7. Source of references in constructing questions for programming exercises

In term of references used in preparing both questions and solution model of programming exercises, Fig. 7 illustrates that the highest rating was recorded by the frequency of *always* with the percentage values of 85% and 78% for syllabus topics and objectives of syllabus topics respectively. However, only 4% of the respondents referred to the online resources of the same frequency. The highest rating was for the frequency of *sometimes*. For self styles and preferences, the same percentage value which is 37% was recorded for both the frequency of *always* and *often*. Thus, it can be stated that the topics in the syllabus and objectives of the topics are the most essential resources used in preparing questions and a solution model for programming exercises. The main reason is definitely to ensure the programming course achieves its goal to ensure students' understanding of programming principles at the end of the course.

3.3 Evaluation of Programming Exercises

This section discusses how educators assess students' programming exercises in terms of what evaluation items they have taken into consideration. In addition, it also relates to how they prioritise the evaluation items in the assessment process. The evaluation items asked to respondents were related to the following factors:

(1) success in program compilation,
(2) success in producing the correct output,
(3) correct logic structure of a code,
(4) conformation to the specified program specifications,
(5) date of submission, and
(6) plagiarism.

The evaluation items comprised both static and dynamic aspects of a program and included some grading criteria [15]. The static aspect of a program basically refers to the syntax or lexical aspect of a code. On the other hand, the dynamic aspect can be divided into either black-box or white-box testing criteria which are mainly based on a program execution. However, the respondents could include any additional evaluation items if possible.

Fig. 8 reveals the frequency of the respondents who were taken into consideration for the specified evaluation items (1) to (6) as listed above. It shows that five evaluation items were ranked as the highest percentage values for the frequency of *always*. They are items (1), (2), (3), (4), and (6), with their respective percentage values of 70%, 74%, 67%, 52%, and 44%. However, the highest percentage value for the date of submission was 48% and it was rated as *often*. Some respondents stated other additional evaluation items that they used in assessing students' programming exercises. Among them were the performance of a program, additional creative features, user interface, program readability and security. This implies that the evaluation items play as additional features to capture students' creativity and capability to produce a better quality program. In overall, it can be concluded that educators commonly apply both static and dynamic analysis criteria in assessing programming exercises.

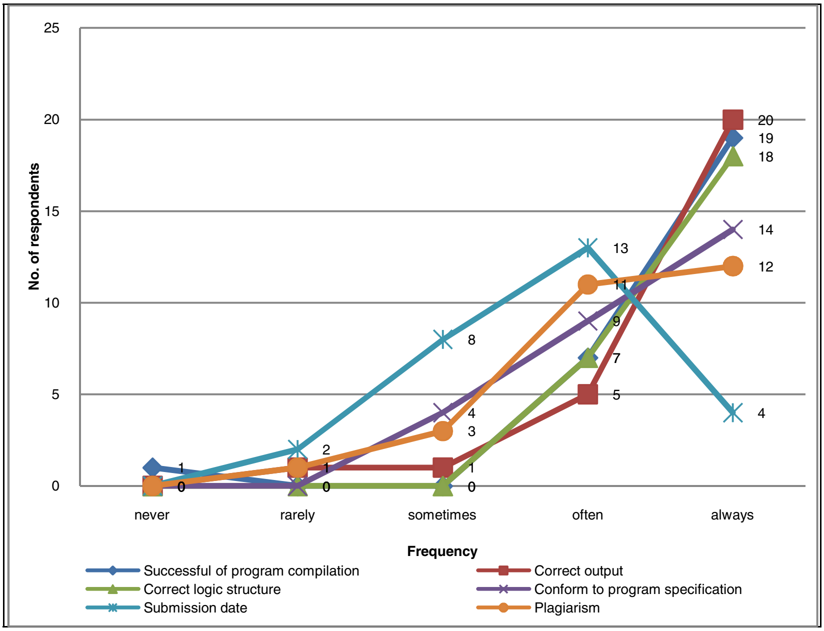


Fig. 8. Factors considered in the evaluation items of programming exercises

Fig. 9 illustrates the result of to what extend educators prioritise the evaluation items in assessing students' programming exercises. The result of the highest percentage values among the evaluation items demonstrates a similar trend as shown in Fig. 8. However, the evaluation item (1) and (2) exchange their positions. The respondents prioritised the evaluation items in the sequence of (2), (1), (3), (4), (6), and (5). It can possibly be stated that the most important criteria of assessing students' programming exercises is by looking into the black-box testing criteria of a program which is particularly referred to the evaluation item (2). In addition, the evaluation item (1) is also necessary to be taken into consideration as it ensures students really understand the syntax of a program. Basically, if a program satisfies evaluation item (2) informally, it can be said that the program should fulfil the evaluation item (3). That is why evaluation item (3) is less important as compared to evaluation item (2). However, it has to be compared to the evaluation item (4) to ensure the program definitely follows the given specifications. Thus, the sequence of evaluation items (2), (3), and (4) does make sense. However, plagiarism and submission date act as complementary aspects that need to be considered in order to offer more reliable assessment results. This is mainly related to the restriction that educators set up for students throughout the course for the learning timeframe.

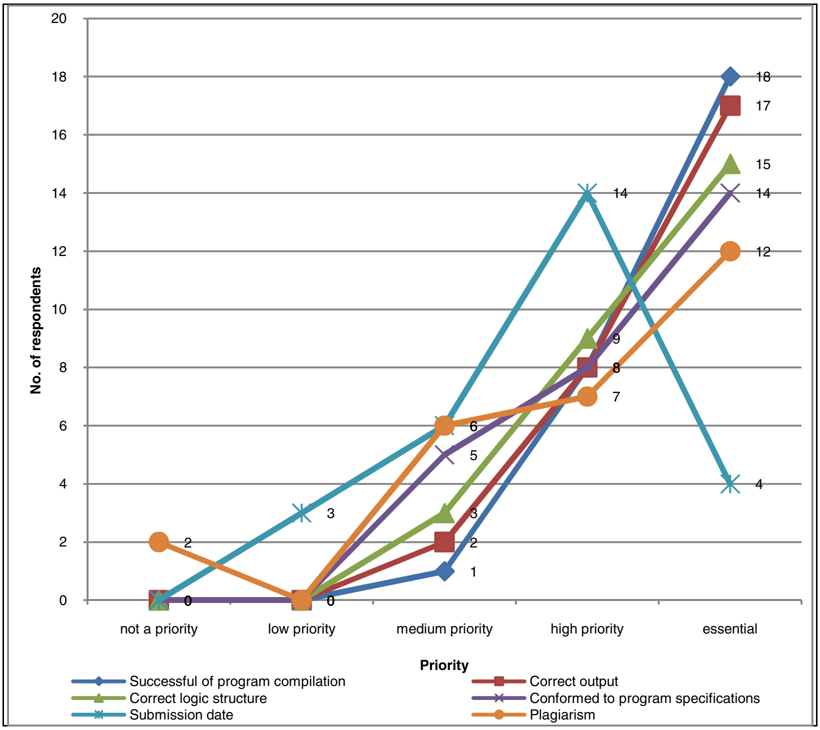


Fig. 9. The frequency of prioritizing evaluation items in programming exercises

Apart from investigating the evaluation items applied in the programming assessment, we also gathered information on the common way educators carried out the process of marking students' programming exercises. Due to the reason that the commercial or research prototype tools to perform automated marking of students' programming exercises are not available, the marking process is in a manual manner. The result in Table 2 supports this point. All respondents stated that they did the marking process manually. On the other hand, some of them complemented it with additional methods such as using a plagiarism tool and/or students should make a presentation for their work.

Table 2. Mode of marking programming exercises

Mode of marking programming exercise	Number of responses
Manually	23
Manually + others	3
Manually + semi-automated tool	1
Semi-automated or fully automated tool or others	0

4 Approaches for Programming Exercises

The practises of programming assessments collectively vary between educators and universities. This relatively focuses on the aspect of the assessment if it has to be carried out manually. Generally, educators rely on some common criteria in assessing certain aspects of quality of student's program regardless the means of assessment carried out. Basically, it relates to the concept of software testing techniques. Thus, the assessment can be done according to whether it needs execution of the program or can be merely based on walk-through or direct inspection of the program code. These are usually so-called static or dynamic assessments respectively. Ala-Mutka [1] and Liang *et al.* [16] provide the details of both assessment approaches. In the subsequent paragraphs, we summarise the approaches.

Dynamic assessment is the most common form of assessment to check that the program operates as the given specifications. It can be done based on merely tests the functionality of the program and/or considering of the internal logic structure of the program code. These are also known as black-box and white-box testing [17] respectively. Black-box testing is yet the most popular strategy employed in programming assessments. For white-box testing, it depends on the criteria that are co-called coverage metrics. These criteria are used to determine the coverage of the program logic and must be executed as least once such as statement coverage, path coverage, branch (or decision) coverage, condition coverage and decision/condition coverage [18]. Both of them usually need a set of test data or inputs to verify whether or not the output produced from student's program consistent to the expected result from the model solution. As such, it is said that the correctness of the program is compared. However, for the higher level programming courses, educators can consider the efficiency of the program as one of the criteria in the assessment such as the running time of the program. Both the correctness and efficiency criteria can be assessed automatically or manually. In addition to automated assessment, students' testing abilities can also be included as a strategy to allow students to design test cases with appropriate test data and test the completeness of their own program prior submitting.

Conversely, static assessment mostly involves an inspection to the program code. In manual manner, the assessment is done by visually work-through the program code. However, it becomes more complicated if it has to be carried out automatically. It requires a kind of extended compiler/interpreter or specific analyser to examine certain static features of the program. As such, taking it in a manual way is the most preferable option used by educators. It can be said that this approach assesses the quality aspect of student's program more thoroughly compared to dynamic assessment. Among the features that commonly look for static assessment are programming errors, programming style, software metrics, design of the program, and some other special features. Programming errors commonly with regard to syntax errors and programming style sees the quality of the program code. If the assessment makes use of software metrics, it can be based on traditional [19] and/or object-oriented metrics [20, 21]. On the other hand, the interface or structural aspect of the program code such as structural similarity analysis can be assessed if it focuses on the design of the program. For the special features, they typically capture certain issues such as plagiarism and capability of providing flexible function such as example keyword search.

Huge class size in computer science or information technology programme and the current practice of programming assessment leads to an extensive workload to educators particularly if it has to be carried out manually. Apparently, manual assessment such as marking printed solutions by hand which is time consuming and requires much effort and attention that are prone to error in any levels of assessment [22]. Therefore, APA has become an important method for grading students' programming exercises and giving feedback to them. Besides, it also improves the consistency, accuracy and efficiency of the assessment [8].

Researches that are mainly related to APA have been of interest to computer science educators since 1960s. Different techniques were employed to judge various aspects in quality of students' programs as above presented. Romli *et al.* [23] summarize the chronology of the studies done by considering the mentioned techniques. Some of the studies developed automated or semi-automated assessment tools as mentioned in Section 1. Instead of providing with a function to automatically assess students' programs, they incorporated an extensive administrative and archive functionalities such as automatic submission, instant feedbacks, running as web based system, supported with multi platforms and tailors to different programming languages.

5 Conclusion and Future Work

In conclusion, educators use different criteria in preparing questions and a solution model for programming exercises as well as the means for evaluating the exercises. The study reveals that a positive testing criterion is necessary to be taken into consideration in preparing both questions and solution model for programming exercises. In addition, inclusion of negative testing criteria in preparing the solution model allows more thorough testing coverage in order to discover more unexpected errors. However, this is particularly related to testing on functional aspects of a program. The main purpose is to provide more efficient and reliable assessment results that require the consideration of in depth program testing coverage. In addition, by integrating both positive and negative testing criteria, educators use a more rigorous testing coverage to judge certain aspects of program quality by considering a program that *does what it is supposed to and not supposed to do* and it *does not do anything that it is not supposed to do*.

In term of evaluation items in marking students' programming exercises, the results show that educators prioritise black-box testing (dynamic testing) criteria as compared to static and structural testing (dynamic testing) criteria. Nevertheless, the highest response was for the static aspect of a program with regard to the evaluation item that influences educators in programming assessment. In overall, we can deduce that both static and dynamic aspects of a program are among the popular and accepted criteria used by educators in marking assessment items. Besides, other additional criteria such as submission date and plagiarism should also be included as they offer a more comprehensive assessment process to produce more accurate results. Unfortunately, educators spend a lot of effort in a marking process as most of the respondents perform the process manually. If there is an automated tool which could offer the desired assessment criteria, then the educators will reduce their workload in the marking process.

As the future work, the results of this study will determine which criteria can be included in our proposed work in APA. Nevertheless, the work will only focus on the dynamic aspect of a program. Typically, most of the existing studies in APA and particularly those that focus on assessing the functional aspect of students' program merely employ the criteria of positive testing rather than negative testing. For a more thorough assessment in the quality aspect of students' programs, we believe negative testing criteria should also be included. This can provide a better program testing coverage which probably cannot be covered by purely depending on positive testing criteria.

Acknowledgement

The authors would like to acknowledge the USM RU Grant: "Development of Variable Strength Interaction Testing Strategy for T-Way Test Data Generation" for funding the related expenses with regard to this paper.

References

1. Ala-Mutka, K.M.: A survey of Automated Assessment Approaches for Programming Assignments. Computer Science Education 15(2), 83–102 (2005)
2. Blumenstein, M., Green, S., Nguyen, A., Muthukkumarasamy, V.: GAME: A Generic Automated Marking Environment for Programming Assessment. In: Proceedings of the International Conference on Information Technology: Coding and Computing ITCC 2004, pp. 212–216 (2004)
3. Choy, M., Nazir, U., Poon, C.K., Yu, Y.T.: Experiences in using an automated system for improving students' learning of computer programming. In: Lau, R., Li, Q., Cheung, R., Liu, W. (eds.) ICWL 2005. LNCS, vol. 3583, pp. 267–272. Springer, Heidelberg (2005)
4. Higgins, C.A., Gray, G., Symeonidis, P., Tsintsifas, A.: Automated Assessment and Experiences of Teaching Programming. Journal of Educational Resources in Computing 5, Article 5 (2006)
5. Auffarth, B., Lopez-Sanchez, M., Miralles, J.C., Puig, A.: System for Automated Assistance in Correction of Programming Exercises (SAC). In: Proceedings of the Fifth CIDUI - V International Congress of University Teaching and Innovation (2008)
6. Truong, N., Bancroft, P., Roe, P.: Learning to Program Through the Web. ACM SIGCSE Bulletin 37(3), 9–13 (2005)
7. Malmi, L., Karavirta, V., Korhonen, A., Nikander, J., Seppala, O., Silvasti, P.: Visual Algorithm Simulation Exercise System with Automatic Assessment: TRAKLA2. Informatics in Education 3(2), 267–288 (2004)
8. Jackson, D., Usher, M.: Grading student programs using ASSYST. In: Proceedings of the 28th SIGCSE Technical Symposium on Computer Science Education, San Jose, CA, pp. 335–339 (1997)
9. Luck, M., Joy, M.S.: A secure on-line submission system. Journal of Software – Practise and Experience 29(8), 721–740 (1999)

10. Baldwin, J., Crupi, E., Estrellado, T.: WeBWork for Programming Fundamentals. In: Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, Bologna, Italy, p. 361 (2006)
11. Joy, M., Griffiths, N., Boyatt, R.: The BOSS Online Submission and Assessment System. ACM Journal on Educational Resources in Computing 5(3), Article 2, (2005)
12. Shukur, Z.: The Automatic Assessment of Z Specification, PhD Thesis University of Nottingham (1999)
13. Sekaran, U.: Research Methods for Business: A Skill Building Approach, 4th edn. John Wiley & Sons Inc., India (2003)
14. IPL Information Processing Ltd., Designing Unit Test Cases (1997), `http://www.ipl.com/pdf/p0829.pdf`
15. Howatt, J.W.: On Criteria for Grading Students Programs. SIGCSE Bulletin 26(3), 3–7 (1994)
16. Liang, Y., Liu, Q., Xu, J., Wang, D.: The Recent Development of Automated Programming Assessment. In: Proceeding of International Conference on Computational Intelligent and Software Engineering, pp. 1–5 (2009)
17. Sommerville, I.: Software Engineering, 7th edn. Pearson-Addison Wesley, USA (2004)
18. Demillo, R.A., McCracken, W.M., Martin, R.J., Passafiume, J.F.: Software Testing and Evaluation. The Benjamin/Cummings Publishing Compony, Inc., California (1987)
19. Tegarden, D.P., Sheetz, S.D.: Effectiveness of traditional software metrics for object-oriented systems, System Sciences. In: Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences, vol. 4, pp. 359–368 (1992)
20. El-Emam, K.: Object-oriented metrics: A review of theory and practice. In: Advances in Software Engineering, pp. 23–50. Springer-Verlag New York, Inc., New York (2002)
21. Xenos, M., Stavrinoudis, D., Zikouli, K., Christodoulakis, D.: Object- Oriented Metrics – A Survey. In: Proceedings of the Federation of European Software Measurement Associations, Madrid, Spain (2000)
22. Jackson, D.: A Software System for Grading Student Computer Programs. Computers and Education 27(3-4), 171–180 (1996)
23. Romli, R., Sulaiman, S., Zamli, K.Z.: Automatic Programming Assessment and Test Data Generation: A Review on Its Approaches. In: Proceeding of 2010 International Symposium on Information Technology (ITSim 2010), pp. 1186–1192 (2010)

Learn with Me: Collaborative Virtual Learning for the Special Children

Nia Valeria and Bee Theng Lau

School of Engineering, Computing, and Science
Swinburne University of Technology Sarawak Campus
nvaleria@swinburne.edu.my, blau@swinburne.edu.my

Abstract. Collaborative learning environment is regarded as stimulating and engaging for normal learners. The main aim of our research is to investigate its effectiveness in assisting the learning of children with disabilities. We developed a prototype, *Learn with Me* and conducted a testing on 6 children who have been diagnosed with cerebral palsy and 7 children who have been diagnosed with autism spectrum disorders. Participants were invited to take part in two tests. Result showed participants learn better with responsive virtual tutor as compared to non-responsive virtual learning.

Keywords: Autism, cerebral palsy, austism, collaborative learning, collaborative virtual learning, emotion, facial expression recognition.

1 Introduction

Communication is one of the important aspects in our daily life. Without communication, people hardly understand other people's thoughts and opinions. Moreover, it is one of the foundations needed in literacy. Nevertheless, many people with disabilities, especially people with cerebral palsy (CP) and autism, found it as a challenge in their lives. Children who have been diagnosed with cerebral palsy and autism mostly have difficulty in developing speech or communication.

Problem faced by CP children mostly is movement difficulty which is caused by the brain injury during or after birth. CP children can have more than one disability, such as speech impairment, hearing impairment, intellectual ability, seizures and other disabilities caused by the brain injury [1], [2]. Disabilities varied from one child to another child. However, studies show that most types of Cerebral Palsy have speech impairment [2].

Speech impairment problem is faced by children with autism as well. Studies show that nearly 50% of children with autism never develop speech [3], [4]. Even though they do not have movement problem, like CP children, but they do have problem in social skills, language, mental, and behavior. Different child may have different level of impairments; therefore autism is called spectrum disorder which affects people differently [5].

Due to their disabilities, children may have some learning difficulties when they come to education. Study shows that some of the children with cerebral palsy and autism have lower academic achievement when compared to normal children [6], [7].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 486–505, 2011.

On top of the speech development problem, they are also slow in understanding what other people express. This also causes problem in their education.

These children require special educators to teach [8]. Special educators have to be extra patient in teaching the slow learner due to their cognitive impairment [8], [9], [10]. The special children require extra time in processing the information compared to the normal children [11]. Moreover, the impairments also cause difficulty in understanding and communicating with others. Educators get discouraged when children make slow improvements in class. Hence, there is a need for having a 1:1 relationship (between the special educator and special child) [12], [13], [14].

Therefore, Collaborative Virtual Learning, CVL has the potential to provide some benefits to the special children as an assistive technology. CVL learning can be used as assistive tool to assist the educator in teaching the children. Through CVL, learners may have companion or virtual tutor to teach and interact to them. CVL offers collaboration during the learning through teacher-learner interaction. CVL fosters 1:1 interaction and repetition to make slow learner master the learning contents.

Many researchers have developed Intelligent Tutoring System (ITS) which adopts the concepts of CVL to help normal students in their learning [15], [16], [17], [18]. Impressive results were obtained to prove that ITS help normal students in their learning [16]. Therefore, the aim of this paper is to present the design and findings of an investigation on the effectiveness of CVL in assisting the special children's learning.

2 Design and Development

In order to solve the problem, we have developed a CVL prototype, *Learn with Me,* for children with disabilities with the intention to assist their daily learning. Through this prototype, children are able to learn through their facial expression to express their thoughts. Educator could use the expression that children made to confirm whether the children truly understand regarding the learning that has been delivered to them. 'Educator' term in the prototype refers to the virtual tutor who is able to response back to the children through the specific conditions (Figure 1).

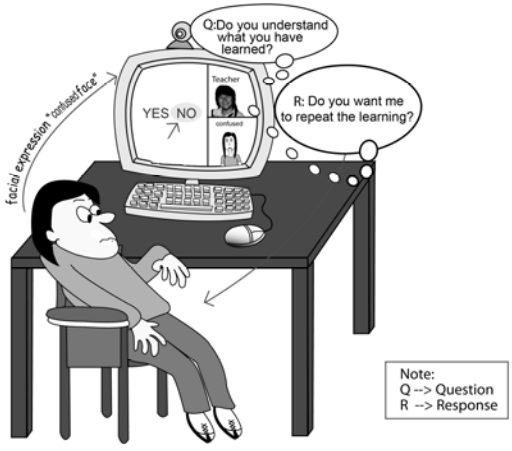

Q:Do you understand what you have learned?
R: Do you want me to repeat the learning?
facial expression "confused face"
YES NO
Teacher
confused
Note:
Q --> Question
R --> Response

Fig. 1. Proposed solution, *Learn with Me*

2.1 Conceptual Modeling of *Learn with Me*

Concept of the program is shown in Figure 2. FacEx-Comm., a facial expression monitoring system, and camera attached are run together with *Learn with Me* during the learning session. The facial expression shown by user is captured through the camera attached on the computer, and sends to FacEx-Comm. to process. The system generates the image that has been captured by compressing the size. FacEx-Comm. matches the generated image with the templates that have been stored previously within user's profile in the local database.

Once the generated image is matched with the template, expression is labeled by the FacEx-Comm. and printed out to the text file. The purpose of printing the expression label to a text file is to allow the *Learn with Me* program to read and obtain the expression label in order to be able to provide a response back to the user. Label that has been obtained is generated and matched with the label within *Learn with Me.* Once it matches, a response will be provided to the user through the virtual tutor. Virtual tutor represents a teacher who is able to deliver learning materials, and give responses back to users according to their expression.

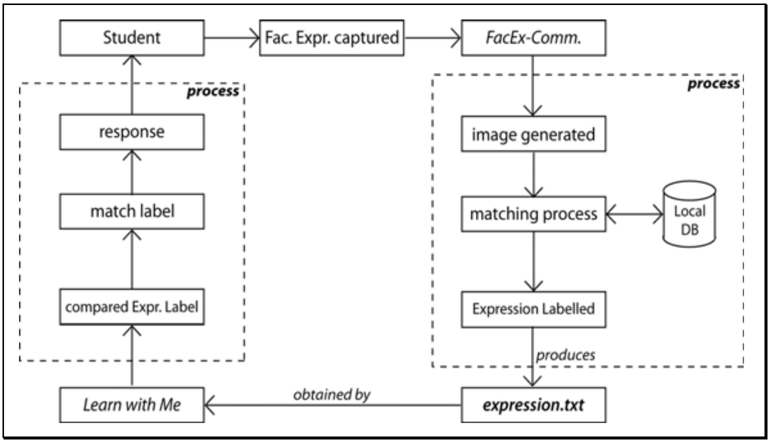


Fig. 2. Concept model of *Learn with Me*

2.2 Components of *Learn with Me*

Figure 3 shows the component diagram of *Learn with Me*. It consists of four layers with small components served different responsibilities in running the program. User interface in ApplicationUI provides multiple languages options for the user to choose. It collaborates with StoryBoardManager to provide options for the users to choose regarding the learning options that they wish to learn.

Service layer responsibles to process the function from the control layer and data is obtained from the data layer. Service layer contains three components and each of the components collaborates with its respective component in the data layer to perform the learning. LearningContentGenerator collaborates with LearningVideoAccessInterface with the intention to display the learning content the screen by gaining the particular related video file. MonitoringGenerator works together with MonitoringAccessInterface to provide and display information regarding the facial expression made by the user. And the last component is VirtualTutorGenerator. It collaborates with ResponsesVideoAccessInterface which responsible to perform the teaching and generate the responses to the user.

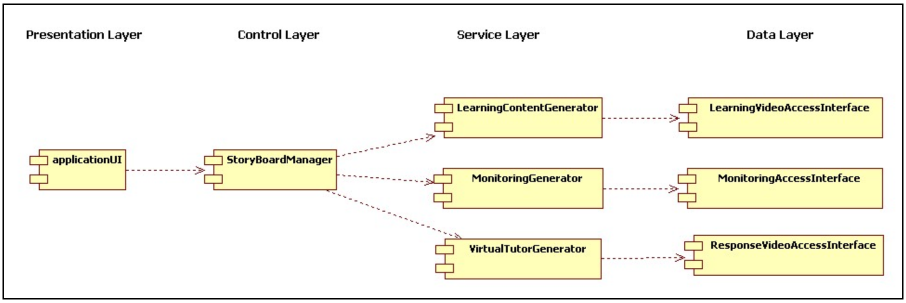


Fig. 3. Components diagram of *Learn with Me*

2.3 Design of *Learn with Me*

2.3.1 Interface

There are 4 components which can be found on the interface at the learning section. The screen in the middle contains the learning contents for the users to learn. At the top of the right side is a placed for displaying the virtual tutor (Figure 4). The purpose of having the virtual tutor is to deliver the learning content to the users. Narration is read by the virtual tutor where at the same time text form is displayed on the screen as well. The reason to have two modes of communication is to train their reading and listening skills since autistic children is poor in reading. Besides reading out the learning materials to the users, virtual tutor is responsible in responding the user according to the facial expression that they made.

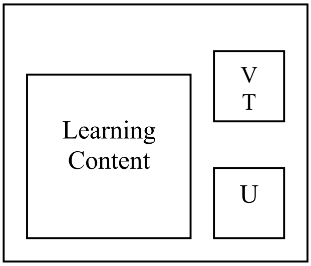


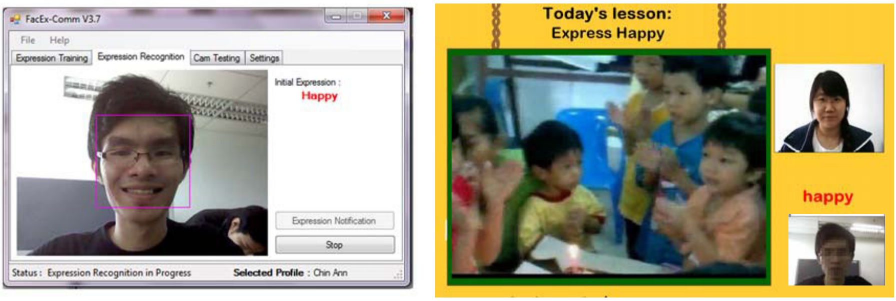
Fig. 4. Screenshot of learning section for *Learn with Me* program

At the bottom of the right side is a placed for displaying the user's facial expression during the learning (Figure 4). Through out the learning, users are able to see themselves on the screen. Lastly is the label of the facial expression which is placed at the top of the user's face (Figure 4). The intention to have these two components on the leaning section is to allow the users to learn their own expression. As we know that, children with autism has deficit in *Theory of Mind*, which include difficult in understanding and recognizing facial expressions. By having this, it hopes that children can learn to recognize other people's facial expression starting from their own face.

2.3.2 FacEx-Comm. Ver.4.0

Facial expression recognition system which is utilized in this learning program is called FacEx-Comm. [19]. It a monitoring system used to detect and recognize human facial expression especially for children with communication and physical disability. The system is built based on the facial expression recognition system consists of three parts which are face detection, facial extraction, and expression recognition or template matching [19].

As it has been discussed previously, in order to identify the expressions shown by users, *Learn with Me* has to work with FacEx-Comm, with the purpose to attain the specific label of expression which is shown during the learning. Label that has been obtained will determine the response that will be given by the virtual tutor. Application needs to be trained before it is used to ensure that the application has the intelligence to recognize the user's expression. Image that has been captured will be compressed to remain the size smaller and classified into individual profile.



Initial Expression : Happy

Status : Expression Recognition in Progress

Selected Profile : Chin Ann

Fig. 5. (left) Screenshot of FacEx-Comm system with "happy" as the label of facial expression, (right) Screenshot of *Learn with Me* which attain the specific label of facial expression from FacEx-Comm

2.3.3 Content

Type of the learning material provided in the *Learn with Me* program is in video format. It is believed video can attract the user's attention more since it is more attractive. Besides, it gives more information compared to image. Content of learning consists of video and singing section. Scenario of the video is created to represent the subject of the learning (*Appendix 3*). As the short video finish, it continues with the singing section related to the current learning. Animation with cartoon as the main character is used in the singing section to attract their intention more. As the users focus on the learning, it is hope that it will help them in answering the assessment questions.

More than that, instead of using avatar for the instructor, *Learn with Me* program facilitates the real person as the virtual tutor by pre-recording all the materials needed in learning and display the video in the learning together with the learning content. Virtual tutor narrates the content throughout the whole learning. The reason behind the use of pre-record video instead of avatar is to give users a familiarity feeling towards to the tutor. Learn together with a person close to the children will improve the learning performance.

2.3.4 Type of Facial Expressions and Responses Given

There are 3 types of expressions to be captured in this learning program, which are confused, bored and happy. According to [17], these are the expressions which have the most occurrences during the learning. Captured image is sent to the recognition system to process if the facial expression of the user does not change within 1 minute. In this program learning, bored and confused facial expression made will generate the responses from the virtual tutor.

Once confused facial expression made, virtual tutor responses with "*Do you understand the question?*". Options of answering the response are in Yes and No type, which is designed to avoid complication in answering the question. If user chooses Yes, program will continue the learning. Whereas, program will play back the learning from beginning if No option is chosen. As for bored facial expression, virtual tutor responses with "*Do you feel bored?*". If user chooses Yes, virtual tutor will ask another question whether he/she wants to stop the learning and continue with the assessment section. If Yes option is chosen, program will stop the learning and direct go the assessment section, else learning will be continued (Figure 6). These two types of responses are the most common responses asked if someone feel confused or bored.

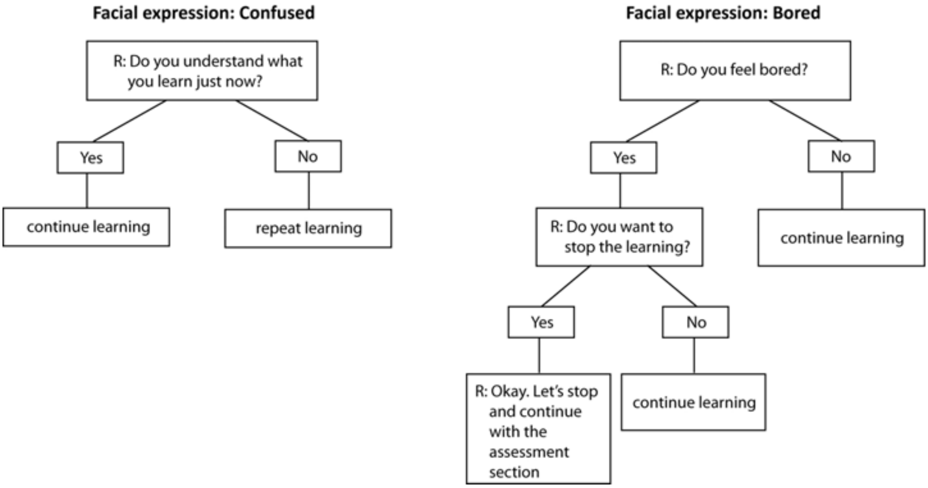

Facial expression: Confused
R: Do you understand what you learn just now?
Yes
No
continue learning
repeat learning
Facial expression: Bored
R: Do you feel bored?
Yes
No
R: Do you want to stop the learning?
continue learning
Yes
No
R: Okay. Let's stop and continue with the assessment section
continue learning

Fig. 6. Flowchart of the responses

2.4 Development of *Learn with Me*

Learn with Me is developed by using Adobe Flash, which is suitable for developing interactive multimedia programs. Flash supports vector and raster graphical, which make the creation of interactive environment to be done easier, compared to others software. Moreover, integration of audio and video are supported by Flash provided by the library code itself. As for the learning content and virtual tutor, they are captured through video recorder in .avi format. It is edited by using Adobe Premiere Pro, software used to edit video. Videos are converted into .flv format by using Adobe Flash CS3 Video Encoder in order to allow the program to read the file.

3 Methodology

There are two research methodologies have been used in this research, which are qualitative and quantitative research, or usually called as mixed-methods designs [20, pp.437], [21, pp.45-48]. According to [20, pp.437], both qualitative and quantitative research is appropriate to be used together. This is useful when doing quantitative research, especially when researcher find a problem in understanding the problem and cannot find the proper solution for it [21, pp.45-48]. Qualitative research can help to identify the problem by using the methodology that qualitative research has such as direct observation, questionnaire, interview, or document review. Quantitative research is used to gather the quantitative data and judge the hypothesis according to the statistic data. Important data can be collected from both methods.

In order to have an in depth understanding on the problem and how to help those targeted children with disabilities, direct investigation to the community settings gives additional information besides reading papers and search through internet. One of the techniques used to gather or collect all the data is by using participant observation at the special education school. It is the social process to assist researchers learning the perspectives held by the special children and teachers [22]. In this method, our participants are approached by the researcher in their own environment instead of having them to come to the researcher [22]. Researcher engaged with participants in order to learn about and get familiar with them.

Knowing the needs of the children is important in this case since they are different from the normal children. Thus, their learning needs and the learning standard are different from the normal children. By involving directly in the community, researcher can have better understanding towards the targeted children. Researcher can participate and observe at the same time [23], [24]. This technique allows the researcher to plan out the appropriate solution before developing the prototype. Moreover, researcher is able to get close to the targeted children to avoid their fear toward stranger so that there is no problem raised when testing is conducted.

Quantitative data is collected through the experimental method applied in conducting the testing. This experimental method is used to fulfill the aim of this research which is to investigate the effectiveness of the Collaborative Virtual Learning towards the targeted children. This method allows data to be analyzed in form of pre and post-test [20, pp.176-181]. In this research, data was collected based on two stages of test, where the 1st test was conducted without running facial expression recognition system and the 2nd test was conducted with facial expression recognition system embedded in the learning.

4 Procedures and Setup

Research conducted in Kuching where the participants were chosen from one of the local schools for disabled children. Before conducting the testing, consent letters were sent out to the children's parents or guardians with the school's principal agreement.

There are some criteria that need to be fulfilled by the participants in taking the test:

1. Participant has to understand brief or simple instructions in English, Bahasa Malaysia, or Mandarin
2. Participant has to be able to focus or give their attention to an object or task for at least 5 seconds
3. Participant has communication problem(s) or incomprehensive speech
4. Participant understands English, Bahasa Malaysia, or Mandarin
5. Participant can control their facial muscles to express emotions
6. Participant has intention to learn
7. Slow in learning

Based on the requirement given, 13 participants were selected to take part in this research. Participants' name remained anonymous in this research to protect their confidentiality.

4.1 Hardware and Software

Participants' expression was captured by using monitoring system (FacEx-Comm). Before capturing, system needed to be trained to distinguish participants' expression. Two USB cameras were attached to laptop/netbook computer, where one was used for capturing the facial expression and another was for displaying participants' face on the screen during the learning. More than that, computer speaker was required in this study as all scenarios were spoken out by the virtual tutor.

4.2 Procedures

In this testing, participants were taken to an unoccupied classroom to prevent any distraction that may affect the result. Participants were asked to point to indicate their answer. Researcher sat next to the participants to give instruction and record the result.

Prior to the study, as has been discussed before in participant observation section, researcher needed to get along with the participants to ensure they get familiar with the researcher. It has come to the stage, where the relationship between researcher and participants were needed because researcher's role in the collaborative virtual learning was as a teacher. Therefore, they needed to understand the participants' responds to develop effective teaching and learning.

There were 2 types of testing being conducted in this research. First testing was conducted *without* running facial expression recognition system (FacEx-Comm).

Second testing was done *with* facial expression recognition system embedded in the learning to achieve the collaboration. In each stage, participants were required to watch the prepared videos, which contain short clip and song related to particular emotion, and answer questions based on the previous video learning. Initially, six (6) emotions were taught in *Learn with Me* such as *happy, sad, angry, scared, surprised,* and *disgust.*

Participants were expected to be able to distinguish each of the emotions after the learning. For instance the video for *Happy emotion* showed a child was celebrating her birthday party, and there were a lot of her friends and family come to celebrate and sing a "happy birthday" song to her. After the celebration they played a game, and everyone felt happy. Right after the video finish playing, it continued with "*If you are happy*" song. At the end of the song, *Learn with Me* showed some questions on the screen. There were 2 questions given in this stage. First question asked in which situation participant will have a happy feeling. Second question asked which face showed the happy face. Each question was given out 3 chances for the participant to answer and it had 2 options for each question. If within 3 chances given, participant was not able to get the correct answer, *Learn with Me* would show the answer directly and gave a message to encourage the participant to try it again next time. Same thing, if participant was able to answer it correctly, "*You are correct, good job!*" message was given. Since there are 6 emotions being taught, 12 questions were prepared to assess the participants.

The procedure to administer the testing was similar for both learning *with* and *without* assistance. Each test was conducted for 3 sessions. Reason for this action is to evaluate the effectiveness of the *Learn with Me* in teaching those children with disabilities. Effectiveness of the collaborative learning is measured by comparing the results between those tests.

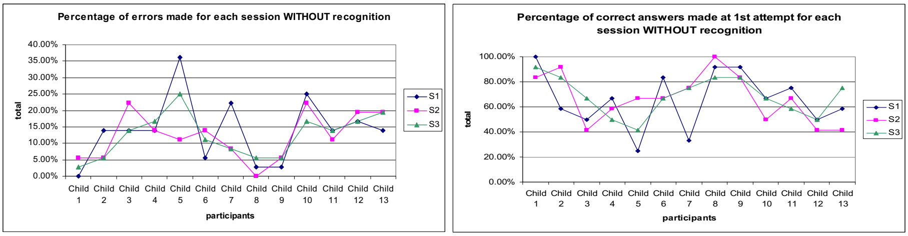
5 Results

A total of 13 participants, 11 males and 2 females with the age from 8 – 17 years old, participated in this research. Six (6) participants were diagnosed as children with autism whereas 7 participants fall under children with cerebral palsy (*Appendix 1*).

5.1 Learning without and with Assistance

Throughout the experiment, result of the tests had been recorded and analyzed. As the data showed in Figure 7a, most of the participants made mistakes in answering the overall questions at learn *without* assistance, which consist of 12 questions. Child 5, 7 and 10 made the most mistakes in the 1st session. Eventually child 7 and 10 made some improvements in 2nd and 3rd sessions. However, Child 5 produced inconsistent results. Minor mistakes were being made in 2nd session, where the number increased again in 3rd session. Child 1, 6, and 13 also made more mistakes in 2nd and 3rd session as compared to the 1st session.

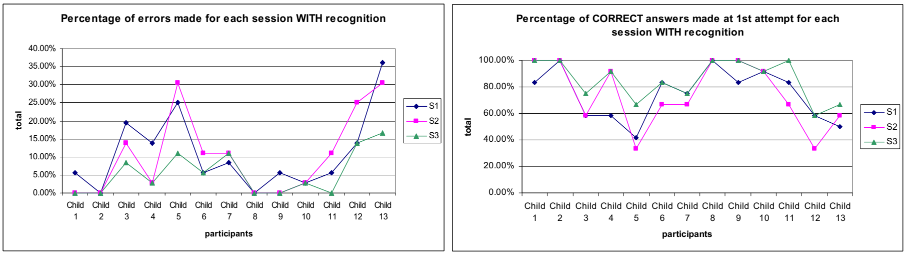
The analysis also showed that, 7 participants get the correct answer below 70% in the first of three attempts given in 1st session (Figure 7b). Child 2, 5, and 7 made some improvements in 2nd and 3rd session. Child 1, 8 and 13 made inconsistent result as they were able to get higher mark in some session. Child 4, 6 and 11 dropped his result at session 2 and 3 compared to the session 1. Overall, children made a little bit of improvement in session 3 compared to the other sessions (s1 x̄ = 65.38%, s2 x̄ = 66.67%, s3 x̄ = 68.59%).



(a) (b)

Fig. 7. (a) Percentage of errors made for each session WITHOUT assistance, (b) Percentage of correct answers made at 1st attempt (chances) for each session WITHOUT assistance

On the other hand, in learning *with* assistance, participants showed some significant improvements as compared to *without* assistance stage (Figure 8a). About 50% of the participants were able to answer the questions correctly without making any errors. Child 5 made great progress in this stage. Lesser mistakes were made as compared to the *without* assistance stage. Most participants get errors below 10% which showed higher improvement. In the last session, child 5 was able to get higher score even though in session 2 total errors were shown higher. Despite of the improvements made, Child 12 and 13 made more mistakes in *with* assistance stage (Figure 8a) as compared to learning without assistance stage. Mistakes made by Child 13 were 50% higher than the previous stage.



(a) (b)

Fig. 8. (a) Percentage of errors made for each session WITH assistance, (b) Percentage of correct answers made at 1st attempt for each session WITH assistance

In Figure 8b, 5 participants were able get 100% correctness in answering the questions at 1st attempt (chances) out of 3 attempts given out to them. This showed a significant result of learning made by the participants towards the effectiveness of the learning program. Graph showed most of the participants made great progress in their learning compare to *without* assistance stage (s1 x̄ = 74.36%, s2 x̄ = 74.36%, s3 x̄ = 85.26%) (Figure 7b). Child 5 made a lot of improvement in getting the correct answer from below 50% to above 60%. More than that, result for Child 2 and 10 showed that they have achieved great assistance in learning *with* assistance compared to learning *without* assistance. It can be seen from the graph shown in Figure 7b and Figure 8b, where result shows that child get more scores which is below 80% in graph Figure 7b, and able to achieve higher scores where mostly is above 80% in Figure 8b.

As a result from both learning *with* and *without* assistance, average errors have been calculated. Difference of the average between both learning can be seen in *TABLE 1*. Most of the participants made lesser errors in the second stage learning. This shows that facial expression recognition in their learning give positive result in their learning. Child 2, 4, 10 and 11 made significant improvements in learning *with* assistance stage (*TABLE 1*). Results showed they were able to answer the questions better than previous stage.

Nonetheless, exception made for Child 13. Data showed child 13 made more mistakes even though learning had been repeated regularly. Child 13 showed inconsistent results in answering the questions. Different time interval of testing between each session which is caused by the long holiday season made the child is not able to remember what they had learnt before. Some possible reasons why this effect where caused are child did not fully understand the learning before, or child could not differentiate the options given. On the other hand, less improvement was made by Child 5. Result showed there was not much difference made in learning *without* and *with* assistance. Throughout the testing, Child 5 was observed and found not paying much attention during the learning. Child 12 didn’t made improvement as well. During the test, child showed interest in using computer. However, less attention is given in the learning compared to the self-camera which was monitoring him on the screen.

Table 1. Result shows improvement according to the errors made from without assistance to with assistance

Participant	Average of errors WITHOUT assistance	Average of errors WITH assistance	Improvement made from WITHOUT learning
Child 1	1.00	0.67	(+)0.33
Child 2	3.00	0.00	(+)3.00
Child 3	6.00	5.00	(+)1.00
Child 4	5.33	2.33	(+)3.00
Child 5	8.67	8.00	(+)0.67
Child 6	3.67	2.67	(+)1.00
Child 7	4.67	3.67	(+)1.00
Child 8	1.00	0.00	(+)1.00
Child 9	1.67	0.67	(+)1.00
Child 10	7.67	1.00	(+)6.67
Child 11	4.67	2.00	(+)2.67
Child 12	6.33	6.33	(+)0.00
Child 13	6.33	10.00	(-)3.67

5.2 Correlation of Each Assistance and Errors in Learning with Assistance

Correlation is analyzed between total of each of recognition that participants shown and errors made in learning *with* assistance. Calculation showed both positive and negative result correlation from overall sessions carried out (*TABLE 2*). There were two significant findings that can be drawn here. “Opt. 2” showed negative result with a quite strong value. Result showed the more participants choose ‘No’ in this option, lesser errors occurred. *Learn with Me* will repeat the learning content if participants choose ‘No’. In addition, “Opt. 4” showed the same result, which is strong negative correlation. In this option, participants showed ‘bored’ expression towards the learning. With the response of ‘Yes’, participants showed much understanding about the learning content. Therefore, lesser errors were made in answering the questions.

Table 2. Correlation between each assistance and errors at learning with assistance

Relations	Overall Corr. (R)
Opt1 / E	0.28
Opt 2 / E	-0.6
Opt 3 / E	0.05
Opt 4 / E	-0.5
Opt 5 / E	0.59

Note:
Opt 1 → Do you understand the learning? , response → Yes
Opt 2 → Do you understand the learning? , response → No
Opt 3 → Do you feel bored? , response → Yes
Do you want stop the learning? , response → No
Opt 4 → Do you feel bored? , response → Yes
Do you want to quit? , reponse → Yes
Opt 5 → Do you feel bored? , response→ No

6 Discussion

Children with disabilities require one-to-one interaction in their learning. However, limited qualified special education teacher are available. It requires high level of patience and skills in teaching or communicating with them due to the disabilities that they suffered. A lot of repetitions need to be done as compared to normal children. Furthermore, communication deficit that they encounter may cause misunderstanding between learner and educator. The children have problem in expressing their thoughts. Hence, this research was carried out to develop a method to assist those special children in their learning. Evaluation of the effectiveness of *Learn with Me* was judged according to the results that participant scored and how the expression monitoring can help in their learning.

A. Participant characteristics

In Figure 8b, graph showed that participants made positive improvement in answering the questions on the 1st attempt given in their learning compared to learning *without* assistance. This shows a great improvement. Previously, in learning *without* assistance, result showed participants need to choose the answer multiple times to get the correct choice. This can be seen from the number of errors they made in their learning (*TABLE 1*).

Each assessment question is given two options to let the participants to choose. By default, if participant made an error in the 1st attempt (chance), he/she could get the correct answer in the 2nd attempt by choosing the alternate answer. However, 76.92% of participants (out of 3 sessions) required 2 or 3 attempts in getting the correct answer. On the other hand, percentage decreased when it comes to learning *with* assistance. Overall, 61.54% of participants required 2 or 3 attempts in getting the correct answer. Decrement of the result showed a positive effect of learning *with* assistance. From the result produced, it was likely that participants may have chosen the answers without understanding the questions given first. Result shows children have weakness in meta-cognitive aspects of self-regulation [25], [26].

Participants with high-functioning (less severe diagnosis) produced better result when compared to participants with low-functioning [11]. For children with cerebral palsy, the main problem that they faced is the hand movement. Most of them faced problem in moving their hand to choose the right answer. Moreover, some participants had problem in controlling their head movement, which caused them not to focus on the screen. Therefore, when it came to the assessment, they could not answer correctly after they missed out some parts of the learning.

Children with autism do not face any problem with their mobility. On the other hand, most of them lack of skill in understanding and recognizing because of the delay of the theory of mind [27], [28]. First type of question is to identify whether participants understand the content that they have learned, while the second type of question is to identify whether participants were able to recognize the facial expression of other people [29]. On first question, participants tend to choose the answer which is more attractive for them. This can be seen from the way they chose the answer, as they always choose the same answer for the first attempt. In addition, study showed more mistakes were found in the second questions. Most of the participants made mistakes in differentiating sad face with scared face, and angry face with sad face.

The similar problem that children with cerebral palsy and autism faced was slowness in learning. Slowness in learning reduces the understanding of the contents which caused errors to increase throughout the learning. Some participants did not want to figure out the correct answer for a question after first attempt. Answers were chosen randomly without having any further thinking. Furthermore, some participants with low-functioning chose the answer without understanding the content. It is merely because the animation graphic for that option is more attractive when compared to the other one. This gave an impact to the learning result.

B. Difficulty level of learning content

According to [30], duration and frequency of the facial expressions can determine the real and fake facial expressions of the users in responding to examinations of pain conditions. In detecting the participants' facial expression, learning with assistance could be used to measure the difficulty of learning content. It can be measured through the occurrences of the responses offered and the answer for that particular response. Responses were offered by responding to participant's facial expression during the learning. According to [17], students may be confused, bored, or excited during the learning.

Throughout the study, overall responses show that learning content regarding *happy* subject is the easiest learning compared to others (*TABLE 3*). This can be proved from the error percentage which was very low, about 5.56%. Study showed that most participants have good understanding towards the learning during the learning time (happy – "Opt 1"). Besides, the high value of "Opt. 4" showed participants' feeling to leave the learning content due to the high understandability towards the content. Response for "Opt. 3" and "Opt. 5" showed that participants has the interest and feeling to learn more about that particular subject.

According to *TABLE 3, sad* subject is considered as the 3rd highest error percentage from overall subject. The higher the number of errors percentage, the more difficult the subject is. This can be judged from the answer of responses given by overall participants. "Opt. 1", which shows the understandability, showed lower result than *happy* subject. Meanwhile, "Opt. 2", which shows the responses for not understand, showed higher result compared to *happy* subject. From these two options, decrease of the ability to understand and increase of the inability to understand, showed the difficulty of the learning. Moreover, response for "Opt. 4 (feeling to leave the learning)" is lower, which means some participants did not have the intention to leave the learning due to the low understandability towards the learning. Result shows the same for *surprised* subject.

If we compared *happy* subject with *angry* subject, the responses ("Opt.1" and "Opt. 2") did not show much difference, which means the learning content was understandable by the participants. However, result of errors in *angry* subject was higher when compared to *happy* subject. The cause of the high error percentage made in *angry* subject was the difficulty that participants faced when choosing the right answer for question number 2, where participants needed to differentiate the right facial expression according the question given. *Sad face* and *angry face* were provided as the options for the *angry* question. 61.5% of total participants chosen sad face as the angry face. This showed that even though learning content was understandable, some of the participants have difficulty applying the learning that they have learned into the question.

Facial expressions can be determined by examining the components parts of facial expression especially on the mouth and eyes [28], for instance, facial expression between happy and sad. The angle of the mouth between both expressions is different. When someone happy, he/she pulls the mouth edge to the back, makes the mouth looks wider. As for sad, a person's mouth tends to be smaller as usual and it looks like a bracket sign '(' which rotate 90 degrees clockwise. When sad and angry face are been placed together, some people can get confused as the facial expression does not merely emphasize on the mouth, but also the eyes. Therefore, if someone does not critically see the whole parts of the face, he/she might get it wrong in determining the facial expressions. As children with CP and autism have difficulty in ToM (Theory of Mind) [5], [28], [31], some of them might be weak in critical thinking which cause them to not be able to differentiate it.

Table 3. Results of responses for each of the subjects in learning with assistance

Subjects	**opt1**	**opt2**	**opt3**	**opt4**	**opt5**	**Errors**
happy	24	2	4	13	16	5.56%
sad	19	7	1	7	9	11.97%
surprised	15	5	3	10	4	10.26%
disgust	13	3	0	9	6	7.69%
scared	15	6	0	9	6	6.84%
angry	22	3	6	8	4	18.38%

Note:
Opt 1 and **Opt 2** represents Confused facial expression
Opt 3, Opt 4 and **Opt 5** represents Bored facial expression

C. Adjustable learning

Study showed that there was a difference of achievement that can be attained by participants between learning *with* and *without* assistance (*TABLE 1*). Most of the participants showed improvement by getting lesser errors compared to previous learning. In learning *with* assistance, *Learn with Me* detects the participant's expressions that showed learning problem that was *confused* and *bored*. Participants' learning would be intervened and attended by the virtual tutor. Through this response, participants were able to continue, pause or terminate the learning by responding to the assistance by the virtual tutor.

However, this could not be done on learning *without* assistance. Participants were not able to control the learning because no 'back' or any buttons given besides the learning content. The reason behind is to get the participants attention while learning, especially children with autism, with the intention that they will not get any distraction. Since there is no buttons or input that can be given, participants need to wait till the learning video finish playing to continue with the assessment part.

Consequently, in the learning content stage, some participants became bored as they had seen the same things previously. Once they felt bored, they would loose their focus and it was hard to get their attention back. Thus, when it came to the assessment, they made a lot of errors / mistakes. Conversely, things can be adjusted in learning *with* assistance. If participants felt bored, virtual tutor would ask them a question whether they want to skip the learning. If they answered yes, *Learn with Me* would leave the learning stage, and go to the assessment directly. The assessment results showed that participants had good result because the prior learning was still fresh in their mind.

7 Conclusions

The aim of this study was to investigate the effectiveness of a CVL design in assisting special children. The findings showed that the use of collaborative learning through expression monitoring was an effective assistance in learning. The results showed positive outcome when compared to non-collaborative virtual learning.

Acknowledgment

We would like to acknowledge the guidance, advice and cooperation from the special school's principal, teachers, and students. We would also like to thank Assoc Prof. Dr Lee Seldon for his advice and feedbacks and Ong Chin Ann for the assistance in the monitoring system.

References

1. Majumdar, R., Laisram, N., Chowdhary, S.: Associated handicaps in cerebral palsy. IJPMR 17(1), 11–13 (2006)
2. Bax, M., Cockerill, H., Carroll-Few, L.: Who needs augmentative communication, and when? In: Cockerill, H., Carroll-Few, L. (eds.) Communication Without Speech: Practical Augmentative & Alternative Communication, pp. 65–71. Mac Keith Press, London (2001)
3. Remus, M.: Autism and school based programming (Strategies for meeting the needs of low functioning autistic children), `http://www.telusplanet.net/public/nremus/communication.htm`, (viewed November 12, 2010)
4. Schneider, E.D.: Communication disorders in children with autism: characteristics, assessment, treatment. In: Gupta, V.B. (ed.) Autistic Spectrum Disorders in Children, pp. 161–174. Marcell Dekker, New York (2004)
5. Adams, J.B., Edelson, S.M., Grandin, T., Rimland, B.: Advice for parents of young autistic children (2004): working paper, `http://www.autismtoday.com/adviceforparents.pdf`, (viewed December 2, 2010)
6. Valente, J.A.: Creating a computer-based learning environment for physically handicapped children, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, USA (1983)
7. Russell, J., Jarrold, C., Henry, L.: Working Memory in Children with Autism and with Moderate Learning Difficulties. Journal of Child Psychology and Psychiatry 37(6), 673–686 (1996)
8. Degreefinders,: How to become a special education teacher, `http://www.degreefinders.com/education-articles/careers/how-to-become-a-special-education-teacher.html`, (viewed January 28, 2011)
9. BecomeaTeacher,: Special education, `http://www.becomeateacher.info/Special-Education.asp`, (viewed January 15, 2011)
10. Meyers, J.:Characteristics of a special education teacher (2001), `http://connected.waldenu.edu/special-education/special-education-teachers/item/1671-characteristics-of-special-education-teacher`, (viewed January 2, 2011)
11. Aleven, V., Koedinger, K.R.: Limitation of student control: do students know when they need help? In: Gauthier, G., Frasson, C., VanLehn, K. (eds.) Proc. 5th Int. Conf. Intelligent Tutoring Systems, pp. 292–303. Springer, Berlin (2000)
12. Fuller, A.: Ten reasons to homeschool your child with special needs (2009), `http://creation.com/images/pdfs/home-school-corner/special-needs/6663ten-reasons-to-hs-your-child-with-special-needs.pdf`, (viewed December 24, 2010)

13. Kilanowski-Press, L., Foote, C.J., Rinaldo, V.J.: Inclusion classrooms and teachers: a survey of current practices. Int. J. Special Education 25(3), 44–56 (2010)
14. Whalen, C., Liden, L., Ingersoll, B., Dallaire, E., Liden, S.: Behavioral improvements associated with computer-assisted instruction for children with developmental disabilities. J. Speech-Language Pathology and Applied Behavior Analysis 1(1) (2006)
15. Alexander, S., Sarrafzadeh, A., Hill, S.: Easy with Eve: a functional affective tutoring system. In: Workshop Motivational and Affective Issues in ITS. 8th Int. Conf. ITS 2006, pp. 38–45 (2006)
16. Sarrafzadeh, A., Hosseini, H.G., Fan, C., Overmyer, S.P.: Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems. In: Proc. 3rd IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2003), p. 336 (2003)
17. Whitehill, J., Bartlett, M., Movellan, J.: Automatic facial recognition for intelligent tutoring systems. In: Proc. CVPR 2008 Workshop Human Communicative Behavior Analysis (2008)
18. Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., Bigdeli, A.: See me, teach me: facial expression and gesture recognition for intelligent tutoring systems. In: Proc. IEEE Int. Conf. Innovations in Information Technology (IIT 2006), pp. 1–5 (2006)
19. Ong, C.A., Lu, M.V., Lau, B.T.: A Face Based Real Time Communication for Physically and Speech Disabled People. In: Lau, B.T. (ed.) Assistive and Augmentive Communication for the Disabled: Intelligent Technologies for Communication, Learning and Teaching, IGI Global Publishing (submitted for publication)
20. Springer, K.: Educational Research: A Contextual Approach. Wiley, USA (2010)
21. Breach, M.: Dissertation Writing for Engineers and Scientists. Pearson Prentice Hall, UK (2009)
22. Mack, N., Woodsong, C., MacQuen, K.M., Guest, G., Namey, E.: Module 2 – Participation Observation. In: Qualitative Research Methods: A Data Collector's Field Guide, Family Health International, USA (2005)
23. Laurier, E.: Participant observation. In: Clifford, N., Valentine, G. (eds.) Research Methods in Human and Physical Geography, Sage, London (2003)
24. Jorgensen, D.L.: Participant Observation : A Methodology for Human Studies. Sage Publications, London (1989)
25. Vockell, E.L.: Educational psychology: a practical approach(1995-2001), `http://education.calumet.purdue.edu/vockell/edPsybook/Edpsy7/edpsy7_meta.htm`, (viewed December 23, 2010)
26. Dawson, P., Guare, R.: Executive Skills in Children and Adolescents: A Practical Guide to Assessment and Intervention, 2nd edn. The Guilford Press, New York (2010)
27. Perner, J., Frith, U., Leslie, A.M., Leekam, S.R.: Exploration of the autistic child's theory of mind: knowledge, belief, and communication. Child Development 60, 689–700 (1989)
28. Ryan, C., Charragáin, C.N.: Teaching emotion recognition skills to children with autism. J. Autism and Developmental Disorders 40(12), 1505–1511 (2010)
29. Fabri, M., Moore, D.J.: The use of emotionally expressive avatars in collaborative virtual environments. In: Proc. Symposium Emphatic Interaction with Synthetic Characters, at Artificial Intelligence and Social Behavior Convention 2005 (AISB 2005). University of Hertfordshire, UK (2005)
30. Hill, M.L., Craig, K.D.: Detecting deception in pain expressions: the structure of genuine and deceptive facial displays. Pain 98, 135–144 (2002)
31. Falkman, K.W., Sandberg, A.D., Hjelmquist, E.: Theory of mind in children with cerebral palsy and severe speech impairment. Göteberg Psychological Reports 34(2), 1–16 (2004)

Appendix 1

Children	Diagnosis	Sex	Age	Language	Learning Ability	Problem faced
Child 1	*CP and Spastic Quadriplegia*	*M*	13	Malay, English	Good in understanding, has high intention to learn	mobility problem, incomprehensive speech
Child 2	*CP (Spastic Quadriplegia)*	*M*	11	English, Malay	Good in understanding, has high intention to learn	Not be able to speak, walking problem
Child 3	*CP both aphyxia - quadriplegia CP - low epilepsy*	*M*	10	Mandarin, Malay, English abit	Moderate in understanding, Slow in learning	simple words like kakak, mum (doesn't have much speech), mobility problem but still can walk
Child 4	*CP with mild learning difficulties*	*F*	12	Malay	Moderate in understanding, Slow in learning	mobility problem, incomprehensive speech
Child 5	*CP*	*M*	13	Mandarin, Malay	Poor in understanding, very slow in learning	mobility problem but still can walk, no speech, hands movement problem
Child 6	*CP*	*M*	10	Mandarin, Malay	Good in understanding, Moderate in learning	mobility problem, unable to control his movement, incomprehensive speech
Child 7	Hyperactive Disorder and Speech delay	M	11	Iban, Malay, English	Moderate in learning and understanding	imitate what other people say, incomprehensive speech
Child 8	ADHD (Autistic Deficit Hyperactive Disorder)	M	10	English, Malay	Good in learning and understanding	incomprehensive speech
Child 9	Autistic and Learning Difficulty	M	17	Mandarin, Malay	Good in learning and understanding	incomprehensive speech
Child 10	Hyperactive Disorders (Autistic)	M	8	Mandarin, English	Slow in learning, Poor understanding	Always repeat what other people say, and talk something unrelated more than once
Child 11	Autistic	M	8	Mandarin, Hakka, Malay	Slow in learning, moderate in understanding	Always repeat what other people say
Child 12	Autistic	M	10	English, Mandarin	Slow in learning, Moderate in understanding	Incomprehensive speech, has interest in computer, lack of focus
Child 13	ADHD, intellectual disability, cognitive and speech delayed	F	9	Chinese, Malay	Slow in learning, Poor in understanding	no speech (communicate use gestures & facial expression)

Appendix 2

Questions in *Learn with Me*

Subjects	Questions	Option 1	Option 2
Happy	1. In which situation you will feel happy?	Animation where a boy been slapped.	A boy receives a present.
	2. Which picture shows the scared face?	Picture shows disgust face.	Picture shows happy face.
Sad	1. I which situation you will feel sad?	Animation where a boy been slapped.	Animation where a boy sees a worm.
	2. Which picture shows the sad face?	Picture shows sad face.	Picture shows fear face.
Disgust	1. In which situation you will feel disgust?	Animation where a boy sees many worms.	Animation where a boy sees rainbow.
	2. Which picture shows the disgust face?	Picture shows surprised face.	Picture shows disgust face.
Surprised	1. In which situation you will feel surprised?	Animation where a boy sees something pops up from a box.	Animation where a boy plays kite.
	2. Which picture shows the surprised face?	Picture shows surprised face.	Picture shows happy face.
Scared	1. In which situation you will feel scared?	Animation where a boy watches horror movies.	Animation where a boy plays kite.
	2. Which picture shows the scared face?	Picture shows happy face.	Picture shows scared face.
Angry	1. In which situation you will feel happy?	Animation where someone takes your thing.	Animation where a boy sees a worm.
	2. Which picture shows the scared face?	Picture shows angry face.	Picture shows sad face.

Appendix 3

Learning content

Scenario for Happy

Today is Jacky's and Nana's birthday. The teacher and their friends are celebrating their birthday. (start singing song)

After celebrating they play a game...and all of them feel happy.

Scenario for Sad

Danny is happy and playing around with his new toy. But suddenly, someone take his toy and he starts crying. He feels sad because his toy has been taken by others.

Scenario for Surprised

Annie and her brother are trying to build a tower. Once it has done, suddenly the tower collapses. Annie feels surprised and her brother is laughing at her because of that.

Scenario for Disgust

There are many worms on the plates. When Annie’s uncle catches the worm, Annie feels disgust and keeps run away.

Scenario for Scared

It happens at night. Annie is watching horror movie alone. Suddenly, the ghost in the movie comes out and she feels very scared. She runs away and almost cries because of scared.

Scenario for Angry

It is during the lunch time. Anna takes Danny's food and Danny is trying to take his food back. However, he loses to Anna. He does not manage to get his food back. Anna, then eats the food. Danny starts to cry. He hits Anna so many times because he feels angry. Anna laughs at him and starts to tease him.

Potential of Using Virtual Environment for Aircraft Maintenance Learning System in Making Tacit Knowledge Explicit

N.S. Jamain and Z.M. Kasirun

Software Engineering Department, Faculty of Computer Science and Information Technology (FCSIT), University of Malaya, Kuala Lumpur, Malaysia
suhaila@fsktm.um.edu.my, zarinahmk@um.edu.my

Abstract. This paper presents an early result of developing a virtual environment for Aircraft Maintenance Learning System in visualizing tacit knowledge to make it explicit. The idea of visualizing tacit knowledge is meant to facilitate efficiently trainers who are dealing with maintenance tasks every day. This virtual environment will provide three principle steps of training session which are to (1) introduce, (2) mentor, and (3) practice. By relying on the Model for Inspection Training Program Development in General Aviation, we emphasize two main activities (training methodology, program evaluation), which all are working with 3-Dimensional graphical images in representing the knowledge; so called VAMLS. VAMLS is associated with Rule-Based engine that used to store and manipulate the collection of knowledge to be interpreted via Role Playing Game – a game that allows learners to presume the role of their 'actual' characters by taking in the responsibility for acting out their role. Therefore, it is anticipated that this tool will help learners in optimizing Tacit Knowledge thus can maintain the standard (equal) knowledge within the process of maintenance learning.

Keywords: Virtual Learning Environment, Role Playing Games, Visualizing Tacit Knowledge, On Job Training, Computer-Based Training, Aircraft Maintenance Learning.

1 Introduction

In the environment of aircraft maintenance training, the involvement of both experience and understanding of the people in the society and the information artifacts available within the organization are the key success in obtaining the greatest value of the knowledge delivered. Knowledge can be categorized into two different types which are tacit and explicit. Currently, the importance of tacit knowledge in terms of how it can be converted into explicit in organizational learning and innovation has become the focus of substantial attention.

According to Watanuki and Kojima (2006) in their model of conceptual for SECI (Socialization, Externalization, Combination, Internalization), they have identified

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 506–516, 2011.

that by using a process so called externalization; it could make the process of transforming tacit knowledge to explicit possible. In this process, the conversion taking place through conceptualization and elicitation by collaboration with others. The externalization happens when there is a conversion or dialog among the team members, and they could respond on the questions asked or the elicitation stories; or in other means, this procedure involves users whom are using a high-level skill transfer system, and they are gaining explicit knowledge through technical document, technical data and so on.

The concept of virtual environment for maintenance learning is relevant for the learners to capture implicit knowledge that has been sent to them while having classroom lectures as well as their practical observation session. This is proven based on the hierarchy of a scientific model that demonstrates the relationship between explicit knowledge and tacit knowledge by Bogue (2006). There are four criteria for making tacit knowledge explicit, which are:

- Step1: Conduct observation to develop experience – The first step requires the expert and novice to determine subjectively the presumption based on their observations; and this illustrates the problem with converting knowledge from tacit to explicit.

- Step2: Hypothesize the observation –When they asked to create an explicit perspective on the presumption, the hypothesis is then erected based on their creativity.

- Step3: Test the hypothesis – The hypothesis is tested to get to the best proper practices, the rules, the guidelines and techniques which lead to good software development.

- Step 4: If incorrect, re-test the hypothesis – The explicit knowledge becomes a periodic set of events that support the underlying observations. The hypothesis is said to be true if the hypothesis gives no real different in the outcome with the existing presumption; as a result tacit knowledge can be converted into explicit. Otherwise, re-test the hypothesis.

The criteria mentioned above were also used in creating a virtual environment of training whereby observation plays a very critical role in ensuring tacit knowledge can be safely converted into explicit. To make this thing possible, visualization is used to describe all those ambiguous knowledge to be understood better. This process is highly recommended by the expert-party of the aircraft maintenance whereby based on the interview which was conducted with the aircraft maintenance expert, they were facing problems in explaining certain parts of the aircraft, and learners easily confused in defining parts of which the experts were referred to (during lecture).

2 Method

In order to liberate some issues encountered while capturing Tacit Knowledge, we proposed a solution which might help learners in gaining systematic and equal knowledge. According to Mulzoff (1990), visual inspections skill can be effectively taught using representative photographic images which show a wide range of conditions and provide immediate feedback on the trainee's decision. Thus, our Virtual Aircraft Maintenance Learning System (VAMLS) will make use the research finding by Mulzoff to support interactive learning in Multimedia.

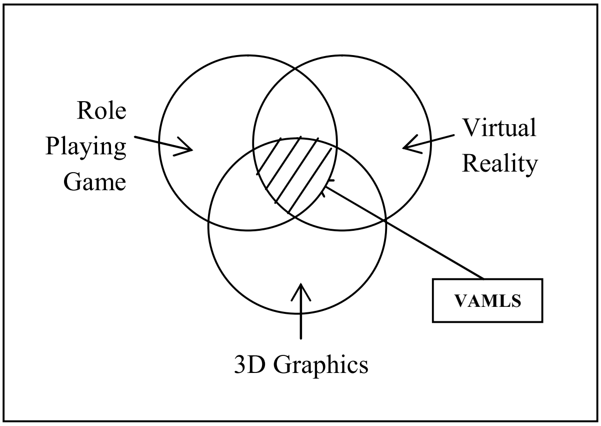


Fig. 1. VAMLS is a subset of RPG, VR and 3D Graphics

Even though there are so many Computer-Based Training System (2D Multimedia Graphical System, Virtual Reality System, Intelligent Tutoring System) are being used in the industry nowadays, VAMLS caters a low cost yet effective tool that can be used for both experts and learners to bring inside themselves in the real scenario of the aircraft maintenance works. The concept of Role Playing Game (RPG) that attached with VAMLS enables learners to be given a set of scenario that corresponding to the real environment of the aircraft maintenance issues. This approach is on the other hand would let trainees to assume the role of their 'actual' character by taking in the responsibility by means of acting out their role through a decision-making process. This method is in fact may help learners in building their character development. As compared to the Virtual Reality (VR) type of a system, VAMLS can alternatively replace the characteristic of VR (realistic, interactive and stereoscopy) by combining its 3D and RPG characteristics, as shown in Figure 1 above.

Other than that, based on the Online Continuum Illustrating Efficacy of Simulation for Learning by Nick Van Dam (2003) and Egdar Dale (1969) shown in Figure 2, most people could remember what they do (90%), and what they write (80%) which are still narrated from the idea of learning for experience. Thus, it shows that learners gain optimum knowledge via simulation or games since these two methods enable learners to look and feel the curriculum informally. However, while working on the technical jobs, mimicking the real world environment works best if the training is conducted in an offline mode. This is possible if a 3D-based inspection environment can facilitate in conducting controlled studies offline and in understanding human performance in aircraft inspection. Even if it is in offline mode, the decision making should come together towards the learning processes. This is added with the rule-based technique whereby it plays as storage for the developed system so that learners may have immediate information or feedbacks for what they did while the system is continuously 'learning' as trainees carrying on their job virtually. Therefore, in general an intelligent machine that is expected can replace human beings or experts in sharing an equal knowledge is a good idea in making knowledge explicit.

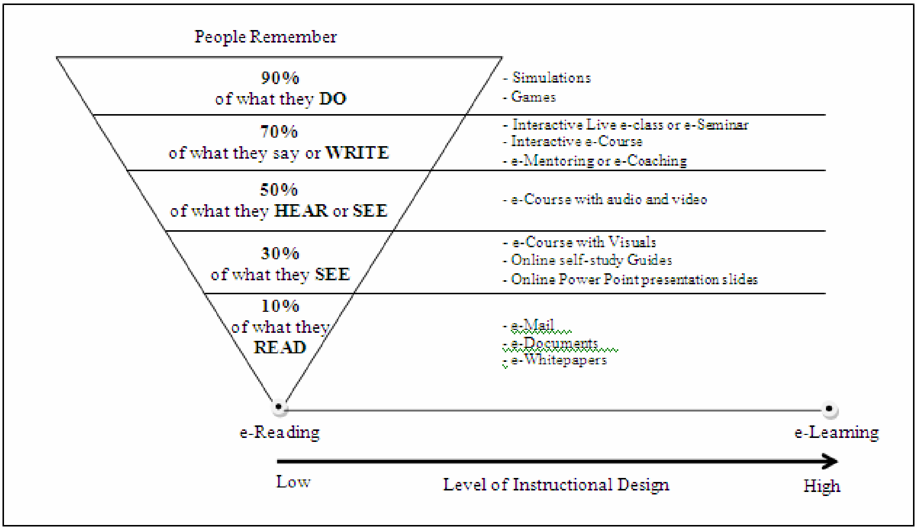

People Remember
90% of what they DO
70% of what they say or WRITE
50% of what they HEAR or SEE
30% of what they SEE
10% of what they READ
- Simulations
- Games
- Interactive Live e-class or e-Seminar
- Interactive e-Course
- e-Mentoring or e-Coaching
- e-Course with audio and video
- e-Course with Visuals
- Online self-study Guides
- Online Power Point presentation slides
- e-Mail
- e-Documents
- e-Whitepapers
e-Reading
e-Learning
Low
Level of Instructional Design
High

Fig. 2. Online Continuum Illustrating Efficacy of Simulation for Learning

In aircraft maintenance industry, observation served as a catalyst in all cases in which they normally use On the Job Training (OJT) as the approach in spreading tacit knowledge. OJT is a technique whereby people are formed into several groups and they meet up together as small gatherings or as break-outs of large meetings which offer many opportunities for creative, flexible interchange or ideas and lively participation; at the same time includes the observation technique which participants are freely observe and comments their thoughts based on what they have seen.

This work is therefore aims to address the problem of how tacit knowledge can be converted into explicit via visualization. In this paper, we introduce our VAMLS conceptual model which is gained by combining the development structure of a Model for Inspection Training Program Development in General Aviation (ITPDGA) with Bogue's Hierarchy of a Scientific Model to make tacit knowledge explicit and its perspectives. Further, the architectural study of the proposed VAMLS is presented, and finally the conclusion along with some suggestions for future works.

3 Results and Discussion

In developing VAMLS, there is a possibility of combining Bogue's Model and ITPDGA, as in ITPDGA, the middle process of its Development Structure can be applied in Bogue's Scientific Model of Tacit to Explicit Knowledge. Based on this arrangement, we have gained the structure of VAMLS that possibly could be used in order to achieve the main objective of this paper; as shown in Figure 3 below.

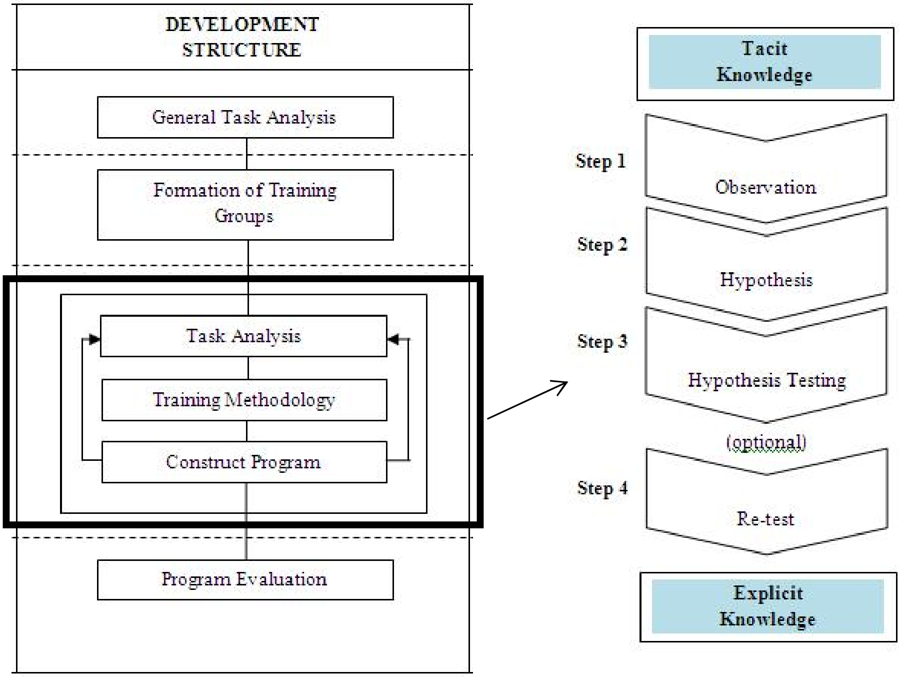


Fig. 3. Applying Bogue's Hierarchical Scientific Model into the middle process in the Development Structure of ITPDGA

By referring to Figure 3 above, observation acts as the key resources in supplying knowledge to the learners. Based on the observation, learners will come out with their own hypothesis and test to verify the accuracy of their hypothesis. In aircraft maintenance industry, most of the comprehension studies applied in the industry is thru observation (Z.M.Kasirun et. al, 2010). So, in explicating tacit knowledge in aircraft maintenance, VAMLS serves as a middle point of this tacit and explicit knowledge. Therefore, the environment of VAMLS based on Bogue's and Development Structure of ITPDGA is shown as Figure 4 below.

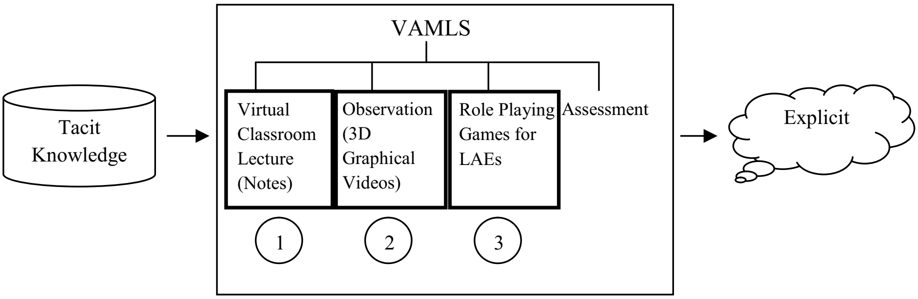

Tacit Knowledge
VAMLS
Virtual Classroom Lecture (Notes)
Observation (3D Graphical Videos)
Role Playing Games for LAEs
Assessment
1
2
3
Explicit

Fig. 4. VAMLS as the middle point (transmitter) in tacit to explicit transformation

3.1 Virtual Aircraft Maintenance Learning System (VAMLS)

Virtual Aircraft Maintenance Learning System (VAMLS) is proposed by having some major enhancement of the existing system; General Aviation Training System (GAITS). The main different between GAITS and the proposed VAMLS is that, VAMLS is fully make use of the OJT concept which allows observation as the key step; and this (in site training) on the other hand is the most important thing of being an LAE Trainees before they are all ready with the real environment of the aircraft industries.

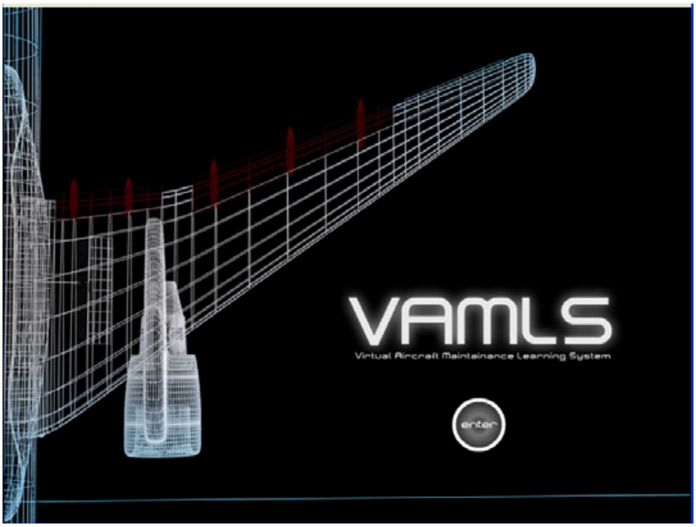


Fig. 5. The VAMLS

The strongest characteristic of this VAMLS as compared to GAITS is, it is using a 3-Dimentional (3D) graphical modeling whereby it contains better and realistic images, backgrounds, and environment. Other than giving scenario, it is also affixed with a Game-likely Learning Tools (GLT) with the intention that trainees can experience at least a minimum authentic environment of the aviation, which is so called RPG.

By having RPG attached in the system, this can help learner to understand better as they will be exposed to the real situation of a Licensed Aircraft Engineer (LAE).

3.2 RPG Infrastructure

Since RPG dealt so much with mind and imagination, it can be said that RPG helps in exercise thinking, reacting, verbal and mathematical skills, and teamwork that can be accomplished while everyone remains sitting. It is a game that can foster ones creativity as gaming requires an active imagination and the ability to think on one’s feet.

Other than that, RPG also teaches problem solving whereby users will be given unexpected situation from the game, so they will just solve the issues given by relying on their problem-solving skills. It may lead to a great way to learn etiquette for the users in such a way that users may listen when others are speaking, take turns during play, assist new players in learning the ropes, and generally try to make sure everyone else at the scenario is having a good time

3.3 VAMLS Architecture

When discussing about the architecture, VAMLS can be embedded into a web-based application or it can be a standalone system and use it in offline mode. It requires a set of rules that can be used as the information storage; and all the information will be executed in terms of graphical images which added with some movements and effects to produce a game. Figure below shows the architecture of VAMLS in both embedded and standalone environment views.

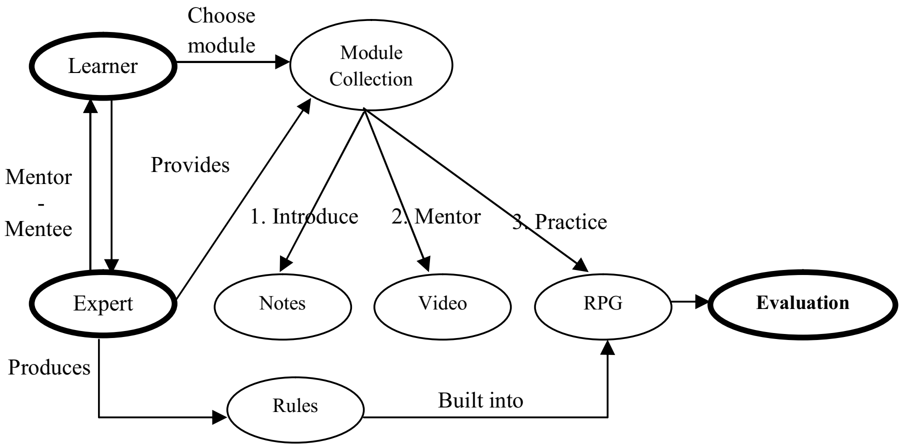


Fig. 6. VAMLS Architecture

Based on Figure 6 shown above, the system starts with a learner having mentor-mentee session with expert via live conversation conference which available in the embedded version of the system. While having this session, learner may raise their thoughts (if any) while the experts can share any tips that might useful while conducting the job routine. When it comes to the standalone (offline) version of this VAMLS, learner may just start choosing their modules available in the system library whereby all those materials were given by the expert and they have been translated into a meaningful version of knowledge (rules, 3D graphics).

There are three main elements in representing VAMLS in ensuring the system really practical in facilitating learners in Aircraft Maintenance job. It starts with Introduction whereby learners are exposed with the interactive notes for the sake of attracting learners in enduring their learning. Next, the virtual mentoring session

(Mentor) that allows learners to see and gain some tips from the video presented by the expert itself. In this part, learners are able to evolve their views in how thing works for each of the module applied.

After all, Practice mode applied which in this VAMLS we used RPG that built in with rules gained from the experts or manuals. In this RPG, learners may have a chance to really make use of the jobs given to them in the future. They will be given any issues related to the module chosen and their job is to solve the issues based on their understanding from step 1 and 2. At the end of the learning session, all those activities will conclude by an Evaluation. In this part, learners knowledge is measured by counting the score collected from the Practice session. If learners managed to resolve the issues accordingly, score will be given. Otherwise they need to retake this session until they succeed.

Based on this, we can say that if learners are able to solve issues given adequately from the VAMLS, and manage to apply the same thing in their actual practical assessment, it means tacit knowledge given from the experts are transformed into explicit properly, or else it can be said there are noises (destructions) in the middle of transformation process.

4 VAMLS Screenshot

Figure 7 shows modules available in the system. As the system meant for testing to measure the efficiency, there are two modules available which are tire and wing; and these followed by a practical assessment (game-based).

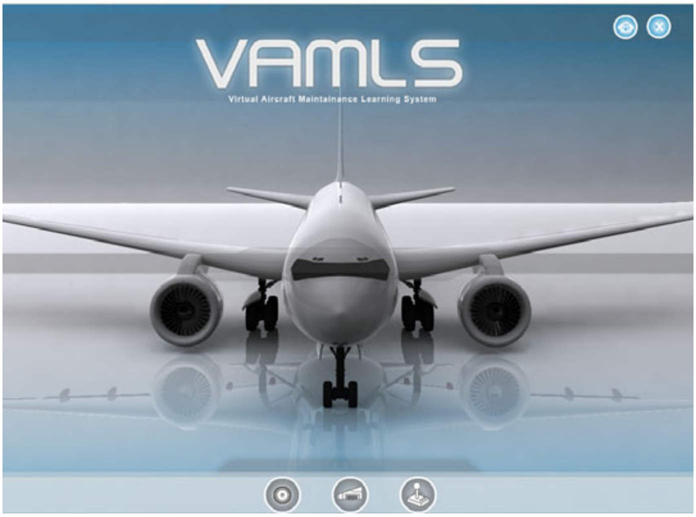


Fig. 7. The main menu of the VAMLS

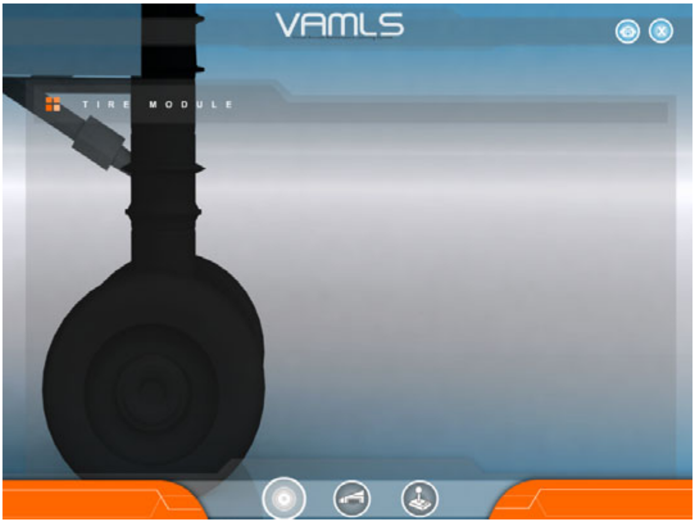
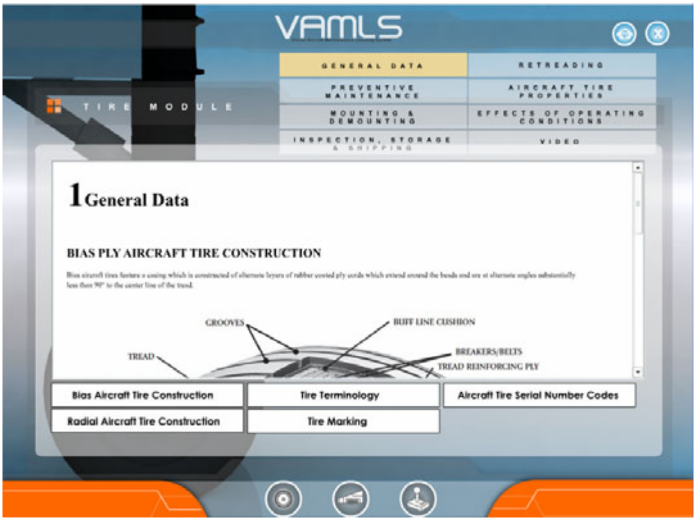


Fig. 8. Tire Module

When the tire symbol at the bottom clicked, it will bring user to the Tire Module part, whereby it shows all the topics available under this module. All these are shown in Figure 9 below.



1 General Data

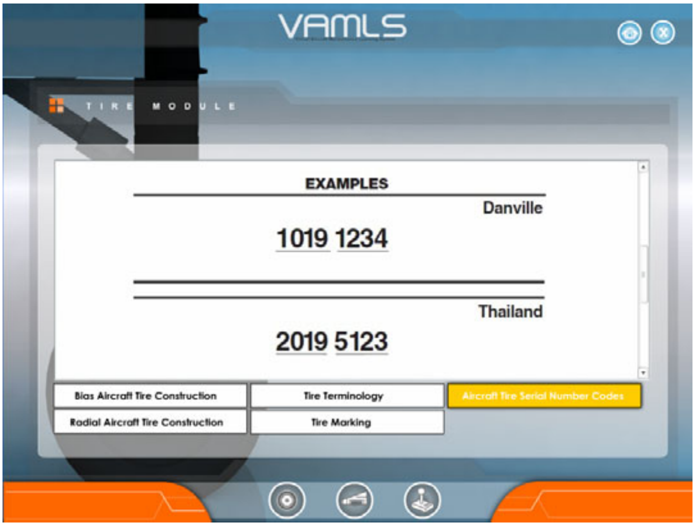
BIAS PLY AIRCRAFT TIRE CONSTRUCTION

Bias aircraft tires feature a casing which is constructed of alternate layers of rubber coated ply cords which extend around the beads and are at alternate angles substantially less than 90° to the center line of the tread.

GROOVES
BUFF LINE CUSHION
TREAD
BREAKERS/BELTS
TREAD REINFORCING PLY
Bias Aircraft Tire Construction
Tire Terminology
Aircraft Tire Serial Number Codes
Radial Aircraft Tire Construction
Tire Marking

Fig. 9. Topics available for the Tire Module

For each of the topic displayed on the screen, all of the topics are again separated into sub topics; as shown in Figure 10. Most of the materials contained in this system are clickable and image-based according to the scenario applied in the industry.


VAMLS
TIRE MODULE
Bias Aircraft Tire Construction
Tire Terminology
Aircraft Tire Serial Number Codes
Radial Aircraft Tire Construction
Tire Marking

EXAMPLES

Danville

1019 1234

Thailand

2019 5123

Fig. 10. Sub topics for the general data of Tire Module: Aircraft Tire Serial Number Code

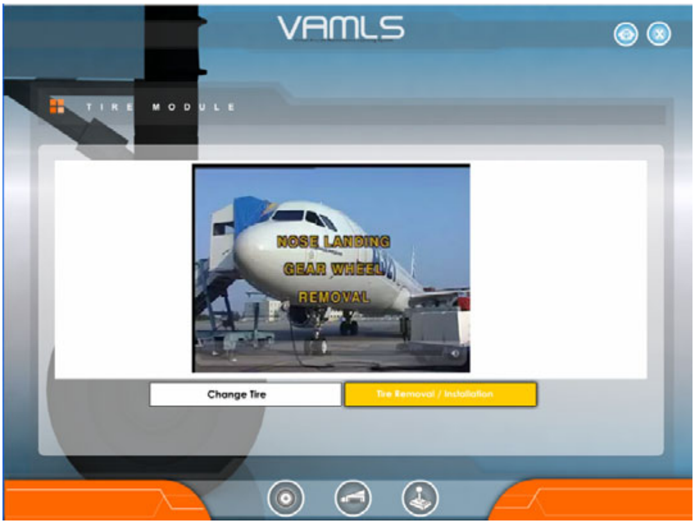


Fig. 11. Videos (OJT)

At the end of the lesson, there is a part whereby learners can learn from the videos that shows the procedure of doing maintenance job based on the module chosen. This is attached in order to fulfill the concept of OJT that applies in the aircraft maintenance industry.

5 Conclusion

We began this research by developing a VAMLS prototype which experimented with the community in MIAT (Malaysia Institute of Aviation Technology), whereby lecturers as the experts and students as the learners. This is done to examine the extent in which this system really works to help ease their problems. By relying on the three most important rules presented in the system earlier (Introduce, Mentor, Practice), we have shown that (indirectly) what are the aspects that we should stressed on in monitoring user's behavior in understanding tacit knowledge given later on. Thus, to set the features that have been designated by the General Aviation, VAMLS are developed according to ITPDGA for the sake of liberating issues encountered while capturing Tacit Knowledge; while at the same time proving that the use of simulations and visualizations can reduce or overcomes noises in the transformation process. However, in future this system can be enhanced by allowing network sharing activity such as Networked VAMLS application that allows for sharing virtual environments or RPG network.

References

[1] Watanuki, K., Kojima, K.: VR-Based Knowledge Acquisition on Job Training for Advanced Casting Skills. In: Proceedings of the 16th International Conference on Artificial Reality and Telexistence–Workshops (ICAT 2006) (2006)

[2] Kasirun, Z.M., Jamain, N., Ahmad, R., Nasir, M.H.N.M., Khamis, N., Abd Latif, B.: An Investigation on Tacit Knowledge Transfer Among Aircraft Engineer: Malaysia Experience. In: Proceedings of The 2nd Int. Conf. Software Engineering and Data Mining. Chengdu China, pp. 197–202 (2010) (Non-ISI/Non-SCOPUS Cited Publication)

[3] Dam, N.V.: The E-learning Fieldbook: E-learning. Editorial Mc Graw Hill, USA (2004)

[4] Mulzof, M.T.: Information Needs of Aircraft Inspectors. Human Factors Issues in Aircraft Maintenance and Inspection. Information Exchange, and Communications, FAA, Office of Aviation Medicine, 79–84 (1990)

[5] Bogue, E.G.: The Context of Organizational Behavior: A Conceptual Synthesis for the Educational Administrator, Graduate Faculty, Memphis State University, Memphis, Tennessee (2006)

[6] Sadasivan, S., Gramopadhye, A.K.: Technology to support inspection trainin. In: the General Aviation Industry: Specification And Design. International Journal of Industrial Ergonomics 39(4), 608–620 (2009)

Trainees' Competency Based-Assessment Methods in Cyber Forensics Education or Training Programmes – A Review

Elfadil Sabeil[1,2], Azizah Bt Abdul Manaf[1,3], Zuraini Ismail[1,3], and Mohamed Abas[1,2]

[1] Universiti Teknologi Malaysia,
[2] Faculty of Computer Science & Information Systems,
[3] Advanced Informatics School, JLN Semarak, 54100, Kuala Lumpur, Malaysia
azizah07@citycampus.utm.my

Abstract. Cyber Forensics Investigations training or education is relatively new. The nature of Cyber Forensics is multidisciplinary, which enforces proliferations to diverse training programmes, from a handful of day's workshop to Masters Degree in Cyber Forensics. Thus, researchers found that the world lacks experts of Cyber Forensics due to some factors. Consequently, this paper focuses to analyze the trainees' Competency Based-Assessment implementation. The study finds that Cyber Forensics training or education has inappropriate trainees' Competency Based-Assessment below 50%.

Keywords: Cyber Forensics Investigations (CFIs), Information and Communications Technologies (ICTs), Cyber crimes, Knowledge, Skills, and Ability (KSA), Competency Based-Assessment (CBA).

1 Introduction

Nowadays, Information and Communications Technology (ICT) has great impacts and contributions to our life, mostly in a positive way [1]. It involves many aspects of our modern lives and most activities of the people around the world. ICT has obvious influences on communication system, transportation, food production, factories, entertainment, etc. In addition, the rapid progress and development of computers, networks and communications increase the ICT domination.

However, these technologies also facilitate criminal activities. In other words, they contribute to various massive Cyber crimes [2]. So, it is hard to find a crime that doesn't potentially involve digital evidence. Researchers [3] estimate of over 85% of criminal and civil prosecution cases have involved digital evidence. These crimes may include fraud, hacking identity theft and other illegal activities.

Consequently, Cyber crimes have a substantial impact on the world economies. Researches carried out in USA and UK companies indicate that the financial losses due to Cyber crimes are very high [4, 5]. As a result, several businesses have lost reputation, while, individual privacy and even public safety exposed to risks. Furthermore, national security concerns have also risen. Securing information assets of organizations also becomes more complex and highly recommended as computer resources become more vital.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 517–526, 2011.

Cyber Forensics Investigations (CFIs) appear to combat crimes committed by computer utilities either as a source of crime, stores or facilitating criminal activities.

Z. Hamzah [6] defines CFIs as "legal aspects of computer investigation and involves the analysis of digital evidence covering the identification, examination, preservation and presentment of potential electronic evidence in a manner that would allow such evidence to be admitted in a court of law".

Thus, CFIs play a major role in computer and network security, information assurance, law enforcement, and national defense [7]. For instance, the digital forensics investigations (DFIs) skills are vital and much required in the first responding team of most National Critical Corporations. That fact comes due to the requirement of both service availability and maintaining evidence when systems are attacked [8]. Therefore, the jobs of the CF Investigators are not only collecting and analyzing, but are also include proper technical testimony presentation in the court.

As such, CF Investigators should be qualified in order to integrate knowledge, skills, and abilities in the identification, preservation, documentation, examination, analysis, interpretation, reporting and testimonial support of digital evidence [9].

In spite of the fact that the CFIs origin goes backs to early 1980s, and new training or education qualifications are run regularly in some countries, researchers have found that many countries around the world lack CF Investigators [10, 11, 12]. The lack of professionalism in CFIs comes due to some factors. For instances, training and education services on CFIs are relatively new. Furthermore, many of the qualifications are presented by vendors and owners of CFIs tools. In addition, most of training or education providers concentrate on the system weakness or training trainees on the use of specific tools or techniques, rather than concentrating on knowledge, skills, ability (KSA) and competency based-assessment needs [13]. Finally, most of qualifications and programmes are not verified by independent external examiners or accreditation bodies [14, 15].

In this paper, the authors are certified CFIs, IT Competency Based Assessor and Internal Verifiers who have contributed in reviewing the current situation of the practices of CF education and training in both academic and non academic institutions, in terms of different Competency Based-Assessment methods.

The paper is organized into four sections. This section introduces the problem behind this paper. Section two presents a critical analysis concerning trainee's competency based- assessment methods implemented in academic and non academic programmes in CF training or education. In section three, the authors discuss the result of section two, while, section four summarizes the study.

2 Competency Based-Assessment

ILO [16] defines Competency Based-Assessment (CBA) as the procedures of gathering evidence about the employee's occupational performance; in order to form judgments about his proficiency with respect to standards identifies the areas of required performance. Competency is defined as "is much more than just a description of a work task or activity. It encompasses measures of the competency and addresses the knowledge, skills and attitudes required for a person to perform a job to a required standard" [17]. Competency is the verified ability to apply knowledge and skills, where is relevant and defines the personal attributes [18]. In other words, it is the ability to perform in the workplace. For instance, collecting product evidence that done by the trainee in actual activities, through the use of simulation in some activities, observing the trainee while skillfully performing his job, oral or written question to test his knowledge, etc. [19, 20]. However, the CBA is totally differing from traditional assessment [16], as in Table. 1.

Table 1. Traditional assessment systems Vs competency based assessment

Traditional assessment systems	Competency based assessment
Parts or programs are assessed by final examination	Gathering and judgment of evidence with objectives.
Success or fail are defined by marking scores	The decision is (competent or not yet competent).
It is done in limited time	It is not subject to predetermine time

An effective CBA is required to be initiated within the training or education primary objectives in order to assure the trainees' proficiency. The CBA idea is based on trainee being *competent or not yet competent*. It is stating that there is sufficient evidence to show trainee is competent or not [16, 21]. In addition, it must be carried continuously during the training or education process in order to remedy the trainee's competency deficiency earlier.

The authors investigation located 27 CF related programmes including two-year associate degrees and diplomas [22,25], four-year undergraduate programmes [22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,34,35], postgraduate degree programmes [22, 36, 37], and professional certification programmes [22, 23, 25, 38, 39, 40].

2.1 Associate Degrees and Diplomas Programmes

In order to meet court rules for evidence admissibility, all of Associate Degree and Diplomas Programmes (ADDPs) offer training of how to set up CFIs procedures properly, and how trainees use commercial CFIs software packages in standard methods and continuous process; as well as the followings:-

- To identify and secure the Cyber evidence at crime scenes.
- To record all procedures carried out at crime scenes and sources of Cyber evidence.
- To maintain the evidence handling and chain of custody properly.
- To ensure valid forensic images are taken to original Cyber device
- To use virtual environment to boot or restore the image of forensic evidence
- To save the multiple copies of potential evidence in a different digital media

The CFIs are known as laboratory procedures that depend on many steps. So, variety of hands-on laboratory exercises is provided. In addition, some of CBA methods are used to ensure the trainees' proficiency, such as programme documentation, objectives evaluation and regular assessments to determine whether the programme objectives and the development needs are met or not [22, 25], see Fig. 1.

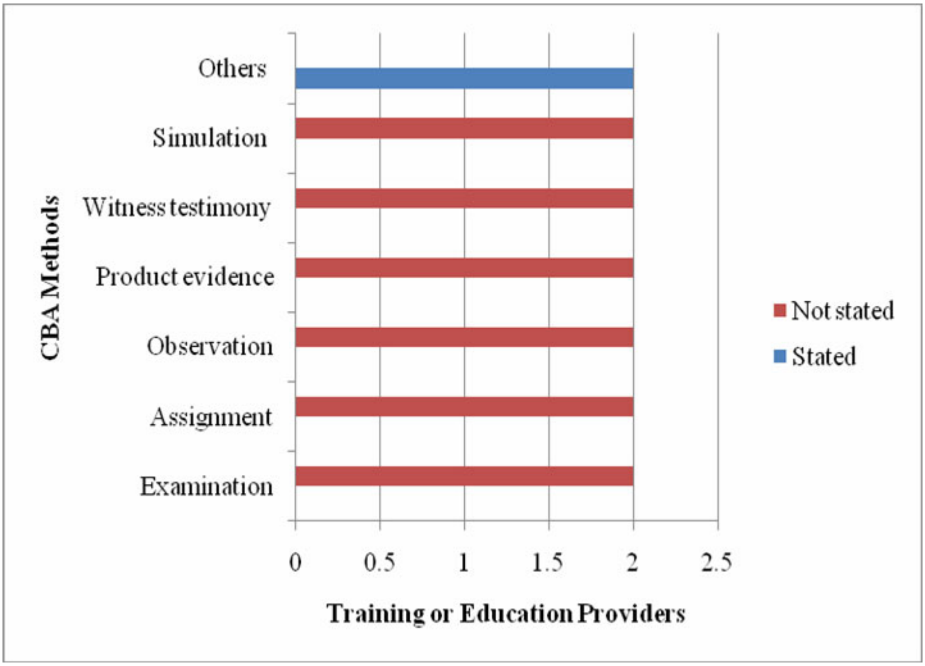


Fig. 1. Associate Degree and Diploma Programs Trainee Competency Assessment

2.2 Undergraduate Programs

A conspicuous similarity of Undergraduate's degree programmes in both training and CBA methods are shown in Fig. 2. They deliver subjects as a combination of lectures and Laboratory sessions. The hands-on exercises take a significant position on training delivery, as well as some types of assessment. Computer Science degree at the University of Western Sydney [10] depends on assignment and exam to evaluate its trainees. While, lab exercise log books and written report are used by faculty for student assessment [32, 33]. The Undergraduate computer forensics program that is part of the Computer Information Systems (CIS) uses three quizzes, midterm exam, and final exam to assess the trainees. Deakin University in Australia adopts two simulated assignments to develop the trainees' skills. Each assignment is case study [34]. Some of lab assignments include:-

- Acquiring data imaging
- Recovery deleted data

- Capture volatile memory data analysis
- Analyzing removable medias
- Analyzing of logs and events viewer
- Password and encryption methods
- Finding concealed data, images and steganography
- Decrypting files

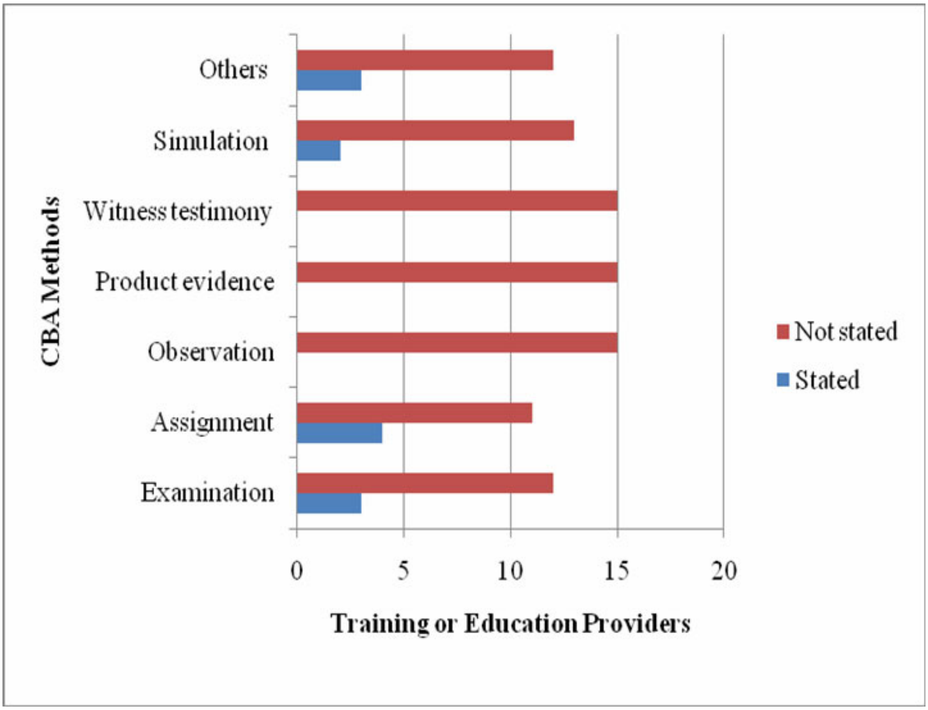


Fig. 2. Undergraduate Trainees' Competency Assessment

2.3 Postgraduate Programs

Most of Postgraduate Computer forensics training or education programmes are at Master of Science (M. S.) degree. A Postgraduate or Master–level on Computer forensics training or education programmes intend to do more than educating trainees in theoretical concepts. They equip the trainees with abilities and skills techniques of how to think critically, solving problem, and adequate knowledge with the forensic science discipline [22]. CFI, a multidisciplinary mix of ethical, legal, technical and Courtroom testimony-based course, is relatively new field of study. Hands-on knowledge is the main laboratory component. Researches, conferences, assignments and simulated courtroom are used to assess the trainees' competency, see Fig. 3.

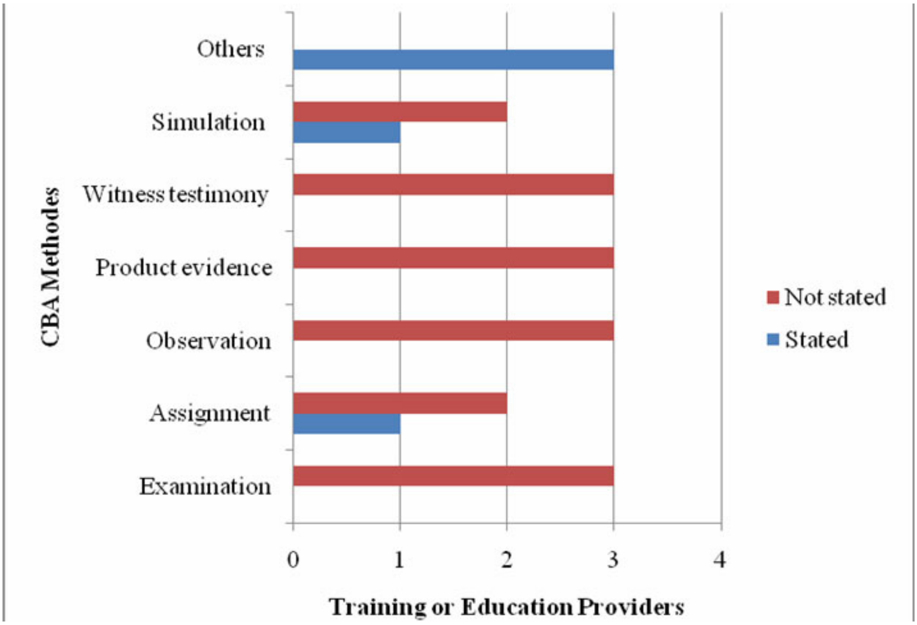


Fig. 3. Postgraduate Trainees' Competency Assessment

2.4 Professional Certification Programs

Although Professional Certification Programmes (PCPs) are qualifications that can be received from academic institutions, they are mainly obtained from professional institutions. For instance, some software vendors offer forensics skill training via some courses, such as Encase software training [3, 38], for more details, see Table 2.

Table 2. Professional Computer forensics certifications

Certification	Training or education providers
Certified Hacking Forensics Investigation (CHFI)	EC-Council
Certified Computer Examiner (CCE)	Int'l Society of Forensics Computer Examiners
GIAC Certified Forensics Analyst (GIAC CFA)	Global Information Assurance Organization
Certified Computer Crime Investigator (CCCI) and Certified Computer Forensics Technician (CCFT)	High Tech Crime Network
Certified Forensic Computer Examiner (CFCE)	Int'l Association of Computer Investigators
Cyber Security Forensics Analyst (CSFA)	Cyber Security Institute

Though the PCPs don't grantee trainees competencies, they establish a knowledge baseline. They provide Computer forensics practicing that show some skills of individuals. [22] states that "Certificates in digital forensics alone may not be sufficient for an entry level position in digital forensics". Examination, assignment, observation, Peer- review publications and Instructor of presenter evaluation are carried out to assess the trainees' competency as shown in Fig. 4.

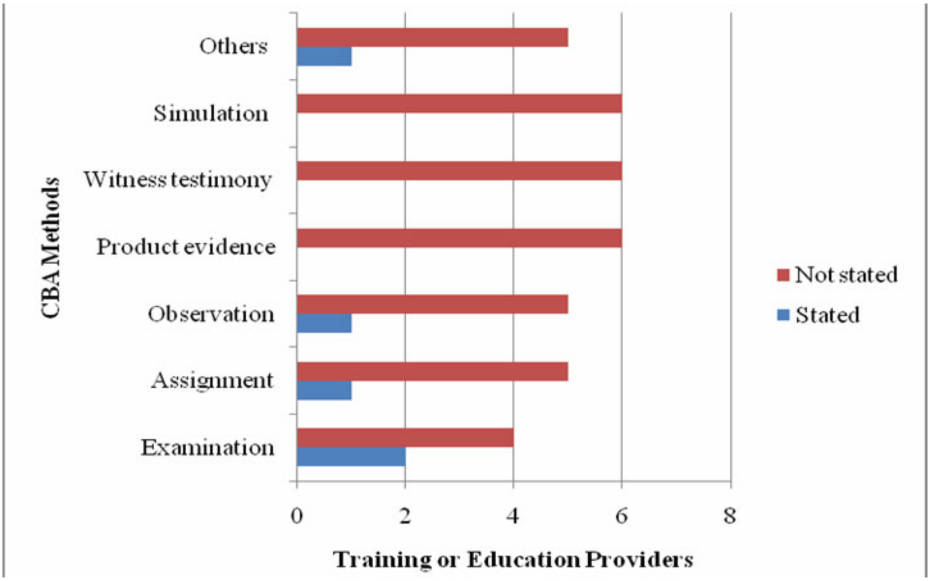


Fig. 4. Professional Certification Programs Trainees' Competency Assessment

3 Discussion

CF professionalism varies from first responders to analysts and experts. However, we are not surprised to find significant differences in curriculum, admission criteria, trainees' competency assessment methods and programme length. Fig. 1 above depicts ADDPs trainees' competency. As seen, most of competency assessment methods are not shown totally, except programme documentation and objectives evaluation.

Fig. 2 presents CF Undergraduate programmes' competency assessment methods. The study shows noticeable likeness of undergraduate's degree programs trainees' assessment methods. All of them show implementation of insufficient competency assessment methods. From Fifteen Undergraduate programmes studied, nine of them [22, 23, 24, 25, 26, 27, 28, 29, 30] don't mention how trainee's competency can be evaluated. One programme [31] shows 14.3% of CBA methods implementation. Four programmes [32, 33, 34, 35] implement 28.6% of CBA methods, and only one program gears over 40% of assessment methods.

DF Postgraduate and diploma programmes CBA methods are shown on Fig. 3 above. Of the three Master programmes studied, the trainee's CBA percentages are below 50%. Two programmes [22, 36] provide 14.3% of CBA methods. They only use researches and conferences to measure their student's competency levels. [37] shows 42.8% of CBA methods are implemented such as assignment, simulation and research.

The authors site 6 programmes of Professional Certification Programmes [22, 23, 25, 38, 39, 40]. Fig. 4 shows insufficient trainees' CBA implementations. Four out of six programmes don't reveal their trainee evaluation system methods [23, 25, 39, 40]. [38] uses only written exam method to assess the trainee. Whereas, [22] uses examination, assignment, observation and peer review publication in trainee's evaluation system.

Overall, the implementations of the CBA methods of CF training or education programmes are below 50%.

4 Conclusion

In summary, Cyber Forensics training or education services proliferate to diverse qualifications, from a handful of day's workshop to Masters Degree in Cyber Forensics. With the immaturity of education or training and practicing programs, most of the training or education providers don't worry about the comprehensive Competency Based Assessment. This may contributes to poor services delivery and outcomes. The study shows the trainees' Competency Based-Assessment methods are below 50%. Overall, the result shows trainees' Competency Based-Assessment has been given inappropriate consideration. These results reflect that the area of Cyber Forensics training or education does not concern about trainees' Competency Based Assessment.

Acknowledgments. This work is of a research that has been done in Universiti Teknologi Malaysia (UTM), under support from Ministry of Science, Technology & Innovation, Malaysia.

References

1. Figg, W., Zhou, Z.: computer Forensics Minor Curriculum Proposal, Department of Computer Information Systems. Dakota State University, Madison, SD 57042, http://portal.acm.org/ft_gateway.cfm?id=1229642&type=pdf (last visit July 10, 2007)
2. Fernandez, J., Smith, S., Garcia, M., Kar, D.: Computer forensics: a critical need in computer science programs. Journal of Computing Sciences in Colleges, 20(4) (2005), Consortium for Computing Sciences in Colleges
3. Taylor, C., Endicott-Popovsky, B., Phillips, A.: Forensics Education: Assessment and Measures of Excellence. In: Proceedings of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering, SADFE 2007 (2007)
4. Venter, J.: Process Flows for Cyber Forensics Training and Operation. Pretoria, South Africa (2006), http://researchspace.csir.co.za/dspace/bitstream/10204/1073/1/Venter_2006.pdf (last visit July 15, 2010)
5. Pidanick, R.: An Investigation of Computer Forensics. Information System Control Journal vol. 3 (2004), http://www.isaca.org/Journal/Past-Issues/2004/Volume3/Documents/jpdf043-InvestigationofComputerForensics.pdf (last visit August 10 2010)
6. Hamzah, Z.: E-Security Law & Strategy. Malayan Law Journal Sdn Bhd, 47301 Kelana Jaya Malaysia, KL, ISBN 967-962-632-6
7. Figg, W., Zhou, Z.: Computer Forensics Minor Curriculum Proposal: Department of Computer Information Systems. Dakota State University, Madison, SD 57042, http://portal.acm.org/ft_gateway.cfm?id=1229642&type=pdf (2007) (last visit August 20, 2010)

8. Hopkins, D.: Innovative Corporation Solutions, Inc. (2006), http://www.michigantechnologyleaders.com/../InnovativeNewsImportanceofComputerForensics.pdf (last visit August 20, 2010)
9. Schroeder, S.C.: How to be a Digital Forensic Expert Witness. In: First International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 69–85 (2005)
10. Bem, D., Huebner, E.: Computer Forensics Workshop for Undergraduate Students (2008), http://crpit.com/confpapers/CRPITV78Bem.pdf, (last visit September 10 2010)
11. Zhijun, L., Ning, W.: Developing a Computer Forensics Program in Police Higher Education. In: ICCSE 2009. 4th International Conference on Computer Science & Education 2009, pp. 1431–1436 (2009)
12. Lim, S.: Demand up, supply short for forensic professionals Anonymous, p. 5. New Straits Times, Kuala Lumpur (2009)
13. Nance, K., Armstrong, H., Armstrong, C.: Defining an Education Agenda. In: 43rd Hawaii International Conference on System Sciences, HICSS 2010, pp. 1–10 (2010)
14. Barbara, J.J.: Certification and Accreditation Overview (2008), http://www.springerlink.com/index/t5852u690qhv5512.pdf, (last visit October 10, 2010)
15. Sabeil, E., Manaf, A.A., Ismail, Z.: Analysing the Quality Assurance of Trainees Competency Assessment and Accreditation of Cyber Forensic Education/Training Programs. In: ICMLC (2011), the 3rd International Conference on Machine Learning and Computing, Singapore (February 2011)
16. ILO (CINTERFOR).: What is labour competencies assessment , http://www.ilo.org/public/english/region/ampro/cinterfor/temas/complab/xxxx/31.htm (last visit, January 1, 2011)
17. Graeme, D.: A Guide to Writing Competency Based Training Materials (2003), http://www.volunteeringaustralia.org/../Revised%20Writers%20Guide%202.pdf (last visit January 3, 2011),
18. EURIM.: Supplying the Skills for Justice (2004), http://www.eurim.org/consult/e-crime/may_04/ECS_DP3_Skills_040505_web.htm, (last visit, June 10, 2010)
19. Greatorex, J.: How can NVQ assessors' judgements be standardised (2003), http://www.cambridgeassessment.org.uk/ca/digitalAssets/113886_How_can_NVQ_Assessors__Judgements_be_Standardised.pdf, (last visit January 3, 2011)
20. NVQs - national vocational qualifications, http://www.businessballs.com/nvqs_national_vocational_qualifications.htm, (last visit January 3, 2011),
21. John, W.: Competency based education and training. British Library Cataloguing in Publication Data Competency based education and training. 1. Competency-based education I. Burke, John 371.3 (2005)
22. West Virginia University.: The Technical Working Group on Education and Training in Digital Forensics (2007), http://www.ncjrs.gov/pdffiles1/nij/grants/219380.pdf, (last visit, (June 15, 2010)

23. Liu, Z., Wang: Developing a computer forensics program in police higher education. In: 4th International Conference on Computer Science & Education. ICCSE 2009, pp. 1431–1436 (2009)
24. Yasinsac, A., et al.: Computer forensics education. Security & Privacy. 1(4), 15–23 (2003)
25. Larry, G., et al.: Computer forensics programs in higher education: a preliminary study. In: Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education. ACM, New York (2005)
26. Stephens, P., Induruwa, A.: Cybercrime investigation training and specialist education for the European Union. In: Proceedings - 2nd International Annual Workshop on Digital Forensics and Incident Analysis, WDFIA 2007, pp. 28–37 (2007), Compendex
27. McGuire, T., Murff, K.: Issues in the development of a digital forensics Curriculum. Journal of Computing Sciences in Colleges 22(2) (2006), Consortium for Computing Sciences in Colleges
28. Irons, A.D., et al.: Digital Investigation as a distinct discipline: A pedagogic perspective. Digital Investigation 6(1-2), 82–90 (2009)
29. Murdoch University.: Cyber Forensics, http://www.murdoch.edu.au/Courses/Cyber-Forensics-Information-Security-and-Management/ (last visit October 15, 2010)
30. Deakin university.: Bachelor of Information Technology (I.T. Security), http://www.deakin.edu.au/futurestudents/courses/course.php?course=S334&stutype=local&keywords=digital%20forensic (last visit 15 October 2010)
31. Louise, L., et al.: Establishing network computer forensics classes. In: InfoSecCD 2004: Proceedings of the 1st Annual Conference on Information Security Curriculum Development. ACM, New York (2004)
32. Luther, T., et al.: Forensic Course Development. In: Proceedings of the 4th Conference on Information Technology Curriculum, CITC4 2003. ACM, New York (2003)
33. Luther, T., et al.: Forensic Course Development – One Year Later. In: SIGITE 2004. ACM, Salt Lake City (2004)
34. Batten, L., Lei, P.: Teaching Digital Forensics to Undergraduate Students, vol. 6(3), pp. 54–56 (2008)
35. Kevin, R., Hongmei, C.: Framework for the design of web-based learning for digital forensics labs. In: Proceedings of the 47th Annual Southeast Regional Conference, ACM-SE 47. ACM, New York (2009)
36. John Jay College of Criminal Justice.: Master of Science in Forensic Computing, http://www.jjay.cuny.edu/academics/690.php, (last visit November 12, 2010)
37. Philip, C., et al.: Master's Degree in Digital Forensics. In: 40th Annual Hawaii International Conference on System Sciences, HICSS 2007 (January 2007)
38. Training Programs.: EnCE Certification Program, https://guidancesoftware.com/training/training_programs.aspx (last visit November 12, 2010)
39. Lane Department of Computer Science and Electrical Engineering.: Computer Forensics Certificate, http://www.csee.wvu.edu/, (last visit July 7 2009)
40. Lee H.C.: College of Criminal Justice and Forensic Sciences.: Forensic Computer Investigation Certificate, http://www.newhaven.edu/5930/ (last visit October 5, 2010)

Survey of Software Engineering Practices in Undergraduate Information System Projects

Nor Shahida Mohamad Yusop[1], Kamalia Azma Kamaruddin[1],
Wan Faezah Abbas[1], and Noor Ardian Abd Rahman[2]

[1] Universiti Teknologi MARA (UiTM),
40450 Shah Alam, Selangor, Malaysia
{nor_shahida,kamalia,wfaezah}@tmsk.uitm.edu.my
[2] Accenture Malaysia, Level 35, The Gardens North Tower,
Midvalley City Lingkaran Syed Putra
59200 Kuala Lumpur, Malaysia
ardian.noor@accenture.com

Abstract. Past research stated that the effectiveness of software engineering (SE) course is determined by theoretical knowledge, added with some practices and improved by the individual experiences [2]. An Information Systems Engineering programme, coded as CS226, has been introduced in Universiti Teknologi MARA in December 2000 with the aim to produce graduates who have extensive knowledge and skills in information technology mainly in software engineering fields. This paper describes a survey done to understand the undergraduate students' habit in applying SE practices in developing IS projects. This research collected data by distributing questionnaires to 78 ISE students. The questionnaire consists of questions related to systems planning, systems requirement and analysis, system design and implementation and system testing activities. Data collected are then analyzed using SPSS and the result can be used to measure the effectiveness of the courses taught and improve flaws that exists in SE curriculum.

Keywords: Information system engineering, software engineering practice survey.

1 Introduction

In line with professional practices, the education of all software engineering courses must include student experiences with the professional practice of software engineering [1]. In their research, Shaw and Dermoudy [2] have identified that the three important elements that contribute to the effective software engineering (SE) course are through theoretic knowledge, practice and experience. Hence, this can be applied in course project to confront students with realistic and challenging software engineering issues.

In compliance with the SE course demand, a CS226 program was introduced in UiTM that was aimed to produce SE undergraduates. To align with the industrial

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 527–537, 2011.

needs, the curriculum was design to support real software development environment through team-based project that exposed the students with practice and experience of SE essentials. The common practice is the project starts from scratch and is to be completed in a semester. However, due to the limited time period, it is constraining the size of the system developed and increasing the risk that the final product may be incomplete, non-functional or lack of quality [3].

The knowledge of whether the students are actually deploying the SE practices in the completion of the project is very vague. Moreover, there is no formal method or guideline to monitor and ensure students follow each of the software engineering principles provided in the curriculum.

This survey is an indication of the undergraduate students' habits in applying software engineering practices in developing their project during undergraduate courses. The outcome from this survey can be used as a benchmark on the effectiveness of the course taught to the students by improving any flaws that exist in the SE curriculum.

This paper is structured as follows; Section 2 describes some literature review whilst Section 3 explains the undertaken methodology in carrying out the survey. Section 4 discusses the results and analysis prior to the conclusion remarks in Section 5.

2 Methodology

This study was very much on exploratory survey to identify students' habits toward software engineering practices. We therefore decided to conduct the survey through questionnaire method which we can tabulate the answer to determine averages and trends. This section explains how we distributed and analyzed the questionnaires.

2.1 Study Design

Data about educational software process practice implemented in university was collected from a paper-based survey questionnaire on the preferred activities that students chose to develop their systems. The sample obtained from 78 Information System Engineering students, although the expected target was 120. The objective of this study is to investigate the student pattern and their perception of software development activities preferences while completing their project. The results were derived from SPSS 16.0 and analyzed via percentage proximity response techniques.

2.2 Participants

The survey involves Information System Engineering undergraduate students of mixed gender groups from Part 3 until 8 from Faculty of Computer and Mathematical Sciences. The students were given less than 30 minutes to answer the questions. Students from different parts were asked to answer different sets of questionnaires according to their experience of learning and developing the group project.

2.3 Instruments

The questionnaire comprise of 123 questions which were divided into 11 parts, and consist of 122 close-ended questions and only one open-ended question. Students who are in their fourth semester or known as Part 4 will only need to answer questions until part 6 as they have learned to do system requirements in the previous semester. Whilst students from Part 5 or fifth semester will answer questions until part 8 which encompasses topic until design. While Part 6, 7 and 8 students need to answer the entire questionnaire which includes how to manage testing questions. The questions are divided into system planning, system requirement and analysis, system design and implementation and system testing activities.

2.4 Data Management and Analysis Techniques

Data was converted into numeric format which used scale of 0 to 2. 0 will represent that the student agree with the question, 1 represent that the student is unsure about the question and 2 represent student disagreement with the question. The results will be presented into a percentage graph of preferable activities that students usually used in completing their project. Most of it will be shown in tabulation format.

3 An Overview of Information System Engineering Program (CS226)

In December 2000, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam offered for the first time, Bachelor of Science (Hons) Information System Engineering (ISE). Coded as CS226, the program was aimed to produce graduate who have extensive knowledge and skills in information technology mainly in software engineering field. This program was designed to incorporate SE knowledge areas such as software modeling and analysis, software design, software verification and validation, software process, software quality assurance and others.

In supporting the learning objectives, the program was structured so that the students have to work in team-based project of 3-4 students for three consecutive semesters. The aim of this project is to build teamwork, enhance communication skills and expose the students on software engineering best practices in bringing the students to the reality of fuzzy and challenging software development. In the National University of Singapore (NUS) [4], the team-based project course is focused on design and implementation phase of the software development lifecycle (SDLC) only, while in CS226 program, the team project emphasized on the whole SDLC activities which consists of planning, analysis, design and testing. Those activities are divided into three courses; ITS470 Object-Oriented Requirement Analysis (semester 3), ITS570 Object-Oriented Design & Implementation (semester 4) and ITS670 System Testing & Evolution (semester 5). Each course is a pre-requisite to the later one. Fig. 1 depicts the workflow of course outline for team-based project in CS226 program.

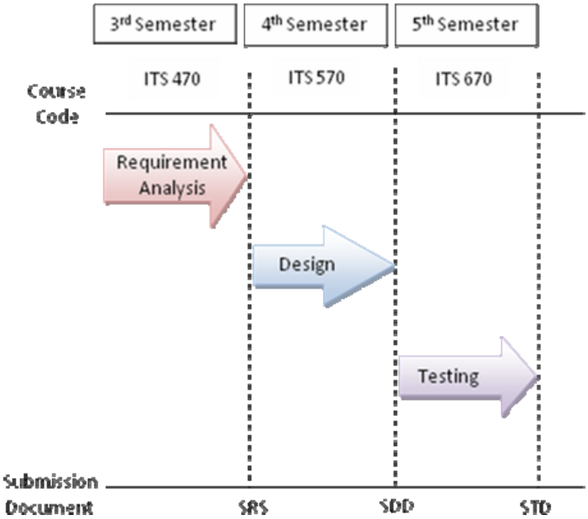


Fig. 1. Main course cutline for team-based project of Information System Engineering program (CS226)

The team-based project is determined by the lecturers. The project differs between each batch or students intake. Project such as e-advising system, e-counseling, e-welfare are examples of previous project developed by students. Each group is assigned to the same particular project and it differs in terms of their functionality based on requirements gathered from different clients. But, basically the outcome are similar as each project must be a web-based project and must be able to do the CRUD functionally (an acronym for *Create, Read* or *Report, Update and Delete*) that must be connected to a database.

At the beginning of the semester, students are required to do project planning to clarify the project objectives, schedule, risk and contingency plan and project methodology that will be used later. Students spend almost 12 weeks on requirement analysis that involved with fact-finding, modeling and documenting. Since ITS470 is designed for object oriented approach, all analysis results are represented through Unified Modeling Language (UML) diagrams and Software Requirement Specification (SRS) is produced in compliance with IEEE standard.

In ITS570, students are exposed in transforming the descriptive analysis models into computational models for development. At this time, students are required to complete the architectural design, detailed design and interface design. This information is documented in Software Design Documentation (SDD). Once they have completed their design, they are required to do coding and implement their system. At the end of the fourth semester, the final product needs to be presented and evaluated by several panels which are among the lecturers.

During ITS670 course, students are introduced with several types of testing technique. For the purpose of testing practices, students are required to prepare

Software Test Design (STD) before running system test. Each team is assigned to another project to evaluate the system based on STD. The system testing is done in two phases to compare the level of bugs found during each phase. Finally, the students will do a user acceptance test (UAT) of their project to verify their system.

In the end, the students should be able to learn and experience the software development life cycle from the planning phase, to requirements and analysis, design and implementation and testing phase. Over the 10-year periods, the program has evolved in large part owing the valuable feedback and high demand enrollment among the students.

4 Results and Discussions

The respondent of this survey consists of students who are pursuing CS226 course and 98.7% of them have had experience in developing system. Therefore, the feedbacks obtained from them are based on knowledge and experiences gain through the courses enrolled. In this section, we summarize our survey results into 4 main software engineering (SE) essentials, which are: documentation, requirements engineering, software design and quality assurance.

4.1 Documentation

Documentation is one of the recommended practices in software engineering that helps in development and maintenance of project [5]. It includes any artifact that is used to communicate project information among the stakeholders from different levels such as managers, project leaders, system analyst, developers and customer [6]. The artifact includes requirement specification, detailed design and testing/verification.

In this research, the respondents used IEEE documentation standard in ensuring the documentation quality. The objectives for this section are to:

1. Investigate the level of student’s understanding/knowledge in documentation.
2. Identify the existence of document review process.
3. Identify student’s writing practices in completing SRS, SDD and STD.

Survey results: On average the respondents did not really understand on how to document the requirement into Software Requirement Specification (SRS), Software Design Document (SDD) and Software Test Design (STD) respectively, but they still put some effort to complete the documents. Referring to Table 1, only 21.8% respondents understand how to document the requirement very well while the other 70.5% were not very sure. Most of them (43.6%) prefer to refer to the previous SRS documentation. For design purpose, survey shows 59.4% respondents did not know how to document the design very well. On the other hand, for testing documentation only 38.3% were able to write test cases well.

Table 1. Students' knowledge on documenting requirement for analysis, design and testing

	Yes (%)	Not Sure (%)	No (%)
Understand how to document the requirement into SRS	21.8	70.5	7.7
Refer to previous SRS documentation	43.6	42.3	23.1
Know how to document SDD	32.8	7.8	59.4
Able to write test cases well	38.3	56.9	2.1

Next, we asked respondents on their common practices in performing documentation process; document review and writing method. In terms of document review, 67.9% of the students had baseline their SRS before proceeding to the design phase. While 54.7% perform SDD review to confirm the requirement before development take place and 42.6% did STD review.

Table 2. Student's common practices in document review

Document review	Yes (%)	Not Sure (%)	No (%)
SRS	67.9	29.5	2.6
SDD	54.7	37.5	7.8
STD	42.6	51.1	6.4

For writing method, SRS is documented at the end of the requirement analysis phase (66.7%). However, for design and testing, survey shows the respondents like to do reverse engineering method, which is performing the task first before documenting the content. 51.6% prefer to design the system first before documenting the design while 48.9 % write the STD after performing system testing.

Table 3. Student's common practices in writing method

Writing Method	*Document*	*Yes (%)*	*Not Sure (%)*	*No (%)*
Forward engineering	SRS	66.7	30.8	2.6
Reverse engineering	SDD	51.6	35.9	12.5
	STD	48.9	42.6	8.5

Discussion: In evaluating students practices in documentation, there are two attributes identified which are understandability and implementation. Understanding the document content is found to be the hardest part as the standard/manual is too general to follow. Holt [7] in his survey also found that most of his participants complaint on the generalization of the standard and they do not know where to start implementing the standards. Due to this problem, most of the students prefer to do reverse engineering method in documenting the design and testing strategy. In motivating students to understand the importance of documentation, each section in the document should be discussed in detail. Having completed and updated documentation could increase future maintainability and provide clear system understanding [8]. Besides that, a detail step-by-step and to-do checklist should be provided by the lecturer in assisting students in completing the documentation.

4.2 Requirements Engineering

Requirement engineering is divided into two main groups which are requirement development that consists of activities such as discovering, analyzing, documenting and validating while requirement management focuses on requirement traceability and maintainability [9]. For CS226 program, students are given more emphasis on requirement development to improve their skill on system analysis activities which focus on object oriented requirement analysis. The objectives of this section are to:

1. Investigate the most frequent used of fact-finding techniques in requirement elicitation.
2. Identify the requirement modeling techniques used in requirement analysis.

Survey results: For requirement elicitation, responses show a high use of interviewing technique (79.5%) followed by observation of the business workflow (50%) and prototyping 40.6%. Also, 38.5% distribute questionnaire in gathering the requirements. Only 37.2% respondent review related documentation to understand problems better.

Table 4. Frequently used of fact-finding techniques in requirement elicitation

Fact-Finding Technique	Yes (%)	Not Sure (%)	No%
Interview	79.5	20.5	0
Observation	50	46.2	3.8
Prototyping	40.6	46.9	12.5
Questionnaire	38.5	29.5	22.1
Review documents	37.2	46.2	16.7

For the requirement analysis, the data revealed that over 50 percent of survey population used object oriented analysis technique in representing the requirement through Use case Diagram (69.2%), Sequence Diagram (71.8%) and Class Diagram (69.2%). Although we expected all of them emphasized on object oriented modeling, but the survey provided evidence that some of them still deployed the structured approach using Data Flow Diagram (DFD) and Entity Relation Diagram (ERD) in modeling the requirements (56.4% and 55.1% respectively).

Table 5. Requirement modeling techniques used in requirement analysis

Modeling Technique	Yes (%)	Not Sure (%)	No%
Use Case Diagram	69.2	30.8	0
Sequence Diagram	71.8	26.9	1.3
Class Diagram	69.2	30.8	0

Discussion: The first step in understanding the project to be developed requires the fact-finding skills. Even questionnaire have a limited and specific use in information gathering, it enables the data collection from a large number of stakeholders [10]. Due to the time constraint, it helps in getting quick response and determines trends. In order to provide a comprehensive fact-finding approach, it would be useful to introduce combination of interview, workshop and iterative development for requirement verification and validation activities [11].

However, referring to Table 4, it shows that the highest fact finding technique that the student use is interview method. This is due to the user for the project is not on a large demographic scale. The project is user specific and it is customized to the potential user requirements which lead to interview method as the best technique for this case.

Also, a formal requirement review session could be included in the curriculum to monitor the requirement analysis activities such as to build models of process, data and behavioral domain [9]. This is to ensure students follow the object oriented approach as per syllabus. Referring to Table 4, it shows that majority of the students are unclear on the concept of suitable fact finding techniques in collecting requirement. In order to overcome this issue, it is advised to include a practical approach in class discussions such as doing mock interview and observation practices in class for the students to understand the different approaches used in requirement gathering.

4.3 Software Design

Software design is aim to define, organize and structure the components of the final system solution that will serve as the blueprint for construction [10]. Through the object oriented approach, several models are produced such as package diagrams, class diagrams, interaction diagrams and object database schema. The objectives of this section are to:

1. Study student's understanding on design concepts
2. Investigate how they deal with project development challenges.

Survey results: This survey shows a weak understanding on design concepts as more than half of the participants responded negatively (Not Sure - 46.9% and No - 7.8%). It can be proven where most of the project failed to be submitted and presented on time (53.1%) even though they have planned the technique, methods and activities before starting to design the system.

In dealing with project development challenges, 57.8% reuse source code from previous project developed by other students. While other 50% use open source code from internet. This analysis shows the practices among the students are not consistent and unsuitable to help them in system design.

Table 6. Student's practices in developing the project

Practice	Yes (%)	Not Sure (%)	No%
Reuse source code from previous system	57.8	35.9	6.2
Use open source code from internet	50	45.3	4.7
Change less than 50% of requirements during design phase	21.9	60.9	17.2

Discussion: Even though a proper documentation and requirement analysis are in practice, the survey shows ineffective design that leads to delivery of non-working system. As shown in Table 6, most of the students prefer to copy source code from previous system done by their seniors because each project has the same basic CRUD functionality. Each project must be able to add, view, update and delete. The only difference is in the project domain, which explains why the practice of reusing and using open source code as their main practice. It would be beneficial to introduce the students with evolutionary design where it can provide opportunity of project growth gradually [12]. The features/functionality of the system could be delivered by releases in ensuring at the end of the project submission; the system is working although not all features have been incorporated.

4.4 Quality Assurance

Quality assurance (QA) practices such as technical reviews and testing are an umbrella process that is performed throughout the software development lifecycle (SDLC) to ensure software quality [10]. A proper testing strategy conducted by students will confirm the project is developed as expected and free of bug. The objectives of this section are to:

1. Investigate the awareness of quality assurance practices.
2. Identify the type of testing conducted by the students.

Survey results: Through the survey, respondents show the awareness of quality assurance practices while developing the project. Even though only, 27.7% respondents begin their testing phase with setting up the organizational objectives, test policy and test strategies but moving forward to the development phase, results show the increase number in testing activity. 34% respondents started with unit testing while 40.4% did functional testing. The remaining respondents - 27.7%, performed integration testing and only 36.2% ended testing activity with system test. Overall, there were 63.7% incomplete project delivered at the end of the semester which results on the low percentage of integration and system testing.

Table 7. Types of testing conducted by the students

Testing Type	Yes (%)	Not Sure (%)	No%
Unit testing	34	59.6	6.4
Functional testing	40.4	53.2	6.4
Integration testing	27.7	61.7	10.6
System testing	36.2	53.2	10.6

Discussion: Unable to complete the project on time has critical implications to the software testing activities, especially in performing integration and system testing. Students are more focused on implementations and less emphasize on QA practices. According to Astigarraga, curricula today emphasize more on requirement gathering, design and implementation and set aside testing essentials [13]. While in a survey done by Chan [14] university curricula was found to have insufficient coverage in preparing graduates for software testing. Although CS226 was designed to focus on software testing; it would be very useful to introduce students with combinatorial testing method instead of common testing type. Besides that, to ensure the effectiveness of QA practices, an inspector could be appointed to monitor and inspect project progress by giving comment, feedback and suggestion to ensure the project could be completed on time and in the right track. This would strengthen their understanding of QA practices.

5 Conclusion

This research served as a benchmark to measure the students' understanding in applying software engineering practices to their team-based projects. It has shown that there are a percentage of ISE students that have not fully applied the theoretical knowledge they have learnt in class to their project works, as discussed in the documentation, requirement engineering, design and quality assurance result above. This maybe due to the understandability and implementation issues where students did not managed to assimilate both concepts together. Based on the finding, 63.7% of the participants were unable to complete their project within the given timeframe. We believed the factors would either due to lack of programming skills, attitude, or poor time management which can be explored further in future research. Taking cue from this survey, improvements can be made in teaching and learning method in the ISE curriculum so that the programme can produce graduates who have extensive knowledge and skill in the software engineering fields.

References

1. Joint Task Force on Computing Curricula, Software Engineering 2004: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering, tech. report, IEEE CS and ACM (2004), http://sites.computer.org/ccse/SE2004Volume.pdf
2. Shaw, K., Dermoudy, J.: Engendering an empathy for software engineering. In: Proceedings of the 7th Australian Conference on Computing Education, pp. 135–144 (2005)
3. Sebern, M.J.: The software development laboratory: Incorporating industrial practice in an academic environment. In: Sebern, M.J. (ed.) Proceeding of the 15th Conference of Software Engineering Education and Training (CSEE&T), pp. 118–127 (2002)
4. Jarzabek, S., Eng, P.K.: Teaching an advance design, team-oriented software project course. In: Proceedings of the 18th Conference on Software Engineering Education and Training (CSEE&T), pp. 223–230 (2005)
5. de Souza, S.C.B., Anquetil, N., de Oliveira, K.M.: A study of the documentation essential to software maintenance. In: Proceedings of the 23rd Annual International Conference on Design of Communication: Documenting & Designing for Pervasive Information (SIGDOC), pp. 68–75 (2005)
6. Forward, A., Lethbridge, T.C.: The relevance of software documentation, tools and technologies: A survey. In: Proceedings of the 2002 ACM Symposium on Document Engineering (DocEng), pp. 26–33 (2002)

7. Holt, J.: Current practice in Software Engineering: a survey. Computing & Control Engineering Journal 8, 167–172 (1997)
8. Umarji, M., Seaman, C., Koru, A.G., Liu, H.: Software engineering education for bioinformatics. In: Proceedings of the 22nd Conference on Software Engineering Education & Training (CSEET), pp. 216–223 (2009)
9. Pandey, D., Suman, U., Ramani, A.K.: An effective requirement engineering process model for software development and requirements management. In: Proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom), pp. 287–291 (2010)
10. Satzinger, J.W., Jackson, R.B., Burd, S.D.: Systems Analysis & Design in a Changing World, pp. 133–631. Course Technology, USA (2009)
11. Mishra, D., Mishra, A., Yazici, A.: Successful requirement elicitation by combining requirement engineering techniques. In: Proceedings of the First International Conference on Applications of Digital Information and Web Technologies (ICADIWT), pp. 258–263 (2008)
12. Reichlmayr, T.: The agile approach in an undergraduate software engineering course projects. In: Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference, vol. S2C13–18, pp. 13–18 (2003)
13. Astigarraga, T., Dow, E.M., Lara, C., Prewitt, R.: The emerging role of software testing in curricula. In: Astigarraga, T., Dow, E.M., Lara, C., Prewitt, R. (eds.) Proceedings of the Transforming Engineering Education: Creating Interdisciplinary Skills for Complex Global Environments, pp. 1–26 (2010)
14. Chan, F.T., Tse, T.H., Tang, W.H., Chen, T.Y.: Software testing education and training in Hong Kong. In: Proceedings of the Fifth International Conference on Quality Software (QSIC), pp. 313–316 (2005)

Health Ontology Generator, Distiller and Accumulator

Yip Chi Kiong, Sellappan Palaniappan, and Nor Adnan Yahaya

Department of Information Technology,
Malaysia University of Science and Technology,
Kelana Square, Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia
yipzqiang@yahoo.com, sell@must.edu.my,
noradnan@must.edu.my

Abstract. Most healthcare institutions such as hospitals and clinics store their data in the form of databases of various formats. The Health Ontology System that we have developed provides a means to integrate these data with concepts and semantics in the form of a shared cumulative ontology for enabling machines to interpret them. This involves three major software tools, namely, Ontology Generator, Ontology Distiller and Ontology Accumulator. The Ontology Generator is used to create an ontology from a selected database using metadata provided by its database management system. In a reverse process, the Ontology Distiller enables a subset of data from an ontology to be distilled into a database for further analysis. The Ontology Accumulator integrates some similar types of ontology to be accumulated in the cumulative ontology. This Integrated Health Ontology System will pave the way for integrating existing data with ontologies that will be useful for developing semantic agents for healthcare domain.

Keywords: Ontology encoding and generation, database schema, ontology viewing, ontology information extraction and integration, extracting ontology into database, knowledge sharing.

1 Introduction

Most institutions in the healthcare industry in this country store their data in various forms of database management systems. While database management systems are efficient in the storage and retrieval of data, they are mostly proprietary and locked into applications that access them. Some of these data may be confidential; especially the health records of certain patients who wish to remain private. However, some of the information should best be shared within the healthcare community. Ontologies provide a standard method for sharing this information, with the added advantage of including semantics and relationships that enable software agents to interpret the information stored in these repositories. The ability to create and evolve an ontology is vital towards developing a system for sharing healthcare knowledge.

An ontology can be viewed as a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and relationships

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 538–546, 2011.

between them. It is typically represented as a knowledge base which then becomes available to applications that need to use and/or share the knowledge of a domain. Within health informatics, an ontology is a formal description of a health-related domain [1]. In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge [2]. Ontologies play a major role in the development of the Semantic Web [6]. In the context of our research, an ontology is a shared conceptualization that describes the terminology used in a particular domain, which in this case is the healthcare domain. This ontology is expressed using the Web Ontology Language or OWL in short [3]. OWL is based upon the Resource Description Framework or RDF in short [4] [5].

The Integrated Health Ontology System consists of a set of tools designed to manage the creation, evolution, merging of ontologies and extraction of their subsets for various applications. The first of these tools is the Ontology Generator [11], which generates an ontology using metadata from a database management system. The generator is used as the first stage in building a healthcare ontology data store. This ontology can then be integrated into a cumulative ontology. The cumulative ontology contains concepts integrated from various kinds of ontologies. There are little or no tools currently available in the market for data mining on ontologies directly. Hence, we have also developed an Ontology Distiller [12], which can extract a subset of an ontology and produce a database which can be the source data for existing data mining tools. This process is the reverse of the Ontology Generator. We are now currently developing the Ontology Accumulator, which functions as a tool to integrate related sets of ontologies into a single cumulative ontology.

2 The Overall Design

Central to our Health Ontology System is the Cumulative Ontology and the Target Database. The Cumulative Ontology integrates some ontologies that are relevant to the domain, which in our current research, covers patient records, doctors, and diseases. The Target Database stores the data which is referenced by the Cumulative Ontology. The initial setup of the system involves the use of the Ontology Generator, which generates an ontology from an initial database that supplies the primary source of information. The instances from the record data are stored in the Target Database. There are few or no tools available currently to perform data mining on the Cumulative Ontology. Hence, some data may be extracted from the Cumulative Ontology to form a subset database from which it may be possible to perform data-mining using existing tools. The overall design of how the three different tools are related is shown in Figure 1.

The Ontology Accumulator, which is currently in the design stage, is used to integrate additional information from other ontologies.

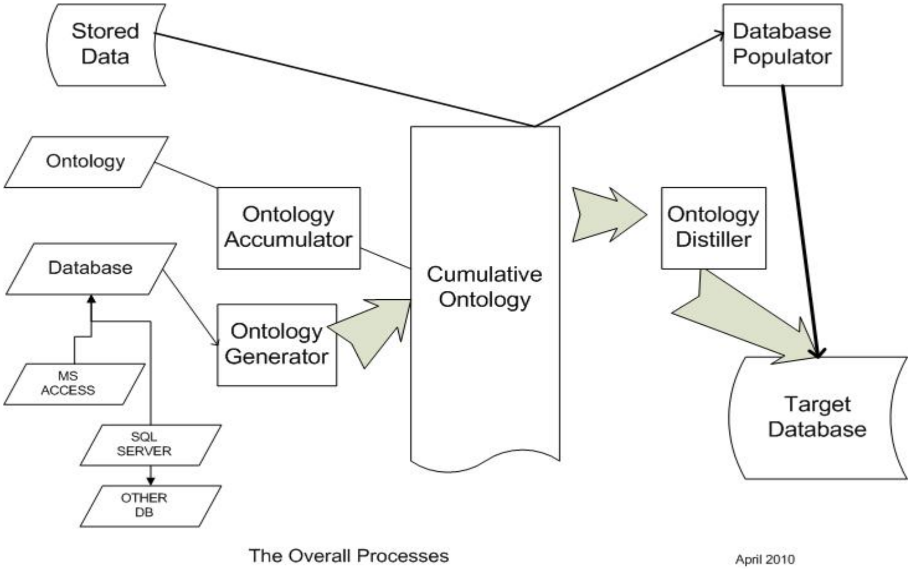


Fig. 1. The Overall Design of the Health Ontology System

3 The Ontology Generator

The Ontology Generator is a tool designed to create an ontology from a selected database. The design is based upon the DataMaster Plugin for Protégé 3.4 [7]. Protégé is written in Java. However, we use a different algorithm that uses C# as the programming language, and the Microsoft Visual Studio as the programming platform. This enables us to use the newer features in C# and rapid prototyping in Visual Studio. We have named it as Health Ontology Generator (HOG) for reference.

3.1 Algorithm to Extract Schema from Database

The first step involves selecting the type of database. We started with Microsoft Access and SQL Server. The connection string is created to connect to the database. For Microsoft Access databases, we use the Microsoft.Jet.OLEDB.4.0 provider; for SQL Server, we use the OLEDB provider. Currently the system supports only Microsoft Access and SQL Server. Support for MySQL and other types of databases will be added later.

HOG uses the schemaTable method to query the tables in a database and returns the result as a DataSet. After the table names are obtained, it extracts the column names and their data types and puts them into another DataSet. Finally, if the user chooses, the row data is extracted into a third DataSet.

3.2 Method for Encoding Ontology

The ontology generation is an automatic process which encodes and stores the ontology physically as an RDF file that includes declarations of classes, properties and instances. In addition, the ontology also includes the semantics that describe the meaning of the data included in it. Typically, the file is given the file extension owl. The first part of the encoding process of ontology is the generation of the header. The body of the ontology includes the classes, the properties and the instances. The final part is the trailer. Figure 2 shows these stages diagrammatically.

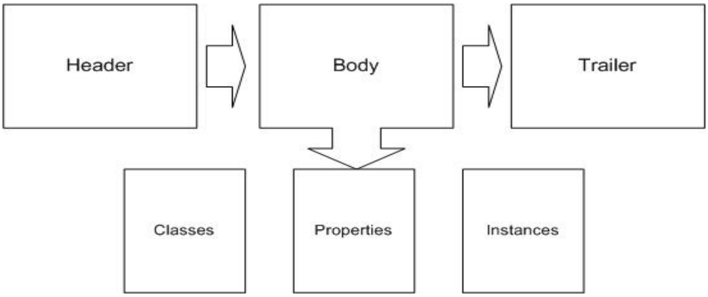


Fig. 2. Stages in encoding the ontology

3.3 Encoding the Header

The header specifies the RDF start tag (with namespace attributes) and the ontology element. It starts with the version information of the XML encoding. This is followed by some standard namespaces, which includes XML schema (for data types), RDF, RDFS and OWL. Each of these standard namespaces is declared using their usual URIs. For example, XMLS is declared as `xmlns:xsd="http://www.w3.org/2001/XMLSchema#"`. The ontology's own namespace is declared as `xmlns: db="http://zhiq.tripod.com /db_table_classes?DSNtype=Access:dbHealth_1#"`, which is a reference to the database to link to the ontology. Finally, the ontology element is declared simply as <owl:Ontology rdf:about=""/>.

3.4 Encoding the Body

In the ontology, tables are converted to classes. This is done by constructing the RDF statement as an OWL class. Field names (or column names) in the table are converted to functional attributes. In addition, other functional attributes are added, which describe the semantics of the ontology. The rows of each table are converted into instances, beginning the first instance in the form of instance_1. The annotation properties, such as #hasForeignKeys are then added.

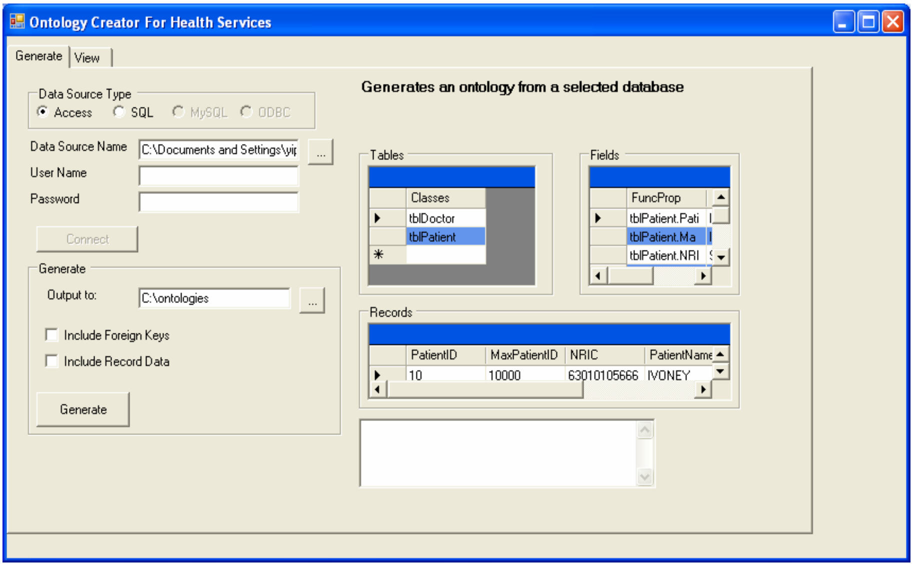
3.5 Encoding the Trailer

The trailer consists of the closing RDF tag and information about the creator of the ontology.

```
</rdf:RDF><!—creator -->
```

3.6 The User Interface of HOG

The User Interface is shown in Figure 3. The user first selects Data Source Type, whether it is MS Access or SQL Server database. Clicking the Connect button will display the Tables, Fields and Records. The user then selects the destination filename to Output the owl file. Clicking the Generate button will generate the owl file in the selected location. The generated owl file is compatible with Protégé 3.4.



Ontology Creator For Health Services

Generates an ontology from a selected database

Data Source Type: ● Access ○ SQL ○ MySQL ○ ODBC

Data Source Name: C:\Documents and Settings\yi

User Name:

Password:

Output to: C:\ontologies

□ Include Foreign Keys

□ Include Record Data

Tables

Classes
tblDoctor
tblPatient

Fields

FuncProp
tblPatient.Pati
tblPatient.Ma
tblPatient.NRI

Records

PatientID	MaxPatientID	NRIC	PatientName
10	10000	63010105666	IVONEY

Fig. 3. The User Interface of HOG

4 The Ontology Distiller

The Ontology Distiller is the reverse process of the Ontology Generator. However, it uses a totally different algorithm. It extracts concepts from an ontology owl file and creates a database. The output database can be a MS Access or a SQL Server database. The current programming requires the database to be created first by MS Access or SQL Server. However, we are working on a direct creation of the database. This database can be operated on by existing data mining tools. We have not been able to access any tool from the Web which can perform a similar task, as we were able to do so with Protégé when we built the Health Ontology Generator.

There is very little work published on the process of creating databases from ontologies. After we have developed the Ontology Distiller, we have found a US patent and a paper describing Knowledge Bus [9], which can create databases from an ontology using the Java platform. This paper uses application program interfaces (APIs) to reflect the entity types and relations (classes and methods) that are represented by the database.

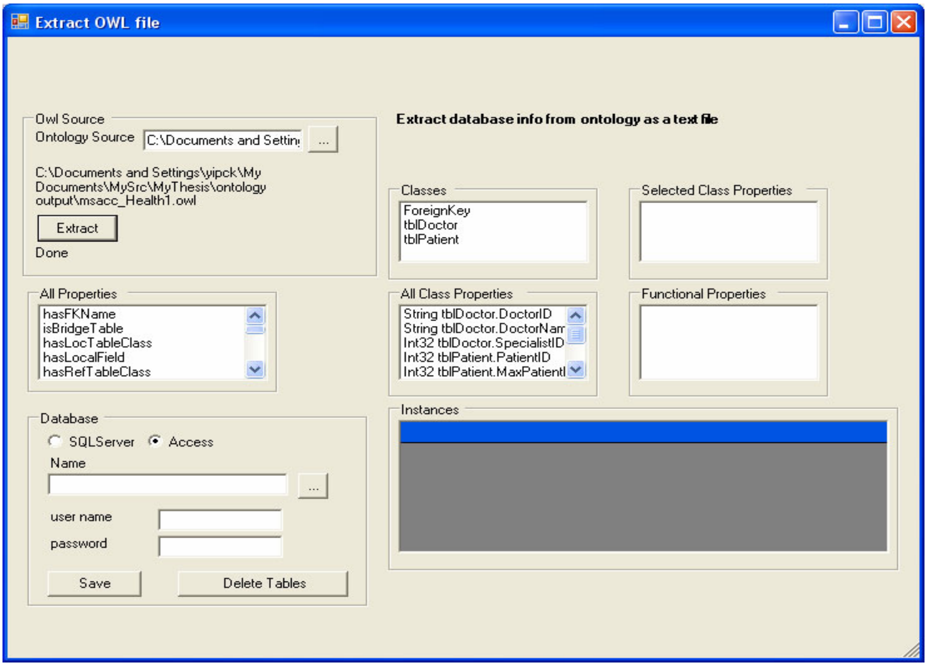
4.1 The Distiller Algorithm

The algorithm used in programming the Distiller involves the following steps:

1. Count the number of classes in the ontology file
2. Get Class properties
3. Build the DataSets
4. Get the Instances
5. Store the data in the database.

Counting the number of classes is necessary to declare the amount of storage space to be allocated for the array of datasets. This is the first pass of the entire encoded owl file before storing any data. The second pass involves allocating the actual classes and their properties and building the datasets. The third pass would read in the instances, their data types and their values. The resulting data would be stored. It may be displayed if necessary.

The process involves some rules that form the basis of the programming. Classes in the ontology are stored as tables in the database. Similarly, functional properties are stored as attributes for each table. Instances from the ontology form the records in the database.



Ontology Source: C:\Documents and Settin

C:\Documents and Settings\yipck\My Documents\MySrc\MyThesis\ontology output\msacc_Health1.owl

Done

Classes:
ForeignKey
tblDoctor
tblPatient

All Properties:
hasFKName
isBridgeTable
hasLocTableClass
hasLocalField
hasRefTableClass

All Class Properties:
String tblDoctor.DoctorID
String tblDoctor.DoctorNam
Int32 tblDoctor.SpecialistID
Int32 tblPatient.PatientID
Int32 tblPatient.MaxPatientI

○ SQLServer ● Access

Fig. 4. User interface of the Ontology Distiller

4.2 The Distiller User Interface

The user interface is shown in Figure 4. First, the user searches for the location of the Ontology Source then clicks the Extract button. The Distiller will extract the classes and the class properties. Then the user can select the type of database to be output. Next is the name of the database.

In the case of MS Access, it is necessary to first create the Access database using Access, and then search for the database. In the case of SQL Server, it is only necessary to name the database. Clicking the Save button will create the tables from the classes, together with the corresponding attributes and records.

5 The Ontology Accumulator

The Accumulator is built as a Microsoft Windows Communications Foundation (WCF) Service and Client. When the Accumulator is initialized, it builds a list of rules for future integration of related ontologies. These are the Accumulator Integration Rules (AIRs). The Accumulator also holds a list of classes that was first built into the cumulative ontology. This is the List of Cumulative Classes (LCCs). Each item in the LCCs points to a List of Cumulative Class Functional Characteristics (LCCFCs). The Accumulator Service provides facilities to read a new ontology and build a list of classes. This is the List of Pending Classes (LPCs). This is followed by procedures to compare the LPCs and the LCCs.

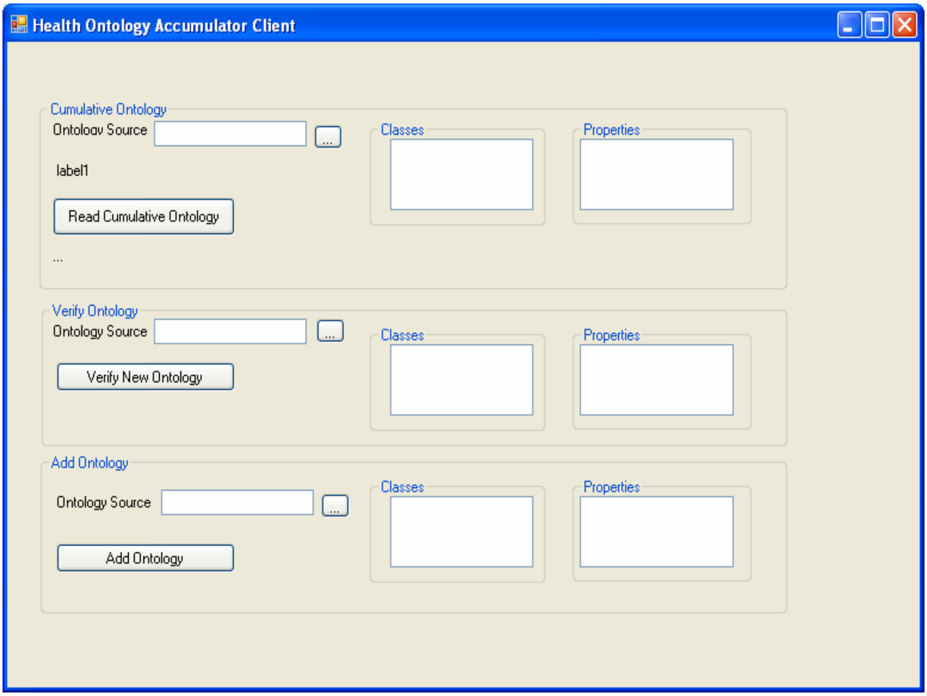


Fig. 5. The Accumulator Client user interface

If an item in the LPC matches one of the items in the LCCs, then no new class is created. The next step is to compare the functional characteristics of the cumulative ontology with the corresponding items in the LCCFCs. Integration will be direct if the items match, and the instances are added to the Target Database. If they do not match, new functional characteristics will be added to the Cumulative Ontology.

Along the same line, if an item in the LPC does not match with any of the items in the LCCs, a new class will be added to the Cumulative Ontology. The rules in the AIRs will be adjusted accordingly.

The Client side of the Accumulator provides the user interface for the Ontology Accumulator, which is shown in Figure 5. The user inputs the name of the cumulative ontology to collect the classes and properties. The user then enters the location of the ontology that is to be verified. Once it is determined that the ontology is suitable for integration, the user can then add the ontology to the cumulative ontology.

6 Conclusion

We have described a Health Ontology System consisting of an integrated set of tools for creating and processing ontologies. The combination of the Health Ontology Generator and the Ontology Distiller that works on the Cumulative Ontology offers the following side benefit. You can start with a MS Access database, use the Ontology Generator to create an ontology, then use the Ontology Distiller to convert that database into an SQL Server database. These are two of the tools of the Integrated Health Ontology System that we have created so far. We are now in the process of the designing and programming of the Ontology Accumulator, which will integrate similar ontologies into the Cumulative Ontology.

References

1. Open Clinical Knowledge for medical care Web Site, http://www.openclinical.org/ontologies.html
2. Protege Overview, http://protege.stanford.edu/overview/
3. McGuiness, D.L., Harmelen, F.V. (eds.): OWL Web Ontology Language Overview, W3C Recommendation, (February 10, 2004), http://www.w3.org/TR/owl-features/
4. Manola, F. and Miller, E (Eds), RDF Primer, W3C Recommendation (February 10, 2004), http://www.w3.org/TR/rdf-primer/
5. Brickley, D. and Guha, R. V (Eds), RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation (February 10, 2004), http://www.w3.org/TR/rdf-schema/
6. Lee, T.B., Hendler, J., Lasilla, O.: The Semantic Web. Scientific American (May 2001)
7. Nyulas, C., O'Connor, M.: Samson Tu, DataMaster – a Plug-in for Importing Schemas and Data from Relational Databases into Protégé, Stanford University School of Medicine, Stanford, CA 94305
8. Wikipaedia, Choosing between versions of Protégé
9. Andersen, W.A., Brinkley, P.M., Engel, J.F., Peterson, B.J.: Ontology for database design and application development, United States Patent 6640231 (2003)
10. Horridge, M., Knublauch, H., Rector, A., Stevens, R., Wroe, C.: Protégé OWL Tutorial, pp. 11–14. The University Of Manchester, Stanford University, United Kingdom (2004)

11. Kiong, Y.C., Palaniappan, S., Yahaya, N.A.: Health Ontology Generator: Design And Implementation, Malaysia University of Science and Technology. IJCSNS International Journal of Computer Science and Network Security 9(2), 104–112 (2009)
12. Kiong, Y.C., Palaniappan, S., Yahaya, N.A.: Ontology Distiller: Extracting Databases From Health Ontologies, Malaysia University of Science and Technology. IJISCE International Journal of Information Sciences and Computer Engineering, to be published 1(2), 68–74 (2010)

A New Model of Securing Iris Authentication Using Steganography

Zaheera Zainal Abidin[1], Mazani Manaf[2], and Abdul Samad Shibghatullah[3]

[1] Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka
[2] Faculty of Computer & Mathematical Sciences, UiTM, Shah Alam, 40450, Selangor
[3] Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka
zaheera@utem.edu.my, mazani@tmsk.uitm.edu.my,
samad@utem.edu.my

Abstract. The integration of steganography in biometric system is a solution for enhancing security in iris. The process of biometric enrollment and verification is not highly secure due to hacking activities at the biometric point system such as overriding *iris template* in database. In this paper, we proposed an enhancement of temporal-spatial domain algorithm which involves the scheme of Least Significant Bits (LSB) as the new model which converts iris images to binary stream and hides into a proper lower bit plane. Here, the *stego key*, *n*, will be inserted into the binary values from the plane which concealed the information; where *n* is the input parameter in binary values which inserted to the *iris codes, m*. These values produce the output which is the new *iris stego* image after binary conversion. Theoretically, the proposed model is promising a high security performance implementation in the future.

Keywords: LSB, PSNR, FAR, stego key, coverimage, message and stegosystem.

1 Introduction

Biometric utilize physical traits (gait and voice recognition) or behavioral characteristics (iris, retina, thumbprint and face) for a reliable identity of authentication. The usage of iris biometric technology and application has increased tremendously for its user friendliness, performance, permanence, accuracy and uniqueness. There are many systems and machines use biometric in daily activities for instance, attendance system, withdrawing money from ATM and thumbprint to switch on laptop. In fact, in biometric, human is the key to access systems. Biometrics data is powerful [1] and useful to the system [2][3]; however, they have no keys [4]. The biometric data is easy to steal [4] or leading to identity theft and not secured [5]. The more a biometric data is used, the less secret it would be [2].

Due to these problems, steganography has been rediscovered and expanding the security performances in iris biometric systems. Steganography is the art and science

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 547–554, 2011.

of hiding information in a cover document such as digital images in a way that conceals the existence of hidden data. The word steganography in Greek means "covered writing" (Greek words "stegos" meaning "cover" and "grafia" meaning "writing"). The main objective of steganography is to hide the true message which is not visible to the observer and securing the database and transmission channel. The intruder should not be able to distinguish in any sense between cover-image and iris stego. Therefore, the iris stego should not diverge much from original cover-image.

Steganography conceals encrypted message from intruders; hiding the information with information to camouflage the observer who unable to comprehend the message.

It differs with watermarking. The technique of watermarking is just to hide the image with another image, with no key. The cryptography itself is insufficient to secure the iris image since it scrambles the message with the encryption algorithm. The encrypted message creates curiosity to the intruders or observers to decrypt the intended message and gives a success to hacker who able to hack the biometric system. Cryptography protects the contents of a message while steganography is to protect both messages and communicating channel. There are cases which, the iris image can be fooled and easily cracked by the hacker.

In [6][7][8][9][10], Discrete Wavelet Transform (DWT) is distinctive and widely used for embedding the iris image which improving the recognition accuracy from tampering. Meanwhile, [6][8] [11][12][13] used the LSB as the embedding scheme. Most of the researchers combine the algorithms to sustain a better security performance. On the other hand, use a single scheme with a different or other property, for instance Haar transformation is used to create the water sign to the respective image [10]. The [11] use the blind extraction with LSB to prevent unauthorized use, and inappropriate user to the system. According to [6], LSB is better scheme comparing with DWT. This is because the PSNR for LSB is 48 while DWT is 9.2.

However, most of the researches are in biometric data watermarking and information hiding. In [11], the iris steganography implementation is only for iris recognition process. The algorithms have been improved [12] however the stego key is not inserted into the binary stream. In this study, we propose a modification to the temporal spatial domain by enhancing the algorithm with the insertion of stego key into the iris codes for both enrollment and verification processes.

2 A General Process of Steganography

The concept of securing the information is almost similar between steganography and cryptography. Cryptography encrypts the information to unreadable codes however; steganography is not only encrypts but hide the information at the same time. Major differences between them are in terms of techniques, applications and technologies.

Three basic techniques used by [15] for steganography are injection, substitution and generation. The first technique of injection is to hides the data (in the form of text, image video or audio) with a cover file (in the form of text, image, video or audio). The second technique is to apply the substitution into the (data+cover file). Here, the Least Significant Bits (LSB) mechanism is useful for embedding information. The least significant bit means the 8th bit of the message is changed to a bit of secret message as in Figure 2. It determines a meaningful content with least distortion. The third technique is generation, which is generating a cover file for solely hiding the information, producing a stego file.

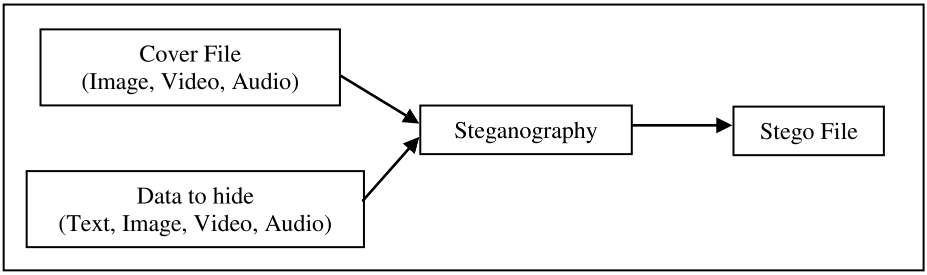
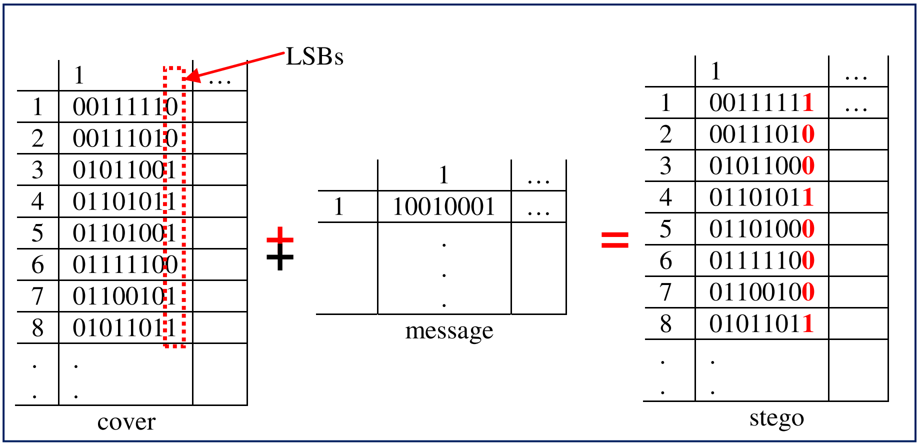


Fig. 1. General process of hiding data [15]



	1	…
1	00111110	
2	00111010	
3	01011001	
4	01101011	
5	01101001	
6	01111100	
7	01100101	
8	01011011	
. .	. .	

	1	…
1	10010001	…
	. . .	

	1	…
1	00111111	…
2	00111010	
3	01011000	
4	01101011	
5	01101000	
6	01111100	
7	01100100	
8	01011011	
. .	. .	

Fig. 2. LSB insertion at the 8th bits

Studies show that there is no established framework of implementing steganography into biometric images. The propose model of securing iris using steganography is explained in section 3.

3 A Propose Model of Iris Steganography

Biometric consists of two general processes which is the enrollment and verification. The enrollment process collects the biometric images, which is the iris image using extraction algorithms. Meanwhile, the verification stage involves matching and detraction algorithms. The integration of the steganography properties is implemented into the biometric enrollment and verification processes, as in Figure 3 and 4; in section 3.1 and 3.2. Steganography has three properties which are:

- key generation algorithm (SK) : takes as input parameter n and outputs a bit string sk, called the stego key.
- steganographic encoding algorithm (SE) : takes as inputs the security parameter n, the stego key (sk) and a message (m), {0, 1} l, to be embedded and outputs an element c of the coverimage space C, which is called iris stego. The algorithm may access the coverimage distribution C.
- steganographic decoding algorithm (SD) : takes as inputs the security parameter n, the stego key (sk), and an element c of the coverimage space C and outputs either a message m {0, 1} l or a special symbol ?. An output value of indicates a decoding error, for example, when SD has determined that no message is embedded in c.

For all sk output by SK(1n) and for all m {0, 1} l, the probability that SD (1n,*sk*,SE(1n, *sk*, m)) 6 = m, must be negligible in n. The syntax of a stegosystem as defined above is equivalent to that of a (symmetric-key) cryptosystem, except for the presence of the coverimage distribution. The probability that the decoding algorithm outputs the correct embedded message is called the reliability of a stegosystem. Indeed, it has no public or private key to be hacked; the key is embedded in the template itself.

The steganographic decoding algorithm is done at the verification process, Figure 4. The verification is a 1:1 matching process, where the user claims an identity and the system verifies whether the user is genuine or vice versa. If the iris code of the claimed identity has a high degree of similarity with the database, then the claim is accepted as “genuine” or else, the claim is rejected and the user is considered as “fraud”. The evaluation testing is going to be implemented to the proposed model. This is for achieving the security performance which is based on the performance parameter, by taking the percentage of False Acceptance Rate (FAR) and the value of Peak-Signal-to-Noise Ratio (PSNR).

In biometric system, FAR measures the percentage of invalid inputs which are incorrectly accepted which it is used to identify between the imposter and genuine user of the system. Meanwhile, in steganography, the performance measurement for image distortion is known as PSNR. The performance of PSNR is in decibels (dB), which providing system’s robustness and the accuracy that benchmarking to the new proposed system. It is expected that the larger PSNR values indicates the higher performance. On the other hand, a smaller PSNR means there is huge distortion between the cover-image and the iris stego. We assume that LSB is better from DWT, since [6] showed in their study, the value of PSNR for LSB is greater than DWT. Therefore, in calculating the value of peak–signal-to-noise-ration (PSNR), which is used for measuring the invisibility of the watermark and normalized correlation (NC).

The definition of PSNR and NC is as follows:

$$PSNR = 10 \text{ X } (\log (255^2/MSE)) \tag{1}$$

$$\text{Where } MSE = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} (\mathrm{f(x,y)} - \mathrm{g(x,y)})^2 / \mathrm{N}^2$$

where f(x,y) and g(x,y) stand for the pixel values of the original iris image and the iris image with secret information.

$$NC(W.W^*) = \frac{\sum_{i=1}^{N} Wi \,.\, Wi^*}{\sqrt{\sum_{i=1}^{N} Wi}^{\,2} \,.\, \sqrt{\sum_{i=1}^{N} Wi}^{*2}} \tag{2}$$

Where W is the original iris code and W* is the extracted iris code. NC is in the range of [0,1], which represents the similarities between W and W*.

The temporal spatial domain is enhanced by inserting the stego key into the binary stream and focus on the BMP format type. [14] provides the formula of brightness : **I = 0.3R + 0.59G + 0.11B**, which define the colour component of green, red and blue. Each colour pixel can be concealed by another data. The effect of brightness is $\mathbf{2^0 \times 0.59 = 0.59}$. If it is the red colour component, the effect of brightness is $\mathbf{2^0 \times 0.3 = 0.3}$. If the fisrt 2 bits of the blue component are altered, the effect on the brightness is $\mathbf{(2^0 + 2^1) \times 0.11 = 0.33}$. Therefore, all of the effects are not greater than the maximum change in brightness in LSB which is 0.59. This means an increase in hiding efficiency about 17% [12]. The stego key is inserted into the binary stream in providing the impact to hiding result.

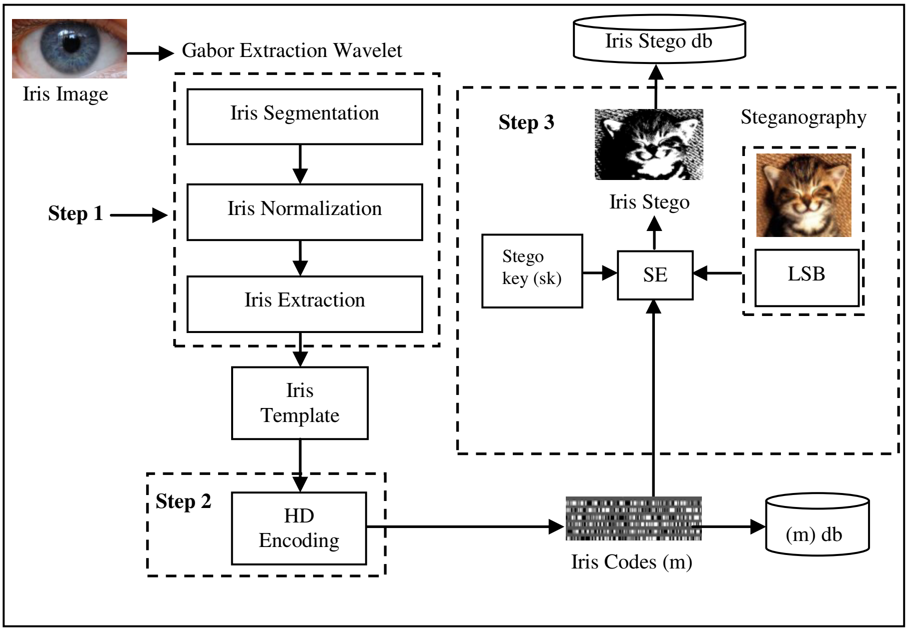


Fig. 3. Iris Enrollment Process

3.1 Iris Enrollment Process

Step 1: The iris image is segmented, normalized and extracted using Gabor Extraction Wavelet. This step produces an iris template.

Step 2: The iris template is encoded with Hamming Distance Algorithm, HD to determine between imposter and genuine. This step produces iris codes.

Step 3: The iris codes is converted in binary before moving to the embedding process. During this step, the cover image (binary) and stego key (binary) is inserted using the LSB algorithm. The value of this combination, gives iris stego in binary.

3.2 Iris Verification Process

Step 1: In de-embedding phase, SD, the iris stego is de-embedded with stego key insertion and LSB algorithm. This process produces iris codes.

Step 2: The iris codes is decoded using Hamming Distance. The decoded iris codes are converted to be iris template.

Step 3: The iris template is extracted using Gabor Wavelet and produce the iris image.

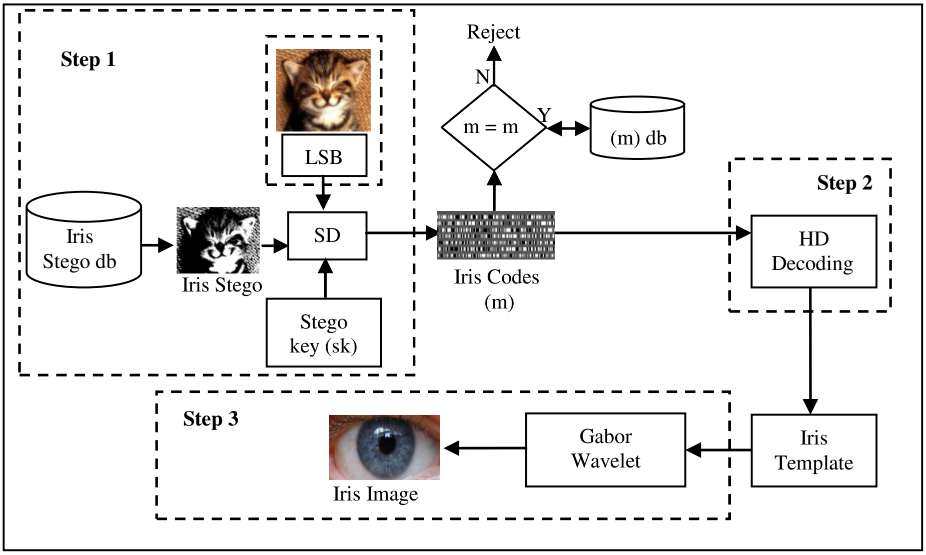


Fig. 4. Iris Verification Process

4 Discussion

How secure is the biometric iris authentication against attacks? Biometric data is easy to steal and has no key. Fake contact lenses with both eyes image on it and use it for illegal activities is one of the trend to attack the system. The biometric threats give different attack at different points in biometric authentication system. The attack launched at the sensor, seize the channel, modify the template and override the biometric templates. Is cryptography process in not sufficient enough to secure the biometric templates, where it has public and private keys to protect the iris template? The encrypted iris code creates a curiosity to the hacker to decrypt the secret keys and most of hackers understand how to do cryptography. Here, we propose the implementation of steganography into the biometric systems which is the new model of temporal-spatial domain algorithm enhancement. A stego-key, coverimage and value of n have been applied to the system during embedment in producing the iris stego. Without the valid key, it is difficult for a hacker to understand the embedded message. However there is a limitation and issues need to be explored in the implementation of LSB on iris images for example the processing time for verifying the genuine. If a longer time takes for a matching process, this gives opportunities for hacker to attack the existing system.

5 Conclusion

Biometric usage is rising and widely accessible in most countries. In fact, the biometric providers have delivered biometric authentication for mobile transactions and client/server based applications. Sustained improvements in the technology will increase performance at a lower cost. However, biometric templates leave traces and traits along the human movements. A new innovation on securing the iris authentication model needs to produce. Therefore, a new model of securing iris is designed with the integration of steganography into the biometric system. The importance of steganography has been apprehended against cryptography and information hiding due to capabilities, security services and performance. Steganography is emphasis on avoiding detection and possibilities of largest hidden massage meanwhile watermarking is robust, emphasis on avoiding distortion of cover file and a small amount of hidden message. The comparison of previous implementations and future model promise a successful achievement. Finally, our contribution of this study is to design and develop iris model for protecting the biometric system against imposter attack, making the iris biometric system more secure with the implementation of steganography.

References

1. Bhargav, A., Squicciarini, A., Bertino, E., Kong, X., Zhang, W.: Biometrics-Based Identifiers for Digital Identity Management. In: ACM International Conference Proceeding Series, Proceedings of the 9th Symposium on Identity and Trust on the Internet, pp. 84–96 (2010)
2. Anderson, R.J.: Security Engineering: A Guide to Building Dependable Distributed Systems. Wiley, New York (2001)
3. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Trans. on Circuits and Systems for Video Technology 14(1), 4–19 (2004)
4. Schneier, B.: The Uses and Abuses of Biometrics. Communications of The ACM 42(8), 136 (1999)
5. Hao, F., Anderson, R., Daugman, J.: Combining Crypto with Biometrics Effectively. IEEE, Transactions on Computers 55(9), 1081–1088 (2006)
6. Xiao, M.-m., Yu, L.-X., Liu, C.-J.: A comparative Research of Robustness for Image Watermarking. IEEE, Computer Science and Software Engineering 6(12-14), 700–703 (2008)

7. VijayKumar, Dinesh: Performance Evaluation of DWT Based Image Steganography. In: IEEE, 2nd International Advance Computing, pp. 223–228 (2010)
8. Fouad, M., Saddik, A.E., Petriu, E.: Combining DWT and LSB Watermarking to Secure Revocable Iris Templates. In: International Conference on Information Science, Signal Processing and their Applications, pp. 25–28 (2010)
9. Zebbiche, K., Khelifi, F., Bouridane, A.: An Efficient Watermarking Technique for the Protection of Fingerprint Images. Journal of Information Security, 20 (2008)
10. Varbanov, G., Blagoev, P.: An Improving Model Watermarking with Iris Biometric Code. In: ACM, International Conference on Computer Systems and Technologies, pp. 5-1 – 5-6 (2007)
11. Das, S., Bandyopadhyay, P., Paul, S., Ray, A.S., Banerjee, M.: A New Introduction Towards Invisible Image Watermarking on Color Image. In: IEEE, International Advance Computing Conference, pp. 1224–1229 (2009)
12. Deng, H., Xie, M., Zhang, L., Yao, Z.: An Improved LSB Information Hiding Algorithm and Its Realization by C#. In: IEEE, International Forum on Information Technology and Application, pp. 759–763 (2009)
13. He, H., Zhang, J., Chen, F.: Block-wise Fragile Watermarking Scheme Based on Scramble Encryption, pp. 216–220 (2007)
14. Lu, H., Wan, B.: Information Hiding Algorithm using BMP Image. Journal of Wuhan University of Technology 28(6), 96–98 (2006)
15. Mehboob, B., Faruqui, R.A.: A Steganography Implementation. IEEE, Los Alamitos (2008)

Tamper Localization and Lossless Recovery Watermarking Scheme

Siau-Chuin Liew and Jasni Mohd. Zain

Faculty of Computer Systems and Software Engineering, Universiti Malaysia Pahang.
Lebuhraya Tun Razak, 26300, Kuantan, Pahang, Malaysia
eliewsc@gmail.com, jasni@ump.edu.my

Abstract. Tamper localization and recovery watermarking scheme can be used to protect the integrity and authenticity of medical images. In this paper, a simple tamper localization and recovery scheme that uses lossless compression was proposed. Lossy compression may also be applied when necessary. The watermarked image has the PSNR of 47.4 dB and the results show that tampering was successfully localized and tampered area was exactly recovered.

Keywords: Lossless compression, tamper localization, recovery, medical image, watermarking.

1 Introduction

Medical images such as radiographs, ultrasound and magnetic resonance images play an important part in the process of diagnosing a patient by medical practitioners. Advancements in medical information system are changing the way patient records are stored, accessed and distributed. Medical images can be stored in a digital form temporarily or permanently on a server along with patient records. Malicious tampering of a medical image for the purpose of insurance claims or to hide a medical condition for personal gains is possible. The integrity of medical images and their information needs to be protected from unauthorized modification or destruction. Current security measures used to protect the integrity of the patient records are such as VPN (Virtual Private Network), data encryption and data embedding [1].

Data encryption is being used on the Internet to protect sensitive data during transmission. It is also being used to protect medical images in the form of digital signature. The problem with digital signature is that it needs to be transmitted together with the image in a separate file or in the image header. Data embedding is where related information such as digital signature can be inserted into the medical images as a watermark. Currently, there is no standard of implementation for digital watermarking.

In practice, diagnoses has been performed on medical images before being directed to long-term storage, thus the clinically relevant part of the image is already been determined by doctors involved in the diagnosing [2]. Here, we refer to the relevant part as the ROI (Region of Interest). Since information in medical images is not to be

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 555–566, 2011.

modified in any way, the watermark is usually being embedded in the RONI (Region of Non-Interest) as this region does not contribute in the diagnosis process.

The ability to localize tampering of a watermarked image is crucial for authentication. Tamper localization is where the area of tampering can be identified. Once tampering is localized, tampered area can be recovered. Wu et al. [3] and Jasni and Abdul [4] divided medical image into blocks and each block is embedded with the authentication message and recovery information of other blocks. Tampered blocks can then be restored using this information. The issue with the current schemes is that the recovery of the tampered area is in the form of average intensity or lossy compression which has lower quality in terms of perception when being compared to the original non-tampered image.

In this paper, a tamper localization and recovery scheme using Joint Photographic Experts Group(JPEG) compression is proposed.

2 Tamper Localization and Recovery Watermarking

One of the requirements of an effective watermark based authentication system is the ability to identify manipulated area or also known as localization where the authentication watermark should be able to detect the location of manipulated areas, and verify other areas as authentic [5]. The tampered area can be recovered using information that is stored as the watermark. The perceptibility of a watermarked image can be judged according to its fidelity. Fidelity measures the similarity between images before or after watermarking. A low fidelity reconstruction is dissimilar or distinguishable from the original. Peak signal-to-noise ratio (PSNR) can be used to measure to the fidelity of the watermarked image with the following equation. A high PSNR represents a high fidelity of a watermarked image.

$$PSNR(dB) = 10\, log_{10} \frac{\max I^2}{MSE}, \tag{1}$$

Chiang et al. [6] proposed a reversible tamper localization scheme with tampered region recovery capability. The image is divided into blocks. The recovery information is generated by taking the average pixel value of each block and embedded as watermark. The watermark is encrypted before the embedding process as a security feature. The whole image can be verified by comparing the retrieved average pixel value from the watermark with the current average pixel value of the image. Any mismatch indicates tampering and tampered region can be localized to an accuracy of 4 x 4 pixels. The tampered block will be recovered using the average pixel value retrieved from the watermark. The advantage of this scheme is that it can be modified to allow a ROI to be defined rather than the whole image for the watermarking process. The recovery information of the ROI is stored as the exact pixel value rather than average pixel values. Images were watermarked and have the PSNR between 36.4 to 40.5 dB.

Osamah and Khoo [7] proposed a scheme that consists of two types of watermark. The first watermark is embedded into spatial domain and the second watermark is embedded into transform domain. The image is divided into 16x16 pixels block. The first watermark consists of patient's data and the hash value of the ROI and is being embedded into the ROI itself by using a modified difference expansion technique. An embedding map of the ROI will be produced to form a second watermark together with compressed recovery information of ROI and average value of each block in the ROI. The second watermark is compressed and embedded into the region of non interest (RONI) using a DWT technique. Tamper localization is done by comparing the average value of each block in the ROI with the retrieved average value from the watermark. Tampered blocks can be recovered using the lossy compressed ROI. It was claimed that this scheme is robust against salt and pepper attack and cropping. A watermarked ultrasound image has the PSNR of 36.7 dB.

Earlier research by Jasni and Abdul [4] had also produced a tamper localization and recovery watermarking. It also uses block based technique where each block consists of 8 x 8 pixels. Each block is then divided into sub-blocks of 4X4 pixels. A three-tuple watermark embedded consists of a two-bit authentication watermark and a seven -bit recovery watermark for the other two sub-blocks. Average intensity of a corresponding block and its sub-blocks is calculated to generate the authentication watermark. Average intensity of a sub-block is embedded as the seven-bit recovery watermark in another block which was predetermined in a mapping sequence. A parity bit is generated based on the seven-bit recovery watermark. Tamper localization is done by comparing the average intensity and the parity bit. Blocks that were marked invalid are recovered using the embedded average intensity of the sub-block. The watermarked ultrasound image has a PSNR of 54.8 dB.

3 Compression

Compression is the process of storing or packing data in a format that requires less space than the initial file. Lossless data compression is a category of data compression algorithm that allows the exact original data to be reconstructed from the compressed data. Lossless compression is suitable for the usage in medical image due to the importance of perceptibility in the process of diagnosis. Some schemes reviewed in this literature had applied lossless compression technique in the watermarking process. Osamah and Khoo [7] compressed the average of block in the ROI for the usage of tamper detection using Huffman coding. They had also used lossy JPEG compression to compress the ROI for recovery purposes. Weng et al. [8] uses arithmetic coding to compress the location map used to facilitate the extraction of the watermark. Other type of lossless compression techniques are run-length encoding (RLE) and JPEG compression.

3.1 Huffman Coding

Huffman coding was developed by David A. Huffman in 1952. Huffman coding is a predictive based compression technique where it removes the redundancy between successive pixels by encoding only the residual between actual and predicted values. Since the residue value usually has a much smaller dynamic range, this leads to fewer encoding bits and causes a compression. Huffman coding is a variable output bit length encoding method in which the input bits are grouped based on their bit value occurrence probability in the signal.

3.2 Arithmetic Coding

Arithmetic coding is a very different algorithm from the Huffman algorithm. Arithmetic coding still uses the probabilities of source symbols. It successively subdivides the interval 0.0 to 1.0 into subintervals based on the probabilities of the source stream. These subintervals are again subdivided as each new source symbols is encountered. The sizes of the subintervals are proportional to the frequency of the symbols in the source stream. As the stream of symbols becomes longer, the interval representing it becomes finer and finer in precision. The number representing the whole stream can be found by choosing a number from this final interval. Arithmetic coding is more powerful than Huffman coding in terms of compression ratio but arithmetic coding is more complex and requires more computer resources.

3.3 RLE

RLE is a simple lossless data compression algorithm. It replaces the sequences of the same data values by a count number and a single value. In a more detailed example, binary data that contains 11111100000111111 is encoded as (6, 1), (5, 0) and (6, 1). This is interpreted as six 1's, five 0's and six1's. The decimal number is then being converted to binary data that reads as (110, 1), (101, 0) and (110, 1). As a comparison, the original binary data has 17 bits and the compressed binary data has only 12 bits. RLE has a very simple algorithm as compare to other compression techniques. But RLE will only work best if it is being applied to images that have large number of identical successive bits.

3.4 JPEG

The JPEG standard is commonly used for lossy compression for digital images. A JPEG file can be created by specifying the degree of compression needed. The highest image quality has the largest file size and vice versa. JPEG has an option to allow lossless compression.

4 Methodology

4.1 Image Preparation

An ultrasound image is divided into ROI and RONI as shown as Fig.1. In this scheme, one rectangle is used for the ROI and eight rectangles for the RONI. Based on the calculations, there are 143,400 pixels in the RONI. If only two bits per pixel are used for watermark embedding, the RONI can only store approximately 286,800 bits of watermark payload. Based on these limitations, the ROI can only be a portion of the non-black area in the image as shown in Fig.1. The ROI has a size of 160 x 300 pixels which equivalent to 384,000 bits. The final total bits from the ROI that needs to be embedded depends on the efficiency of the compression technique used. The RONI is divided into blocks of 2 x 2 pixels.

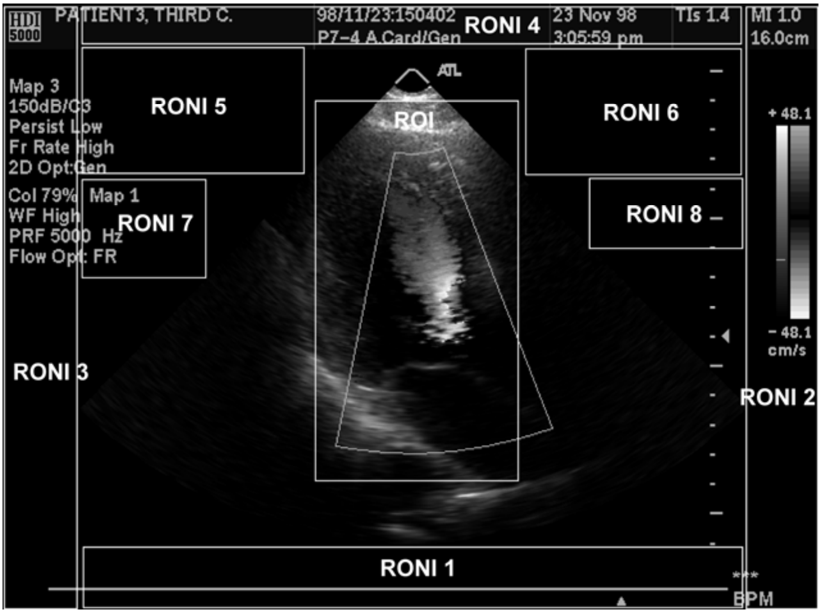


Fig. 1. Ultrasound image is divided into ROI and RONI

4.2 Embedding

The watermark consist of compressed ROI pixels and its hash value. The watermark will be embedded in the LSB and second LSB of each pixel in the RONI.

i. Compressed ROI

The ROI are compressed with JPEG compression algorithm available in MATLAB. The compressed ROIs were saved in a file with a JPEG extension. The file will be embedded in the RONI.

ii. Hash Function

SHA-256 is used to hash the JPEG file before it is being embedded in the RONI. SHA-256 is a non-collision hash function. The hash value can be used to authenticate the JPEG file and other more secure hash function may be use. The hexadecimal hash value, denoted as JPEG_hash_A will be embedded in the RONI together with the JPEG file.

4.3 Tamper Localization And Recovery

The RONI is divided into blocks of 2 x 2 pixels similar to the embedding process.

i. Tamper Localization

The ROI in the form of JPEG file that was embedded in the RONI is retrieved. The file is hashed using the same hash function used in the embedding process, producing hash value, JPEG_hash_B. The embedded JPEG_hash_A is retrieved and compared with JPEG_hash_B. A positive result indicates that the JPEG file retrieved from the RONI is authentic or the RONI had not been tampered.

The retrieved JPEG file is decompressed to form a block of pixel values, denoted as ROI_A. The current pixel values of the ROI are gathered and denoted as ROI_B. ROI_A and ROI_B are compared and difference in value indicates tampering.

ii. Recovery

Tampered pixel in the ROI is being replaced with its corresponding pixel value from ROI_A. The LSBs in the RONI that were used for watermarking embedding are set to zero.

5 Experiment Results

A 8-bit monochrome grayscales ultrasound image measuring 640 x 480 pixels in size. The watermarked image has the PSNR of 47.4 dB. The ROI was losslessly compressed and has an approximate compression ratio of 0.57 or compression output of 216,784 bits. The total watermark payload is 217,040 bits. Fig.2 and Fig.3 shows the original image and watermarked image.

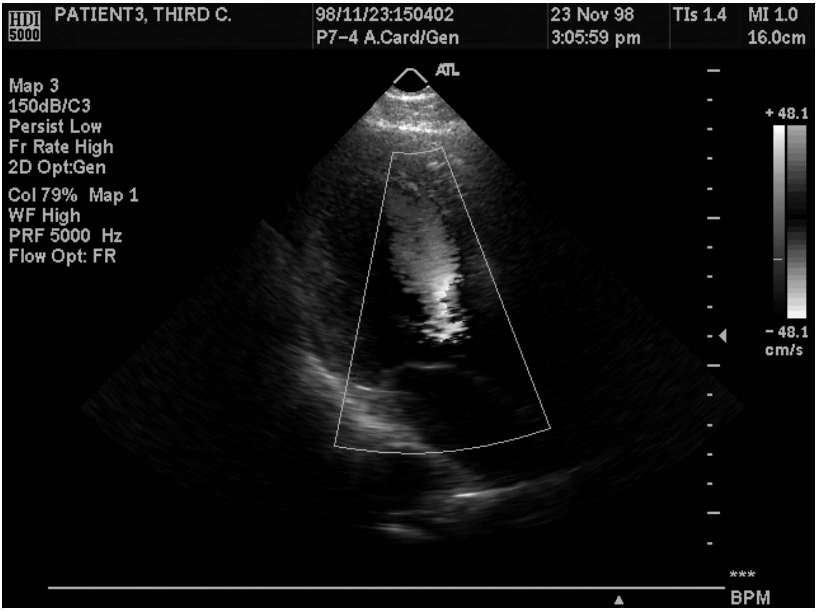


Fig. 2. Watermarked image, PSNR= 47.4 dB

i. Tamper Localization and Recovery

The watermarked image was tampered by modifying the ROI by cloning an area measuring 60 x 90 pixels using ImageJ as shown in Fig.3. The tampered image was recovered as shown in Fig.4. The ROI was fully recovered using the embedded JPEG file. The ROI was also tampered where one pixel value was changed to black as shown in Fig.5 in a magnified version. Fig.6 shows the tampering was detected and the original pixel was recovered.

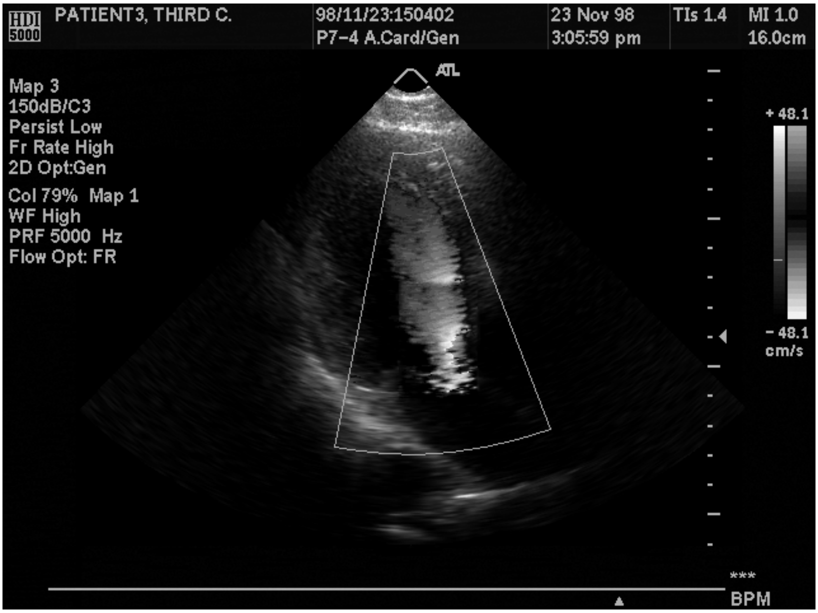


Fig. 3. ROI was tampered by cloning

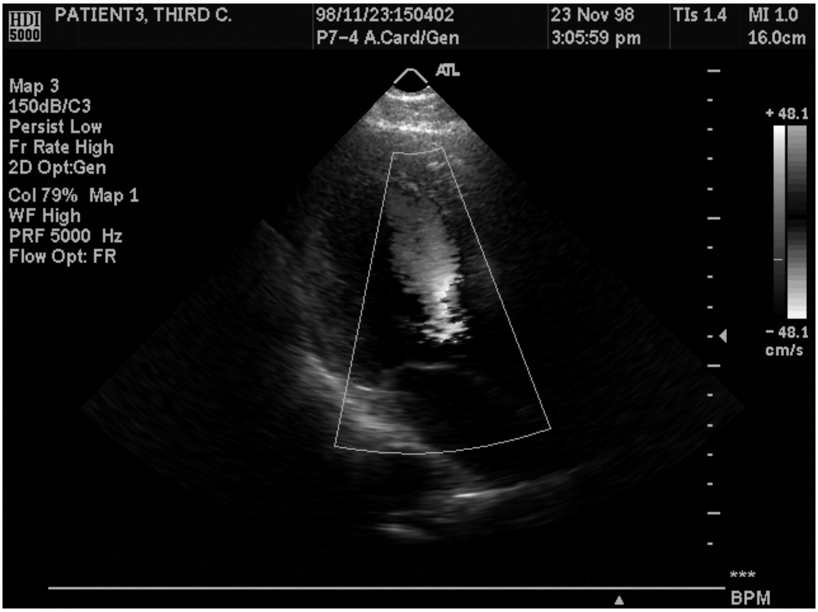


Fig. 4. Recovered tampered image

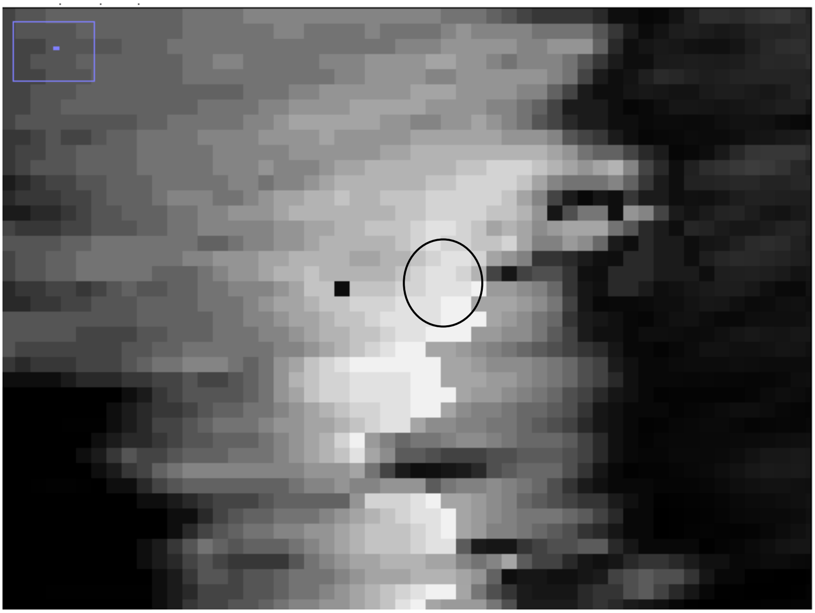


Fig. 5. ROI with 1 pixel tampered in black as highligted

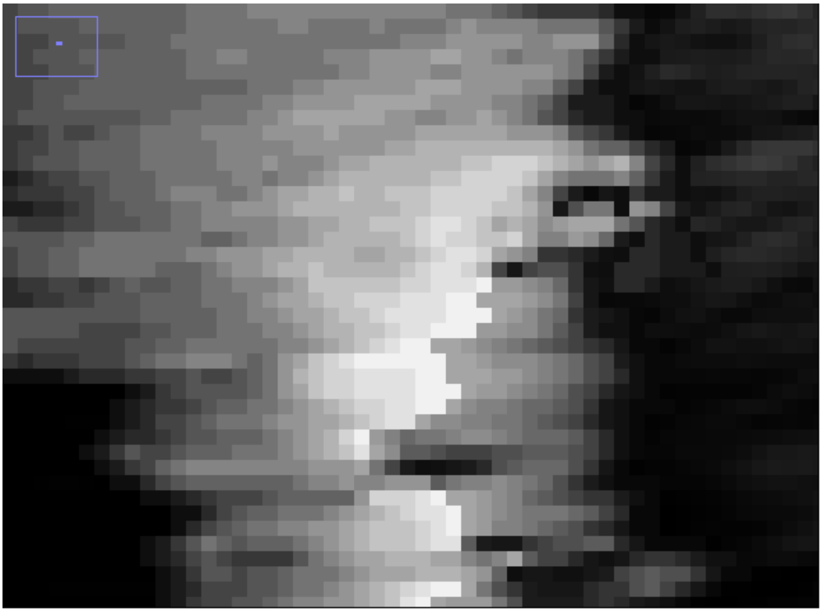


Fig. 6. Recovered tampered ROI

ii. Hash Function Test

The authenticity of the embedded JPEG file in the RONI can be verified by comparing the hash values. The watermarked image was tampered by modifying all RONI pixels to black as shown in Fig.7. This is to demonstrate one of the worst tampering scenarios that may occur in the RONI. The hash value for the JPEG file, JPEG_hash_A embedded in the RONI was retrieved as shown Fig.8. The JPEG file embedded in the RONI was hashed to produce JPEG_hash_B as shown in Fig.9 and being compared to JPEG_hash_A. The hash values were not identical and this indicates tampering of the JPEG file.

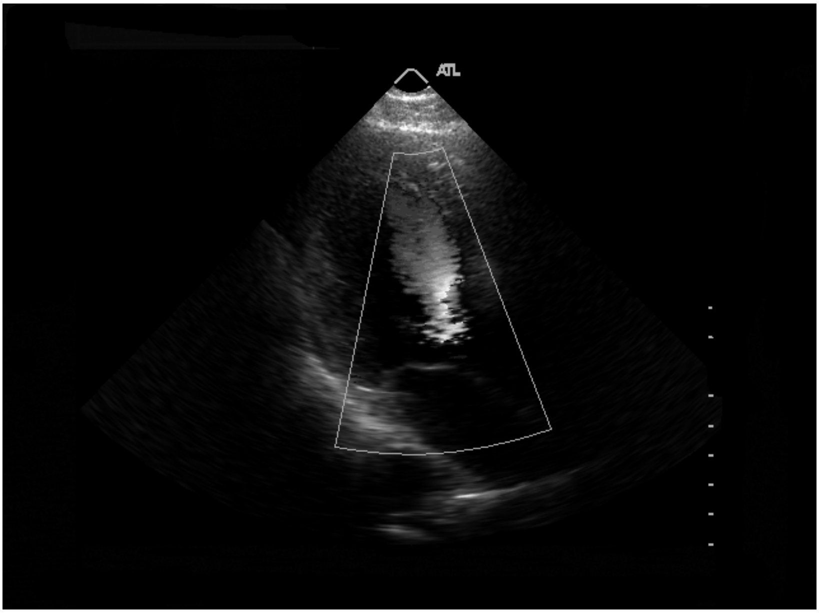


Fig. 7. RONI painted in black

```
e3219b10bb592a18393eb7ec0e0d57a2fbeecf17fc948c994ec9dbe7c8c2a116
```

Fig. 8. JPEG_hash_A retrieved from the RONI

```
dadecd45c7e1f32618f8a641eef75ef6fc48c87b6de09cc6ccb2a1c93833c5dd
```

Fig. 9. JPEG_hash_B computed from JPEG file retrieved from the RONI

iii. Lossy Compression

The ROI is also compressed with lossy JPEG compression that has a quality scale between 0 to 100. The highest scale value applied will produce the highest image quality. The following Fig.10 shows the results of the ROI applied with different quality scales and its output size.

6 Discussion and Conclusion

The fidelity of the watermarked is good with the PSNR of 47.4 dB. The watermark embedding occurs only in the RONI to maintain the originality of the ROI. The RONI of the watermarked image is identical with the RONI of the original image.

The tampered ROI was localized and recovered with a 100% success rate as shown in Fig.4 and Fig.6. The quality of the recovered area was high where the pixels values were retrieved from the JPEG file which was losslessly compressed. The pixel values were the exact values originated from the non-tampered ROI. The recovered ROI may be used for diagnoses purposes due to its high quality.

The RONI was tampered as shown in Fig.7. The comparison result between JPEG_hash_A and JPEG_hash_B were negative. There are two possible scenarios in this situation. Firstly, the possibility of JPEG_hash_A had been tampered and cannot be used to authenticate the embedded JPEG file. The second scenario is that the embedded JPEG file had been tampered and the recovery of the ROI is not be authentic. In either scenario, it can be concluded that tampering in the RONI was detected successfully.

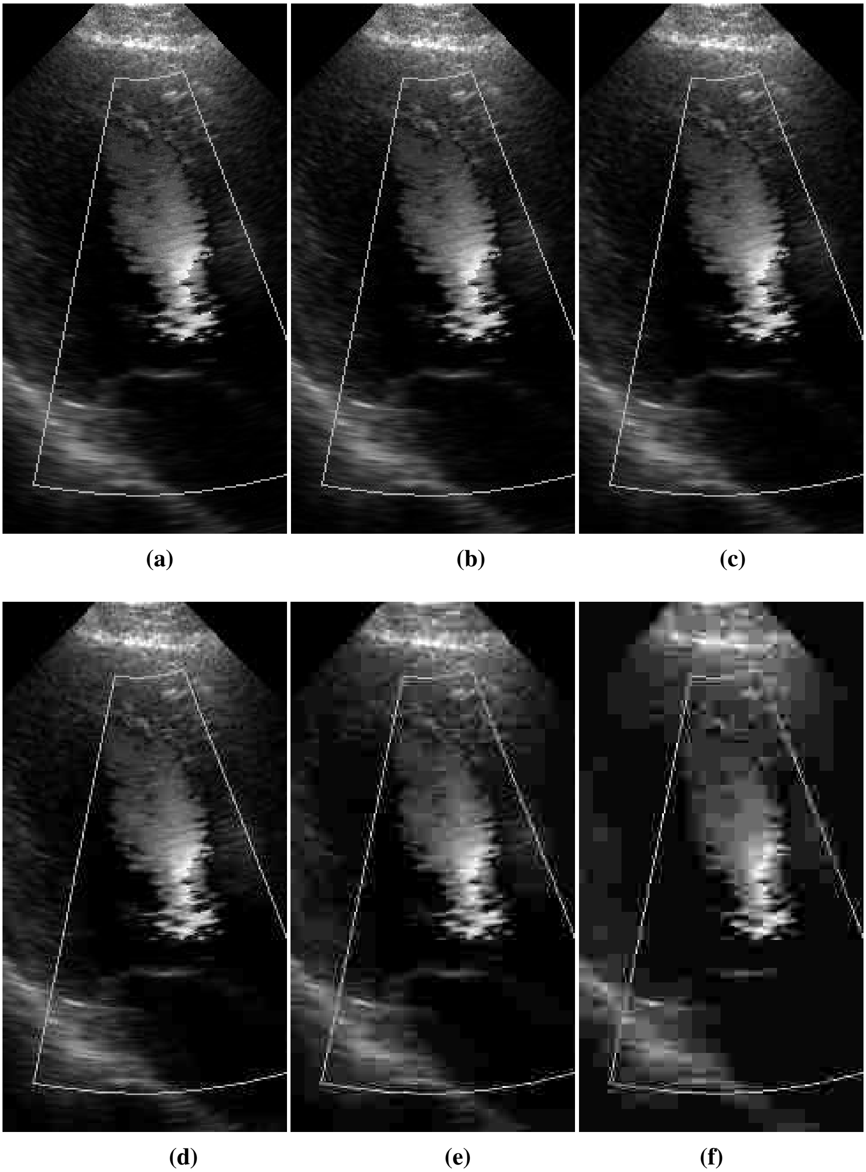
The ROI was losslessly compressed and achieved an approximate compression ratio of 0.57. The ROI was also compressed using lossy JPEG compression with different compression scales. Based on the compression results, there were no noticeable difference between the image quality of the lossless compressed ROI and lossy compressed ROI at the scale of 90. The compression ratio for lossy compression at the scale 90 is approximately at 0.29 as shown in Table 1 which is a significant difference when being compare to the ratio achieved by using lossless compression. In fact, the image quality degradation is only noticeable when the ROI is compressed at the quality scale of 25 as shown in Fig.10. This will allow a larger ROI to be defined if the recovered tampered ROI in a lossy compressed quality is acceptable.

Table 1. ROI applied with different lossy compression quality scale

	Total input (ROI) bits = 384000					
Quality scale	90	75	50	25	10	5
Output size(bits)	110936	68480	47176	31264	17776	12624
Compression Ratio	≈0.29	≈0.18	≈0.12	≈0.08	≈0.05	≈0.03

The performance of the proposed scheme in terms of PSNR is better than the scheme developed by Chiang et al.[6] and Osamah and Khoo[7]. The proposed scheme provides exact recovery for the tampered area which is superior compare to other schemes mentioned in the literature that provides approximate recovery such as average intensity and lossy compressed ROI. The proposed scheme also achieved an excellent tamper localization accuracy of 1 pixel.

As a conclusion, the proposed scheme successfully localizes tampering and recovers the tampered ROI using the exact pixel values which was losslessly compressed. It also has the option to allow lossy compression to be applied when necessary.



(a) **(b)** **(c)**

(d) **(e)** **(f)**

Fig. 10. ROI compressed with different scales **(a)** Scale=90 **(b)** Scale=75 **(c)** Scale=50 **(d)** Scale=25 **(e)** Scale=10 **(f)** Scale=5

Acknowledgments

We would like to thank Research and Innovation Centre of University Malaysia Pahang for the financial support provided for the research work.

References

1. Cao, F., Huang, H.K., Zhou, X.Q.: Medical Image Security in a HIPAA Mandated PACS environment. Computerized Med. Imaging and Graphics. 27, 185–196 (2003)
2. Wakatani, A.: Digital Watermarking for ROI Medical Images by Using Compressed Signature Image. In: 35th Annual Hawaii International Conference on System Sciences (HICSS-35 2002), pp. 2043–2048. IEEE Press, New York (2002)
3. Wu, J.H.K., Chang, R.-F., Chen, C.-J., Wang, C.-L., Kuo, T.-H., Moon, W.K., Che, D.-R.: Tamper Detection And Recovery for Medical Images Using Near-Lossless Information Hiding Technique. J. of Digit. Imaging 21(1), 59–76 (2008)
4. Jasni, M., Zain, A.R.M.: Medical Image Watermarking with Tamper Detection and Recovery. In: 28th Annual International Conference of the IEEE EMBS, pp. 3270–3273. IEEE Press, New York (2006)
5. Liu, T., Qiu, Z.-d.: The Survey of Digital Watermarking-Based Image Authentication Techniques. In: 6th International Conference on Signal Processing, pp. 1556–1559. IEEE Press, New York (2002)
6. Chiang, K., Chang, K., Chang, R., Yen, H.: Tamper Detection and Restoring System for Medical Images Using Wavelet-Based Reversible Data Embedding. J. of Digit. Imaging 21(1), 77–90 (2008)
7. Osamah, M., Khoo, B.E.: Authentication and Data Hiding Using a Hybrid ROI-Based Watermarking Scheme for DICOM Images. J. of Digit. Imaging 24(1), 114–125 (2011)
8. Weng, Shaowei, Zhao, Pan, Y., Jeng-Shyang, Ni, R.: A Novel High-Capacity Reversible Watermarking Scheme. In: IEEE International Conference on Multimedia and Expo., pp. 631–634. IEEE Press, New York (2007)

A New Performance Trade-Off Measurement Technique for Evaluating Image Watermarking Schemes

Mir Shahriar Emami[1], Ghazali Bin Sulong[1], and Jasni Mohamad Zain[2]

[1] Department of Computer Graphics and Multimedia,
Faculty of Computer Science and Information Systems
University Technology Malaysia
81310, Skudai, Johor, Malaysia
shemami85@yahoo.com
[2] Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang

Abstract. In digital watermarking performance evaluation at least three major metrics: imperceptibility, robustness and capacity have been widely used by researchers to analyze the performance of watermarking schemes; however, they constantly conflict with each other. In this paper we propose an effective method to evaluate the trade-off balanced degree among these measures using three threshold conditions. These thresholds comprise of three factors: imperceptibility effect 'before attack and after watermarking', perceptibility effect 'after attack', and robustness 'after attack'. As a result of this study, the performance trade-off of a watermarking scheme can be stated based on degrees. Moreover, we proposed Reset Removal Attack as a severe geometric attack. Finally, the experimental investigation of the proposed technique using the bit-plane watermarking algorithms under several intensities of Reset Removal Attack revealed that the 3rd bit-plane algorithm behaved a better compromise among robustness, image quality, and capacity.

Keywords: Watermarking, Performance Measurement, Security, Robustness, Imperceptibility.

1 Introduction

Nowadays, digital multimedia with the rapid growth of multimedia technologies is largely vulnerable to replication and redistribution via simply accessible networks. In such a situation, digital watermarking techniques have become widely recognized as effective solutions for copyright protection of digital multimedia properties. In this sense, digital watermarking as an approach for owner identification of digital multimedia has become a hot topic of research in computer science which has attracted the interest of several researchers both in industry and academia.

Meanwhile, several digital watermarking technologies and applications have been presented until now which successfully have dealt with robustness, quality and capacity requirements separately; however, none of them gave concrete attention to manage the

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 567–580, 2011.

trade-off among these requirements. In fact, prior to mange this trade-off, an effective technique must be developed to measure this trade-off. Unfortunately, a mechanism to measure the trade-off among watermarking performance metrics has not been appeared yet. The major goal of this paper is to propose a mechanism to measure this trade-off. But let us first discuss watermarking system, attacks and watermarking performance metrics in the next sections.

1.1 Watermarking System

Figure 1 demonstrates the embedding stage in a common watermarking system.

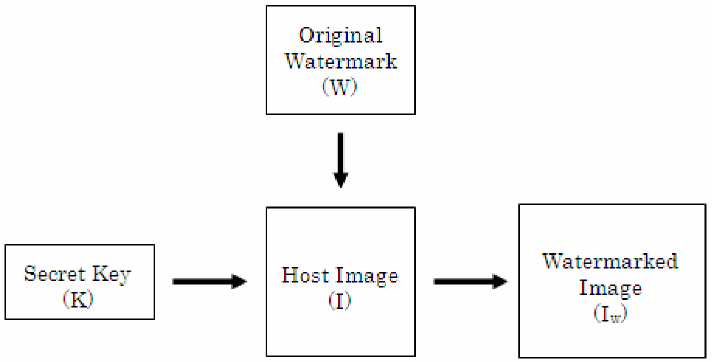


Fig. 1. Embedding Stage in a Watermarking System

In general, a watermarking system is comprised of a host image, a watermark, a secret key, and a watermarked image [1, 2]. Basically, a watermarking system is comprised of two processes: embedding and extracting. At embedding stage the watermark or the encrypted version of it is inserted into the host media. This normally is achieved using a secret key which is predefined confidentially. Let us notify original host image, watermark, secret key and watermarked image by I, W, K, I_w respectively.

Mathematically, a watermark embedding algorithm is a function E such that

$$I_w = \mathrm{E}(I, W, [K])$$

where K is an optional secret key which must be determined confidentially. Watermarking schemes normally applied this secret key; however, considering such a secret key in a watermarking scheme is optional.

At extracting stage the main goal is to extract the embedded watermark using the secret key in order to identify the ownership of the watermarked media (Figure 2).

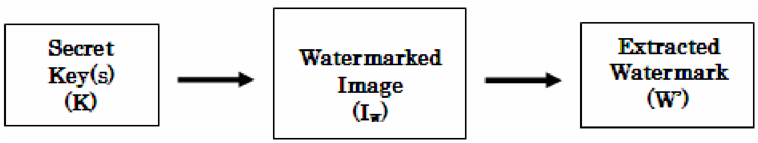


Fig. 2. Extracting Stage in a Watermarking System

Mathematically, a watermark extracting process using blind approach is function T such that

$$W^{'} = \mathrm{T}(I_{w},[K]) \tag{1}$$

where W' is the extracted watermark.

For the non-blind approach the function T can be considered as Equation 2.

$$W^{'} = \mathrm{T}(I_{w},I,[K]) \tag{2}$$

where I, the host image, is available in the time of extracting process.

1.2 Watermarking Performance Measures

Contrary to the steganography in which the embedded information in the host image is important, in watermarking the host image is in the most of the importance so keeping the quality of the host image after watermarking is essential. Moreover, the watermarking algorithm must be robust and able to withstand against intentional and unintentional attacks. Otherwise, embedded owner information hidden in the watermarked multimedia content can easily be detected and destroyed or replaced by malicious users or some software tools intentionally or unintentionally. The common watermarking performance measures are as bellows:

1.2.1 Peak Signal to Noise Ratio (PSNR)

With the purpose of evaluating the imperceptibility of watermarked image, PSNR metric given in Equations 3 has been widely used by other researchers [4, 7, 8, 12-14, 16, 18-20, 24].

$$PSNR = 10 \times \log_{10}\left(\frac{MAX_{I}^{2}}{MSE}\right) \tag{3}$$

where MAX_I denotes the maximum of pixel value in the host image. In an 8-bit gray scale image MAX_I is as Equation 4.

$$MAX_{I} = 2^{8} - 1 = 255 \tag{4}$$

The MSE is the mean square error and is defined as Equation 5

$$MSE = \frac{1}{m \times n} \sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left[(I_{ij} - K_{ij})^{2}\right] \tag{5}$$

where I_{ij} is the value of the pixel (i,j) of the host image, and K_{ij} is the pixel(i,j) value of the watermarked image.

1.2.2 Correlation Factor (ρ)

With the purpose of evaluating the performance of a watermarking system one of the important questions is: Is $W = W^{'}$? This is the question of any robustness metric in a digital watermarking system.

With the aim of evaluating the robustness of the proposed algorithm, NCC[1] parameter has been widely applied [4, 7, 8, 12-14, 16, 18-20, 24]. Other researchers applied another measure to assess the robustness of a watermarking system. Hameed Kamran and et al [23], and Ali Al-Haj [24] used correlation factor (correlation coefficient) ρ (Equation 6) in order to measure the similarity rate between original watermark and the extracted one in their proposed watermarking schemes.

$$\rho(w,w') = \frac{\sum_{i=0}^{L-1} w_i w_i'}{\sqrt{\sum_{i=0}^{L-1} w_i^2 \sum_{i=0}^{L-1} w_i'^2}} \qquad (6)$$

Where w is the original and w' is the extracted watermark bit streams with the size of L.

1.2.3 Usage Capacity Measure

Capacity of a watermarking approach is the maximum data which can be inserted into the host image. Normally this measure represented by percent (%).In a bit-plane algorithm in spatial domain, 12.5% is the maximum capacity when one bit per byte is used for watermark embedding in an 8-bit grayscale host image. Of course if two bit-planes is used the capacity will be doubled i.e. 25%. Basically, the spatial domain algorithms provide more capacities rather than transform domain techniques.

1.3 The Trade-Off Triangle for Watermarking Performance Measures

As mentioned earlier, in digital watermarking schemes at least three major metrics are used: Quality (Imperceptibility), Robustness, and Usage Capacity. Most of these three measures have been used by many of the researchers [2, 5, 6, 16, 12, 13, 17, 19, 21, 37] in order to analyze the performance of the under consideration watermarking scheme; however, robustness and imperceptibility are the most important metrics [37]. All of these researchers implicitly or explicitly notified that there is a trade-off among those three major measures but they did not proposed a mechanism to measure this trade-off. Figure 3 depicts the trade-off triangle of the watermarking performance metrics.

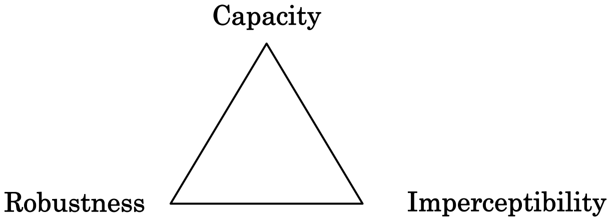


Fig. 3. Watermarking Performance Measures Trade-Off Triangle in Watermarking System

In Conclusion, new approaches are needed on the way to measure the trade-off among the watermarking performance metrics. Now the question is: Is it possible to obtain a technique to measure the trade-off among imperceptibility, robustness, and capacity?

[1] Normalized Cross Correlation.

1.4 Attacks on Watermarking Schemes

Watermark attacks are intentional or unintentional removing, modification or replacement of the embedded watermark. In addition, an unauthorized duplication of a watermarked image can be considered as a watermark attack. The quality of the watermarked image normally is degraded if it is attacked; however, the major goal of the watermarking attacks is to preserve the image quality [22] in order to make possibility of piracy. This can be shown in Figure 4.

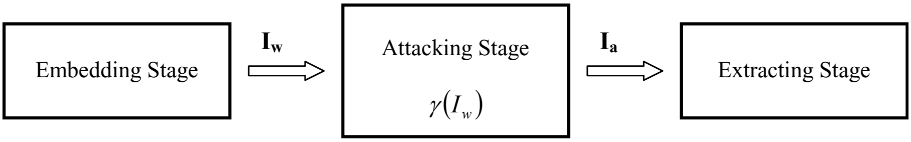


Fig. 4. Attacks on Watermarking Scheme

Mathematically, the attack function γ can be represented as bellow,

$$I_a = \gamma(I_w) \tag{7}$$

where I_w and I_a denote the watermarked and the attacked images respectively.

In general, the attacks on watermarking schemes can be categorized into two main groups: geometric and non-geometric attacks as bellow.

1.4.1 Geometric Attacks

Watermarking geometric attack refers to varieties of attacks which are able to displace all or some of the digital image pixel values by a new amount intentionally or unintentionally [38, 41]. Such attacks are capable of destroy the original watermark so that the ownership of the image can not be identified. Some of the geometric attacks are cropping, scaling, rotating, shifting, bending, warping, perspective projection, collusion, template, Gaussian noise, print-photocopy-scan, etc. One of the severest geometric attacks can be created by resetting all or some of the bits in the bit-plane that is used for watermark embedding. We proposed this attack and called it “Reset Removal Attack”. We used this attack in our experiments.

1.4.2 Non-geometric Attacks

Non-geometric attacks as a wide class of attacks can be classified as Removal, Cryptographic, Protocol and physical attacks [38, 39].

2 Proposed Scheme to Measure the Watermarking Performance Trade-Off

In general, watermarking performance metrics constantly conflict with each other. In this section we propose an effective mechanism to evaluate the trade-off among digital watermarking performance metrics.

In order to measure this trade-off, in this study, different intensities of reset removal attacks were used to attack on the subject watermarked image. In order to analyze the worst case scenarios, it was supposed that the attacker have an aforementioned knowledge of the beginning point of the embedded watermark in all attacking scenarios. By the way, this work attempted to find the PSNR values for both scenarios before attack (PSNRw) and after attack (PSNRa).

After occurring each of attacking procedures the three following threshold conditions should be analyzed in order to measure the performance trade-off balanced degree of the watermarking scheme:

Threshold condition 1:

$$\begin{cases} \rho \geq \rho_{Acceptable} \\ PSNR_w \geq PSNR_{Acceptable} \\ For \quad any \quad PSNR_a \end{cases}$$

In this situation, the watermarked image quality is acceptable and the extracted watermark has a great probability to be identified. Thus the watermarking algorithm can be considered robust.

Threshold condition 2:

$$\begin{cases} \rho \prec \rho_{Acceptable} \\ PSNR_w \geq PSNR_{Acceptable} \\ PSNR_a \prec PSNR_{Acceptable} \end{cases}$$

In this scenario, although the extracted watermark may not be possible to be identified, it is very unlikely that it can be available for piracy because its PSNR value is within the unfavorable level. Thus the watermarking algorithm can be considered robust. In this scenario the watermarked image quality is acceptable too.

Threshold condition 3:

$$\begin{cases} \rho \prec \rho_{Acceptable} \\ PSNR_a \geq PSNR_{Acceptable} \\ For \quad any \quad PSNR_w \end{cases}$$

Here, on the one hand the extracted watermark was hardly recognized, and on the other hand the value of PSNRa is within the acceptable level. This means that the watermarked image is still useable and can be pirated. Thus, the algorithm is non-robust.

Table 1 shows the watermarking performance balanced and unbalanced conditions. A watermarking performance trade-off regarding a watermarking scheme can be considered balanced if it satisfies either threshold conditions 1 or 2. For example, say an algorithm behavior satisfies 30% of threshold condition 1 and 20% of threshold condition 2 so the ‘balanced degree’ of the algorithm will be 0.5 i.e. 50%.

Table 1. Watermarking Performance Balanced and Unbalanced Conditions

	Threshold Condition 1	**Threshold Condition 2**	**Threshold Condition 3**
Watermarking Performance Trade-Off	Balanced	Balanced	Unbalanced

3 Proposed Prerequisites and Experimental Settings

For the evaluating purpose many of the image processing researchers use standard images to be able to compare their results. Among several standard images, in this research with the aim of testing the proposed scheme precisely, the 8-bit gray-scale standard image "Lena" with the size of 512×512 was used. Several smooth areas of this famous image make it enough sensitive for obtaining better results. In this reason, many of the image processing researches use this standard image. Jerome M. Shapiro [40] noted that there are many versions of the Lena. But only one version is an accepted standard. Using non-standard version of testing image may bring about obtaining the results imprecisely. For this reason we used the original version of Lena (Figure 5.a). This image can be downloaded from the website of electrical and computer engineering department of RICE University in USA (http://www.ece.rice.edu/~wakin/images).

In our experiments the host image size, standard Lena, was $512\times512=262144$ pixels. Therefore the maximum capacity of watermark embedding was $\frac{262144}{8}=32768$ pixels when we used only one bit-plane.

Size: $512\times512=262144$ pixels
(a)

Size: $53\times100=5300$ pixels
(b)

Fig. 5. Subject Host Image (Standard Lena) and the Embedding Watermark

The size of the watermark image was $53\times100=5300$ pixels (Figure 5.b). So the maximum capacity was approximately 6 times of such size of the watermark because $\frac{32768}{5300}\cong6$. Table 2 shows the host image and bit-plane usage capacities while one bit per pixel value of the host image has been used to embed the watermark. This table also shows that in a bit-plane algorithm, 12.5% is the maximum capacity when one bit per byte is used for watermark embedding in an 8-bit grayscale host image. Because $\frac{100\%}{8}=12.5\%\cdot$

Table 2. Host Image and Bit-plane Usage Capacities

Host Image Usage Capacity	Bit-plane Usage Capacity	Number of Embedding Watermarks
12.5%	100%	6
10.4%	83%	5
8.4%	67%	4
6.3%	50%	3
4.1%	33%	2
2.3%	17%	1

In this study we applied Enhanced ISB Watermarking Algorithm [13] or EISB for short as the testing platform. EISB approach is based on spatial domain watermarking. In this approach in order to embed a bit of watermark information, the nearest pixel value to the original pixel is chosen whereas this pixel value delivers the watermark bit. For this purpose for the chosen bit-plane algorithm, a specific number of sub ranges must be produced in advance.

Table 3 shows the sub ranges concern the eight bit-plane algorithms based on the embedding position values. Now suppose that we use 3rd bit-plane algorithm, the embedding watermark bit is zero and the embedding position value is 66 then the pixel value of 66 is located inside the sub range of [64..95]. This embedding position delivers this watermark bit automatically because the pixel value in this position is zero i.e. 66=(01000010)B. Therefore the algorithm “does nothing” here.

Now consider the algorithm and the embedding pixel value to be the same but the watermark bit to be one, it can be drawn that this pixel value is not located in any of the sub ranges whose embedding position delivers value of one. Nevertheless, this pixel value is close to the sub range of [32..63] (Table 3). In this sub range the value of 63 is the nearest value to the pixel value of 66. Therefore the value of 63 can be inserted into the original pixel. Thus, the value of 63 is the new value of the pixel in the watermarked image which delivers the watermark bit of one. But if the enhanced ISB watermarking algorithm was not used, the original pixel value of 66 (01000010) would be converted to the new pixel value of 98 (01100010).

Table 3. Pixel value Sub Ranges in Enhanced ISB watermarking algorithm

Bit-plane	Number of Ranges	Pixel Value Sub Ranges	
		Sub Ranges in which Embedding position has value of 0	Sub Ranges in which Embeddingposition has value of 1
1st (MSB)	2	[0..127]	[128..255]
2nd	4	[0..63] [128..191]	[64..127] [192..255]
3rd	8	[0..31] [64..95][128..159][192..223]	[32..63][96..127][160..191][224..255]
4th	16	[0..15][32..47][64..79]…[224..239]	[16..31][48..63][80..95]…[240..255]
5th	32	[0..7][16..23][32..39]…[240..247]	[8..15][24..31][40..47]…[248..255]
6th	64	[0..3][8..11][16..19]…[248..251]	[4..7][12..15][20..23]...[252..255]
7th	128	[0,1][4,5][8,9][12,13]…[252,253]	[2,3][6,7][10,11][14,15]…[254,255]
8th (LSB)	256	[0][2][4][6][8][10][12]…[254]	[1][3][5][7][9][11][13][15]…[255]

In this study, in order to evaluate the proposed scheme a trade-off among digital watermarking performance measures is investigated. Measuring the balanced degree regarding the trade-off among these metrics was the favorable point of this study. The pre-determined boundaries are necessary to measure this trade-off. For this reason several thresholds for each of these metrics must be obtained.

In addition, although a specific international standard value for the PSNR has not been determined yet, some researchers recommended 34 dB (decibel) [13-14], while others suggested 30 dB [15-16]. In this research, based on the previous research in [13-16], we considered (8) and (9) for “favorable” and “unfavorable” PSNR thresholds respectively.

$$PSNR\Big|_{Favorable} \geq 34dB \tag{8}$$

$$PSNR\Big|_{Unfavoarable} \prec 29dB \tag{9}$$

The range between (8) and (9) was considered “acceptable” as it can be shown in (10).

$$29dB \leq PSNR\Big|_{Accaptable} \prec 34dB \tag{10}$$

Finally, according to Man Xuan and Jia-Guo Jiang, if the similarity rate is greater than 0.6, a watermarking algorithm can be considered robust [12]. Similarity rate of about 0.75 is considered acceptable by other researchers [17-19]. In fact the higher the similarity rate value, the more desirable it would be. Similarity rate grater than or equal to 0.7 is considered acceptable in this paper. In this study, the “acceptable” range of ρ is given in (11).

$$0.7 \leq \rho\Big|_{Acceptable} \leq 1 \tag{11}$$

4 Analysis and Discussions

In this study 30 different attacking scenarios (Points) for each bit-plane watermarking schemes had been tested. Figure 6 depicts the graphical representations of our experiments on the subject image using eight bit-plane watermarking algorithms: 1st bit-plane (MSB), 2nd bit-plane, 3rd bit-plane,…, 8th bit-plane (LSB).

Figure 6.h depicts the LSB algorithm behavior against reset removal attacks. This figure revealed that when 17% or less of bit-plane capacity was used, the watermarked image is not able to resist against any reset removal attack. The rationale was that the ρ values were 0.0 for any level of attacks. Likewise, the same condition was also fulfilled by the following scenarios:

(i) The bit-plane capacity usage was 33% and the attack level was more than 20%, the maximum ρ was about 0.0 but the minimum PSNRa was 59.

(ii) The bit-plane capacity usage was 50% and the attack level was more than 40%, the maximum ρ was about 0.0 but the minimum PSNRa was 51.

(iii) The bit-plane capacity usage was 67% and the attack level was more than 60%, the maximum ρ was about 0.0 but the minimum PSNRa was 51.

(iv) The bit-plane capacity usage was between 83% and 100% and the attack level was more than 80%, the maximum ρ was about 0.0 but the minimum PSNRa was 51.

Similarly Figures 6.e, 6.f and 6.g revealed that the 5th to 7th bit-plane algorithms behaved in the same manner as that of the 8th bit-plane (LSB). Therefore, these algorithms were also considered non-robust.

Moreover, Figure 6.d revealed that the 4th bit-plane algorithm performed much better than all of 5th to 8th bit-plane algorithms in terms of robustness. When less than 50% level of reset removal attack had been used the maximum ρ values were more than acceptable-level. But for attack level of 50% or more than that, the ρ values were less than the acceptable level. In this situation, although the watermark information was far from recognition, the PSNRa was within the $\text{PSNR}_{\text{Unfavorable}}$. Therefore the attacked images were considered useless and unproductive for piracy.

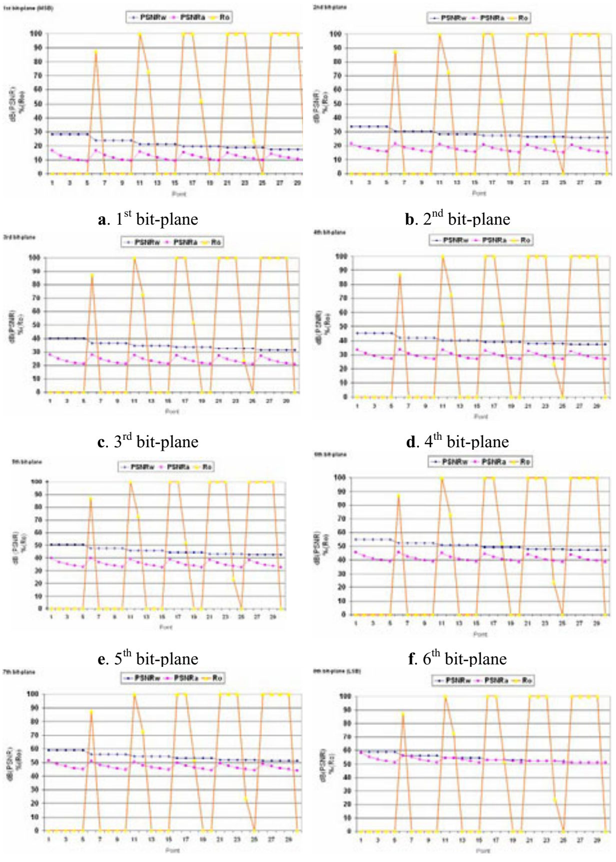
Furthermore, Figure 6.c proved that the 3rd bit-plane algorithm performed extremely well for most situations. Therefore it was stronger than 4th bit-plane algorithm in terms of robustness. Moreover, Figure 6.c also shows that the 3rd bit-plane algorithm was acceptable enough in terms of degradation because in most situations the PSNR ‘after watermarking before attack’ was more than $\text{PSNR}_{\text{Acceptable}}$.

In addition, Figure 6.a and 6.b illustrate that both of 1st and 2nd bit-plane algorithms performed well in terms of robustness; however, the PNSR values ‘after watermarking before attack’ (PSNRw) decreased, and dropped down below the acceptable level. It can be interpreted that during the watermarking process, both of 1st and 2nd bit-plane algorithms drastically degraded the quality of the host image.

Finally, from Figure 6, it can be understood that there is a direct relationship between robustness and usage capacity if we embed several copy of the same watermark. This means that the robustness will grows when the usage capacity increases by embedding more than one copy of an arbitrary watermark. Let us see the following scenarios in Figure 6.d. (i) The bit-plane capacity usage was 17%, for any attack level the maximum ρ was 0.00%, (ii) The bit-plane capacity usage was 33% and the attack level was 20% then the maximum ρ was about 0.87, (iii) The bit-plane capacity usage was 50% and the attack level was 20% then the maximum ρ was about 0.73, (iv) The bit-plane capacity usage was 67% and the attack level was between 20% and 40% then the maximum ρ was 1.0, (v) The bit-plane capacity usage was 83% and the attack level was between 20% and 60% then the maximum ρ was 1.0, (vi) The bit-plane capacity usage was 100% and the attack level was between 20% and 80% then the maximum ρ was 1.0.

Legend

Scenario No (Point)	Capacity	Attack
1	17%	20%
2	17%	40%
3	17%	60%
4	17%	80%
5	17%	100%
6	33%	20%
7	33%	40%
8	33%	60%
9	33%	80%
10	33%	100%
11	50%	20%
12	50%	40%
13	50%	60%
14	50%	80%
15	50%	100%
16	67%	20%
17	67%	40%
18	67%	60%
19	67%	80%
20	67%	100%
21	83%	20%
22	83%	40%
23	83%	60%
24	83%	80%
25	83%	100%
26	100%	20%
27	100%	40%
28	100%	60%
29	100%	80%
30	100%	100%



a. 1st bit-plane

b. 2nd bit-plane

c. 3rd bit-plane

d. 4th bit-plane

e. 5th bit-plane

f. 6th bit-plane

g. 7th bit-plane

h.8th bit-plane

Fig. 6. Attacking scenarios for each bit-plane watermarking schemes

Figure 7 which summarizes the results in Figure 6.a to 6.h depicts the balanced degrees of the watermarking performance trade-off regarding the mentioned algorithms. This figure obviously proved that the 3rd bit-plane algorithm performed extremely well for most situations because the corresponding performance trade-off balanced degree was equal to 0.90. This means it satisfied 90% for both threshold conditions 1 and 2. Figure 7 also illustrates that the 4th bit-plane algorithm satisfied 83% for both conditions 1 and 2 because the balanced degree was 0.83. Moreover, Figure 7 revealed that 1st bit-plane algorithm could not balance the trade-off among watermarking performance metrics. Finally, this figure illustrates that the 2nd, 5th, 6th, 7th and 8th bit-plane algorithms weakly balanced the performance trade-off.

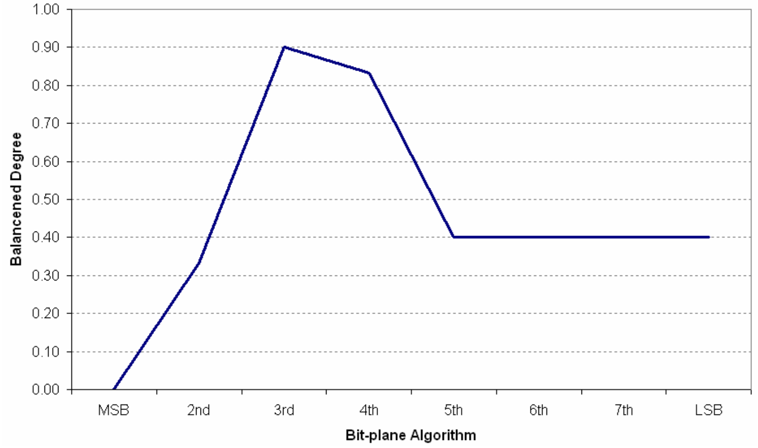


Fig. 7. Watermarking Performance Trade-off Balanced Degree of the eight bit-plane algorithms

5 Conclusions and Future Works

This paper proposed a technique to evaluate the balanced degree regarding the trade-off among watermarking performance metrics. The proposed technique revealed that in order to make a balanced trade-off among the quality, robustness and capacity, three factors viz. PSNR 'before attack after watermarking', PSNR 'after attack', and ρ or NCC 'after attack' should be considered all together using three threshold conditions. We tested the proposed method on the bit-plane watermarking algorithms under one of the severest geometric attacks. The results of the experimental investigations on the proposed method illustrated that the 3rd bit-plane algorithm is the most superior among the bit-plane algorithms which has produced optimal results that provides a balanced trade-off among robustness, imperceptibility and capacity.

In addition, as the future perspective, this study can be continued by testing other attacks such as pepper and salt, blurring, JPEG, etc. Finally, the proposed approach can be applied in evaluating the performance trade-off of spectral domain watermarking approaches such as DCT[2], FFT[3], and DWT[4].

References

1. Taoa, P., Eskicioglub, A.M.: A robust multiple watermarking scheme in the Discrete Wavelet Transform domain. In: Proc. SPIE, Internet Multimedia Management Systems V, Philadelphia, PA, USA, vol. 5601, p. 133 (2004)
2. Al-Otum, H.M., Samara, N.A.: A robust blind color image watermarking based on wavelet-tree bit host difference selection. In: Signal Processing, vol. 90, pp. 2498–2512. Elsevier, Amsterdam (2010)

[2] Discrete Cosine Transform.
[3] Fast Fourier Transform.
[4] Discrete Wavelet Transform.

3. Schyndel, R.G.V., Tirke, A.Z., Osborne, C.F.: A digital watermark. In: Proceeding of the 1st IEEE Image Processing Conference, Houston TX, November 15-17. RMIT, pp. 86–90 (1994), `http://goanna.cs.rmit.edu.au/~ronvs/papers/ICIP94.PDF`
4. Eugene, P.G.: Digital watermarking of bitmap images. In: Proc. International Conference on Computer Systems and Technologies, June 14-15, pp. 1–6. ACM Press, Rousse (2007)
5. Shieh, J., Lou, D., Chang, M.: A Semi-blind digital watermarking scheme basedon singular value decomposition. In: Computer Standards & Interfaces, vol. 28, pp. 428–440. Elsevier, Amsterdam (2006)
6. Wu, N.I., Hwang, M.: Data Hiding: Current Status and Key Issues. International Journal of Network Security 4(1), 1–9 (2007)
7. Burdescu, D.D., Stanescu, L., Ion, A., Mihaescu, C.M.: Spatial Watermarking Algorithm for Video Images. In: Burdescu, D.D., Stanescu, L., Ion, A., Mihaescu, C.M. (eds.) Computer Network Security- Communications in Computer and Information Science, vol. 1, pp. 402–407. Springer, Heidelberg (2007)
8. Ozturk, M., Akan, A., Cekic, Y.: A Robust Image Processing in the Joint Time-Frequency Domain. EURASIP Journal on Advances in Signal Processing, Hindawi Publishing Corporation (2010)
9. O'Ruanaidh, J., Pun, T.: Rotation Scale and Translation Invariant Digital Watermarking. In: Proc, IEEE Int' Conf. on Image Processing, pp. 536–538. IEEE Computer Society, Los Alamitos (1997)
10. Barni, M., Bartoloni, F., Cappellini, V., Piva, A.: A DCT-domain System for Robust Image Image Watermarking. Signal Processing 66(3), 357–372 (1998)
11. Voloshynovskiy, S., Deguillaume, F., Pereira, S., Pun, T.: Optimal Adaptive Diversity Watermarking with Channe State Estimation. Proc,SPIE: Security and Watermarking of Multimedia Contents (2001)
12. Zeki, A.M., Manaf, A.A.: Robust Digital Watermarking Method based on Bit-Plane Ranges. Studies in Informatics and Control Journal, Romania (September 2007)
13. Zeki, A.M., Manaf, A.A.: A Novel Digital Watermarking Technique Based on ISB (Intermediate Significant Bit). International Journal of Information Technology 5(3) (2009)
14. Yang, C.: Inverted pattern approach to improve image quality of information hiding by LSB. Pattern Recognition 41, 2674–2683 (2008)
15. Chan, C., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recognition 37, 469–474 (2004)
16. Maity, S.P., Kundu, M.K.: Robust and blind spatial watermarking in digital image. Proc. of the 3rd Indian Conference on Computer Vision, Graphics and Image Processing (December 2002)
17. Dejun, Y., Rijing, Y., Yuhai, Y., Huijie, X.: Blind Digital Image Watermarking Technique Based On Intermediate Significant Bit and Discrete Wavelet Transform. In: Proc. of the International Conference on Computational Intelligence and Software Engineering, CISE 2009, IEEE Computer Soceity, Los Alamitos (2009)
18. Habes, A.: Information Hiding in BMP image Implementation, Analysis and Evaluation. Information Trnsmissions In Computer Networks, Tom. 6(1) (2006)
19. Mehemed, B.A., El-Tobely, T.E.A., Fahmy, M.M., Said Nasr, M.E.L., El-Aziz, M.H.A.: Robust digital watermarking based falling-off boundary in corners board-MSB-6 gray scale images. International Journal of Computer Science and Network Security 9(8), 227–240 (2009)
20. Rabah, K.: Steganography – The Art of Hiding Data. Information Technology Journal 3(3), 245–269 (2004)
21. Fazli, S., Khodaverdi, G.: Trade-off between Imperceptibility and Robustness of LSB Watermarking using SSIM Quality Metrics. In: Proc, Second International Conference on Machine Vision, IEEE Computer Society Press, Los Alamitos (2009)

22. Voloshynovskiy, S., Pereira, S., Pun, T.: Watermark Attacks. In: Proc. Erlangen Watermarking Workshop (1999)
23. Kamran, H., Mumtaz, A., Gilani, S.A.M.: Digital image watermarking in the wavelet transform domain. In: Proc. World Academy of Science, Engineering and Technology, vol. 13, pp. 86–89 (2006)
24. Al-Haj, A.: Combined DWT-DCT digital image Watermarking. Journal of Computer Science 3, 740–746 (2007)
25. Piao, C., Woo, D., Park, D., Han, S.: Medical Image Authentication Using Hash Function and Integer Wavelet Transform. Proc, IEEE Computer Society, Congress on Image and Signal Processing (2008)
26. El-Ghoneimy, M.M.: Comparison Between Two Watermarking Algorithms Using DCT Coefficient, And LSB Replacement. Journal of Theoretical and Applied Information Technology, 132–139 (2008)
27. Kung, C., Chao, S., Tu, Y., Yan, Y., Kung, C.: A Robust Watermarking and Image Authentication Scheme used for Digital Contant Application. Journal of Multimedia 4(3) (June 2009)
28. Cheung, N.W.: Digital Image Watermarking in Spacial and Transform Domain. In: Proc, TENCON, pp. 374–378 (2000)
29. Eggers, J.J., Su, J.K., Girod, B.: Robustness of a Blind Image Watermarking Scheme. In: International Conference on Image Processing (ICIP 2000), Canada (2000)
30. Bennour, J., Dugelay, J.L., Matta, F.: Watermarking Attack: BOWS Contest. In: Proceeding of SPIE (February 2007)
31. Wu, N.: A Study on Data Hiding for Gray-Level and Binary Images, Master Thesis, Chaoyang University of Technology, Taiwan (2004)
32. Xuan, M., Jiang, J.: A Novel Watermarking Algorithm in Entropy Coding Based on Image Complexity Analysis. In: Xuan, M., Jiang, J. (eds.) Proc. of the International Conference on Multimedia Information Networking and Security, MINES 2009, pp. 128–129 (2009)
33. Muhammad, S.S., Dot, Y.: A watermarking scheme for digital images using multilevel wavelet composition. Malaysian Journal of Computer Science 16, 24–36 (2003)
34. Voloshynovskiy, S., Pereira, S., Herrigel, A., Baumgartner, N., Pun, T.: A generalized watermark attack based on stochastic watermark estimation and perceptual remodulation. In: Wong, P.W., Delp, E.J. (eds.) SPIE Proceedings IS&T/SPIE's 12th Annual Symposium, Electronic Imaging 2000: Security and Watermarking of Multimedia Content II, San Jose, CA, USA, vol. 3971, pp. 23–28 (2000)
35. Voloshynovskiy, S., Pereira, S., Iquise, V., Pun, T.: Attack modelling-towards a second generation watermarking benchmark. Signal Processing 81, 1177–1214 (2001)
36. Voloshynovskiy, S., Pereira, S., Iquise, V., Pun, T.: Attack modelling-towards a second generation watermarking benchmark. Signal Processing 81, 1177–1214 (2001)
37. Song, C., Sudirman, S., Merabti, M., Llewellyn-Jones, D.: Analysis of Digital Image Watermark Attacks. In: 2010 7th IEEE Proc. Consumer Communications and Networking Conference (CCNC), IEEE Computer Society Press, Los Alamitos (2010)
38. Kougianos, E., Mohanty, S.P., Mahapatra, R.N.: Hardware Assisted Watermarking for Multimedia. Computers and Electrical Engineering 35, 339–358 (2009)
39. Meerwald, P., Pereira, S.: Attacks, applications, and evaluation of known watermarking algorithms with Checkmark. In: Proc. Security and Watermarking of Multimedia Contents IV. SPIE, vol. 4675, pp. 293–304 (2002)
40. Shapiro, J.M.: Embedded Image Coding Using Zerotrees of Wavelet Coefficients. IEEE Transactions on Signal Processing 41(12) (1993)
41. Licks, V., Jordan, R.: Geometric Attackes on Image Watermarking Systems. IEEE Multimedia, 68–78 (July-September 2005)

Malaysian Car Plates Recognition Using Freeman Chain Codes and Characters' Features

Nor Amizam Jusoh[1] and Jasni Mohamad Zain[2]

[1] School of Science & Technology, IKIP International College, Kampus 3
25050 Kuantan, Pahang, Malaysia
[2] Faculty of Computer Systems & Software Engineering, University Malaysia of Pahang,
25000 Kuantan, Pahang, Malaysia
amizamjusoh@yahoo.com,
jasni@ump.edu.my

Abstract. Car plate recognition (CPR) system is an important application of image detection and recognition used to overcome the challenges of monitoring modern day traffic. The Freeman chain codes (FCC) is applied in this research to study on the accuracy and efficiency of this technique as an alternative recognition approach to recognize characters on car plates because of its ability to recognize characters and digits successfully. Therefore, this paper is mainly focused on conducting an experiment using FCC to test on its accuracy and efficiency in characters recognition with the support of characters' features because FCC alone does not provide an accurate and efficient recognition result. The result shows that the combination of FCC with characters' features (FCCwF) yields 95% recognition accuracy. As a conclusion, FCC can be an accurate and efficient technique only if the image quality is high and without any noise which might disturbing the recognition process or if it is combined with another technique.

Keywords: Image processing, recognition, chain codes, segmentation, features.

1 Introduction

Many applications have been developed and implemented which applies image processing and character recognition technology. One of the most popular and important applications of digital image processing which focus on image detection and recognition is car plate recognition (CPR) system which is used to overcome the challenges of monitoring modern day traffic. This technology identify vehicles automatically by reading and recognizing their car plates which may involve in traffic violations, crime scenes, car parks area determination, statistical and etc. The plate recognition is done based on the given conditions and instructions. This recognition system is installed in many places such as toll gates, parking lots and also entrance of highly secured buildings. Police are using this application because they can detect speeding vehicles from distance away and summon tickets will be issued later to the vehicle's owner. These systems are beneficial because it can automate car park

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 581–591, 2011.

management, improve the security of car park operator and the users as well, eliminate the usage of swipe cards and parking tickets, improve traffic flow during peak hours, detect speeding cars on highways, and detect cars which run over red traffic lights. Furthermore, when the data gathered by a CPR system is stored and organized in a database, more complex information-driven tasks may be achieved, such as vehicle travel time calculations, marketing analysis, and border control.

The research and development (R&D) for car plate recognition system has been carried intensively by many researchers all over the world. Normally, when conducting R&D in this area, a general methodology which consists of several phases is applied together with the appropriate techniques in order to ensure that the R&D will meet its objectives. The general phases are image acquisition, image pre-processing, image segmentation and image recognition and each phase is performed continuously. Image acquisition is a phase for obtaining samples of images while image pre-processing is a phase for enhancing the quality of images obtained from the first phase. Image segmentation is a phase to segment images whether to isolate object from background or to separate characters on car plates. The final phase is image recognition where in this phase, a recognition technique or combination of several techniques is applied to recognize the characters on car plates. The recognition techniques are varied from simple to complex techniques and most researchers are focusing on the application of template matching technique [1, 2, 3, 4, 5] and neural networks [6, 7, 8, 9, 10, 11]. One of the CPR researches has been conducted using stroke analysis or Freeman chain code (FCC) [12]. However, according to the findings, this work has not been published anywhere. Due to that, another research using FCC technique has been conducted by [34]. Recent researches on CPR have been performed using neural networks by [13] which achieved high recognition rate for normal characters. However, recognition by using this technique for Chinese characters was failed. So, this has become a motivation to continue research on CPR since FCC is a shape-representation technique which is very good in representing shapes or objects information with curves. The advantages of this technique also can be considered in proposing FCC as an alternative recognition technique.

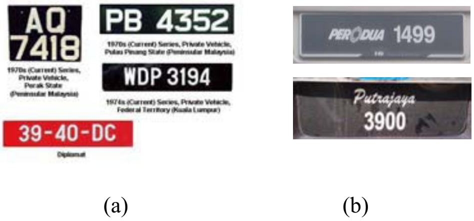

AQ 7418
1970s (Current) Series, Private Vehicle, Perak State (Peninsular Malaysia)
PB 4352
1970s (Current) Series, Private Vehicle, Pulau Pinang State (Peninsular Malaysia)
WDP 3194
1974s (Current) Series, Private Vehicle, Federal Territory (Kuala Lumpur)
39-40-DC
Diplomat
PERODUA 1499
Putrajaya 3900
(a)
(b)

Fig. 1. (a) Samples of common Malaysian car plates. (b) Samples of special Malaysian car plates with various styles of fonts and sizes

The Department of Road and Transport of Malaysia has endorsed a specification for car plates that includes the font and size of characters that must be followed by car owners. However, there are cases where this standard is not being followed. Private car owners tend to use various kinds of fonts and sizes for their car plates. Fig. 1

below shows samples of common and special Malaysian car plates. This various fonts and sizes of characters will lead to problems during recognition phase. One of the factors that contribute to the failure in achieving 100% accuracy in recognition was unable to recognize similar pattern characters such as in the case of recognizing character 'B' or '3' as '8', 'O' and 'D' as '0', 'G' and '5' as '6', 'A' as '4' and 'I' as '1' [7,8,9,10].

Therefore, this paper will focus on conducting an experiment by applying chain codes technique for recognition to be considered as an alternative solution for recognizing characters in Malaysian car plates. Single line Malaysian car plates will be used as testing images.

2 Related Works

There are many techniques applied by different researchers in the recognition phase. The research in this area started with the application of syntax forcer by [24]. The researches keep on growing progressively due to the need to improve the recognition rate with less processing time and to be applied with latest technology for real time use. According to [14], recognition is the process that assigns a label to an object based on its descriptors. This is the last phase in car plate recognition and it is really dependent on how segmentation is done. Thus it is really important to ensure that the earlier phases performed have produced accurate results or else it will give bad effect during recognition.

The mostly used and popular techniques for recognition are template matching [5, 7, 15] and neural networks [11, 16, 17] while other applied techniques are such as multi-methods [4, 18], hybrid methods [19], Hausdorff Distance [20], combination of minimum distance classifier and neural network method [21], combination of statistical model and template matching [22], statistical feature extraction [23] and Freeman chain codes [12, 34]. Due to the different types of characters used in car plates such as the combination of Chinese and Roman characters [4, 19] makes the use of multi-methods or hybrid methods become popular in recognizing car plates. Other reasons why the combination of several techniques becomes a preference are the ability to improve the computation time which is very essential in any real-time car plate recognition system. Besides that, another advantage is the ability to show high percentage of recognition accuracy.

[4] claims that the recognition using multi-methods can increase the correction rate and reduce the computation time. The multi-methods are the combination of projection sequence feature matching and template matching to recognize Chinese characters and Roman characters in car plate while hybrid method [19] are the combination of statistical and structural recognition methods. The result shows that the method applied is more effective and robust while recognition using multi-methods suffers in terms of processing time [18]. The combination of proposed three steps in the recognition approach: the character categorization, topological sorting and self-organizing (SO) recognition only improved the recognition rate but not the time complexity. This was due to the neural-based OCR process running on a sequential computer.

Many researchers applied single technique to perform recognition of characters in car plates such as Hausdorff Distance [20] to recognize Thai car plates proven to achieve high percentage of recognition accuracy but needed more research to get high performance in recognizing poor quality images and among groups of similar-pattern characters. Another research used neural networks [21] to test on Malaysian car plates however does not shows a very impressive result due to having some kind of 'mustache' portions generated in the thin line formation and less training samples were used. [23] proposed on recognition using statistical feature extraction rather than artificial neural networks because this technique considers on hardware and time optimization. The recognition result by using this technique is 85% with 2 seconds taken for recognition time.

The studies show that recognition techniques applied are varied from a simple technique such as template matching to a complex one such as neural networks as well as the recognition using distance measurement. Neural networks and template matching are the most popular techniques to be used since both deliver its own advantages to show high recognition rate. However, the studies also show that some techniques have their drawbacks such as artificial neural network [21, 25] which shows quite good recognition rate but long processing time and it needs periodical training for better accuracy while [7] claims that RBF neural network has disadvantage on processing Chinese character because it has to consider more interference such as fouling and occlusion. [26] in their research found that template matching technique shows high recognition rate but it requires efficient searching method and it needs large storage area to save all the numbers and characters templates. In other major studies, fuzzy logic system [26, 27] has been used to recognize the plate's segmented elements and this technique shows high performance and recognition rate and less processing time but sensitive for the noise and distortion.

The evolution of the research and application of various techniques was due to the need to study on the ability of the techniques to improve recognition performance by recognizing similar pattern characters. The combinations of recognition techniques or modifications for certain techniques become the solutions for some researchers to recognize various types of car plates in the whole world or to apply with real time systems in order to gain processing speed without losing too much performance.

3 Chain Codes

Chain codes are one of the shape representations which are used to represent a boundary by a connected sequence of straight line segments of specified length and direction. This representation is based on 4-connectivity or 8-connectivity of the segments [14]. The direction of each segment is coded by using a numbering scheme as shown in Figure 2 below. Chain codes based from this scheme are known as Freeman chain codes.

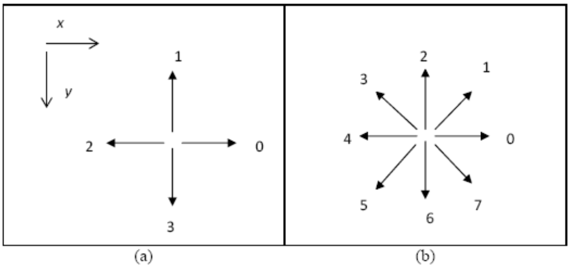
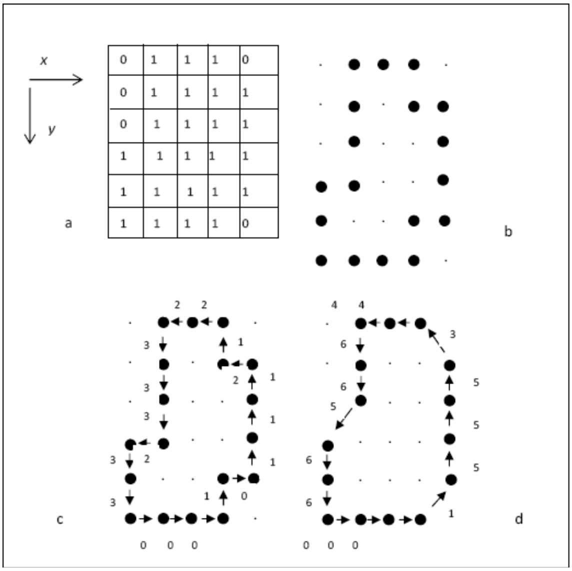


Fig. 2. Direction numbers for (a) 4-directional chain codes, (b) 8-directional chain code

A coding scheme for line structure must preserve the information of interest, must permit compact storage and convenient for display and must facilitate any required processing [28]. According to [29], chain codes are a linear structure that results from quantization of the trajectory traced by the centers of adjacent boundary elements in an image array. A chain code can be generated by following a boundary of an object in a clockwise direction and assigning a direction to the segments connecting every pair of pixels.

First, we pick a starting pixel location anywhere on the object boundary. Our aim is to find the next pixel in the boundary. There must be an adjoining boundary pixel at one of the eight locations surrounding the current boundary pixel. By looking at each of the eight adjoining pixels, we will find at least one that is also a boundary pixel. Depending on which one it is, we assign a numeric code of between 0 and 7 as already shown in Figure 2. For example, if the pixel found is located at the right of the current location or pixel, a code "0" is assigned. If the pixel found is directly to the upper right, a code "1" is assigned. The process of locating the next boundary pixel and assigning a code is repeated until we came back to our first location or boundary pixel. The result is a list of chain codes showing the direction taken in moving from each boundary pixel to the next. The process of finding the boundary pixel and assigning a code is shown in Figure 3. However, this method is unacceptable for two main reasons; the resulting chain of codes tends to be quite long and any small disturbances along the boundary due to noise or imperfect segmentation cause changes in the code that may not be related to the shape of the boundary.

Chain codes have been claimed as one of the techniques that are able to recognize characters and digits successfully [30]. This is because of several advantages possessed by this technique as listed by [31]. The first advantage over the representation of a binary object is that the chain codes are a compact representation of a binary object. Second, the chain codes are a translation invariant representation of a binary object. Due to that, it is easier to compare objects using this technique. The third advantage is that the chain code is a complete representation of an object or curve. This means that we can compute any shape feature from the chain codes. According to [32], chain codes provide a lossless compressing and preserving all topological and morphological information which bring out another benefit in terms of speed and effectiveness for the analysis of line patterns.



0	1	1	1	0
0	1	1	1	1
0	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	0

Fig. 3. (a & b) A 4-connected object and its boundary; c & d) Obtaining the chain code from the object in (a & b) with (c) for 4-connected and (d) for 8-connected

4 Proposed Approach

The research methodology has been divided into two modules where the first module consists of image acquisition, data definition, pre-processing and segmentation phases and three phases are included in this module; the chain codes derivation, characters' features extraction and shape recognition. For recognition purpose, the technique applied is Freeman chain code with characters' features (FCCwF). Freeman chain codes have shown to be very effective in recognizing hand-written digits, and it is assumed that the technique might work well on car plate characters.

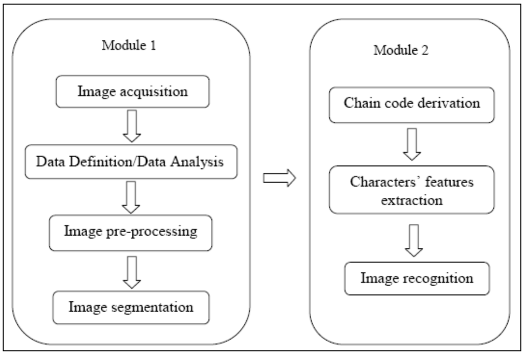


Fig. 4. The proposed methodology

4.1 Segmentation

Two processes have been done in this phase; the boundary extraction and segmentation. The boundary image of car plate extraction is done in order to ease the process of deriving the chain codes. The segmentation phase or character isolation takes the region of interest (from the boundary image) and attempts to divide the region into individual characters. To help in detecting the characters, the plate image is divided into seven images where each will contain one isolated character. For the purpose of research, only car plate images which contain 7 characters, with 3 letters at the position of C1, C1 and C3 and 4 numbers at the position N1, N2, N3 and N4 will be used. The segmentation has been done using the pixel count technique first while the connected component labeling technique has been performed for other images which were failed to be segmented using previous technique.

4.2 Chain Code Derivation

This phase is to derive the chain codes for each character or number in the specific region which is the result from image segmentation phase. The algorithm for extracting chain codes for 8-connected boundaries is as follows [33]:

1. Find the leftmost value in the topmost row pixel in the object name this pixel P_0.
2. Define a variable *dir* (for direction), and set it to equal to 7(since P_0 is the top-left pixel in the object, the direction to the next pixel must be 7).
3. Traverse the 3x3 neighborhood of the current pixel in a counter-clockwise direction, beginning the search at the pixel in direction dir + 7 (mod 8) *if dir is even* or dir + 6 (mod 8) *if dir is odd.* This will sets the current direction to the first direction counter-clockwise from dir:

dir	0	1	2	3	4	5	6	7
dir + 7 (mod 8)	7	0	1	2	3	4	5	6
dir + 6 (mod 8)	6	7	0	1	2	3	4	5

4. The first foreground pixel will be the new boundary element. Update *dir.*
5. Stop when the current boundary element P_n is equal to the second element P_1 and the previous boundary pixel P_{n-1} is equal to the first boundary element P_0.

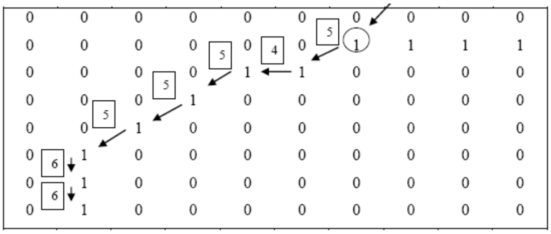


Fig. 5. The initial location, P_0 and the direction to derive chain codes

```
Columns 1 through 19
 5  4  4  5  4  5  4  5  5  6  6  5  6  5  6  6  6  6  6
Columns 20 through 38
 6  6  6  6  6  6  6  6  6  6  6  6  6  6  7  6  6  6  6
Columns 39 through 57
 6  7  6  7  7  7  7  7  0  0  7  0  0  0  0  0  0  1  0
Columns 58 through 76
 1  0  1  1  1  2  1  2  2  1  2  2  2  3  4  4  4  4  5
Columns 77 through 95
 6  6  6  5  5  5  5  4  4  4  4  4  3  3  3  2  3  2  2
Columns 96 through 114
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
Columns 115 through 133
 2  1  1  1  1  0  0  0  7  0  7  7  7  6  7  6  7  0  0
Columns 134 through 152
 0  1  2  2  2  2  3  2  2  3  3  3  3  4  3  4  4  4  3
Columns 153 through 154
4  4
```

Fig. 6. The chain code extracted from the boundary image of character ‘C

4.3 Characters’ Features Extraction

Each character has its own features and these features can be extracted and used as contributing factor in recognizing characters. Based on the experiments, three features are suitable to be applied in this research since the combination of them gives a unique feature for each character. The features require that the characters have the same size in pixels, and therefore the images with the characters are initially normalized to the same height and width. The shape of each character is still preserved even when it is normalized into smaller size. The features that are used in this experiment are listed below.

a. The number of pixel ‘1’ in the new segmented image
b. The total number of all pixels in the new segmented image
c. The average value of new segmented image.

To extract the unique features for each character, a technique is proposed by preparing a constant matrix with value ‘0’, ‘0.5’ and ‘1’. This constant matrix has the same size as the smallest segmentation image. Then, each segmented character’s image is normalized to the same size of constant matrix to ease the process of finding the result of multiplying this constant to each character’s image. The process of multiplication is done between the constant matrix and the segmented image and it is done by column and row. From the multiplication process, a new matrix of segmented image is produced with new values. Based on this new segmented image, the three features can be calculated which later is used for recognition purpose. Table 1 below shows sample of features values extracted for characters.

Table 1. Features extracted for some characters

Character	Total pixel '1'		Total all pixels		Average value	
	Min	Max	Min	Max	Min	Max
A	56	83	386	583.5	6.64	7.48
B	59	119	436.5	797	6.77	7.75
C	40	86	230.5	658.5	7.08	8.11
0	51	88	333	613	6.43	7.24
1	64	111	314.5	660.5	4.26	6
2	55	99	340	678	6.1	7.5

From the table, we can see that different techniques give different result in terms of the segmentation accuracy. Pixel count technique can only achieved up to 80% while connected component labeling is up to 93%. Eventhough the rate is quite high, however the best technique should be able to reach almost 100% accuracy. However, the proposed segmentation technique able to increase the segmentation rate by more than 96%. In terms of recognition accuracy rate, using FCC only achieved up to 80%. This is considered as low rate since our aim is to achieve almost 100% accuracy. Recognition time for each character is about 0.01s to 0.07s. and varied with the differentiation in types of fonts but most characters have same recognition time regardless its font types. For example, character '1' has recognition time of only 0.01s for all types of fonts. Template matching technique only able to show recognition rate of 90.63% while the proposed technique able to increase the rate of up to 95%.

From the analysis done in previous work [34], the used of FCC alone does not able to yield a high accuracy rate since this technique easy to be disturbed by noises occurred in the segmented images. The noises here are the pixels '1' which is considered as misallocated where its location is not where it is supposed to be. These noises affected the chain codes derived thus will affected the recognition process directly. Therefore, the proposed technique, Freeman chain codes with characters' features (FCCwF) is used to increase the recognition accuracy rate by extending the recognition process using the features extracted in previous phase.

In general, as the conclusion of this research is, the input images need to be ensuring as high in quality or pre-processing stage must be performed so that the noise could be removed. Freeman chain code can be considered as an alternative recognition technique because of its ability in low recognition time which is also as important as recognition accuracy rate. To improve the recognition accuracy rate, the Freeman chain code can be applied with other technique and in this case, by using characters features such as total of pixel '1', accumulated sum of all pixels with value greater than '0' and the result of division between total of pixel '1' and accumulated sum of all pixels with value greater than '0'. By combining this technique with other technique, the recognition accuracy rate can gain high recognition rate of near 100 %.

References

1. Naito, T., Tsukada, T., Yamada, K., Kozuka, K.: Robust License-Plate Recognition Method For Passing Vehicles Under Outside Environment. IEEE Transactions on Vehicular Technology 49(6), 2309–2319 (2000)
2. Sarfraz, M., Ahmed, M.J., Ghazi, S.A.: Saudi Arabian License Plate Recognition System. In: Proceedings of the 2003 International Conference on Geometric Modeling and Graphics, pp. 36–41 (2003)
3. Wang, P., Zhang, W.: Research And Realization Of Improved Pattern Matching In License Plate Recognition. In: International Symposium on Intelligent Information Technology Application Workshops, pp. 1089–1092 (2008)
4. Tao, Q., He, X., Luo, D., Wu, W.: A New Car Plate Recognition Method Based On Fuzzy Entropy. In: Fifth World Congress on Intelligent Control and Automation, vol. 5, pp. 4054–4056 (2004)
5. Quan, J., Quan, S., Ying, S., Xue, Z.: A Fast License Plate Segmentation And Recognition Method Based On The Modified Template Matching. In: Proceedings of 2nd International Congress on Image and Signal Processing, pp. 1–6 (2009)
6. Tindal1, D.W.: Application Of Neural Network Techniques To Automatic Licence Plate Recognition. In: Conference of European Convention on Security and Detection, pp. 81–85 (1995)
7. Sirithinaphong, T., Chamnongthai, K.: The Recognition Of Car License Plate For Automatic Parking System. In: Fifth International Symposium on Signal Processing and its Applications, pp. 455–457 (1999)
8. Abdullah, S.N.H.S., Khalid, M., Yusof, R.: License Plate Recognition Using Multi-Cluster And Multilayer Neural Networks. In: 2nd International Conference on Information and Communication Technologies, pp. 1818–1823 (2006)
9. Vázquez, N., Nakano, M., Pérez-Meana, H.: Automatic System For Localization And Recognition Of Vehicle Plate Numbers. J. Applied Research and Technology 1(1), 63–77 (2003)
10. Zidouri, A., Deriche, M.: Recognition Of Arabic License Plates Using NN. In: First Workshops on Image Processing Theory, Tools and Applications, pp. 1–4 (2008)
11. Feng, J., Li, Y., Chen, L.M.: The Research of Vehicle License Plate Character Recognition Method Based on Artificial Neural Network. In: Proceedings of 2nd International Asia Conference on Informatics in Control, Automation and Robotics, pp. 317–320 (2010)
12. Paras Ram, A.S.: Design Of A Recognition System For Special Malaysian Car Plates Using Stroke Analysis. M.Eng. Thesis. University of Technology Malaysia, Malaysia (2005)
13. Shan, B.: License Plate Character Segmentation and Recognition Based On RBF Neural Network. In: Shan, B. (ed.) Second International Workshop on Education Technology and Computer Science (ETCS), vol. 2, pp. 86–89 (2010)
14. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice-Hall, Inc., Upper Saddle River (2002)
15. Yang, H., Xu, L., Shi, L.: Design and Implementation Of License Plate Recognition System. In: Design and Implementation Of License Plate Recognition System, pp. 602–605 (2007)
16. Fan, X., Fan, G.: Graphical Models For Joint Segmentation And Recognition Of License Plate Characters. IEEE Signal Processing Letters 16(1), 10–13 (2009)

17. Ma, X., Pan, R., Wang, L.: License Plate Character Recognition Based On Gaussian-Hermite Moments. In: Proceeding of 2010 Second International Workshop on Education Technology and Computer Science, pp. 11–14 (2010)
18. Chang, S.L., Chen, L.S., Chung, Y.C., Chen, S.W.: Automatic License Plate Recognition. IEEE Transactions on Intelligent Transportation Systems 5(1), 42–53 (2004)
19. Pan, X., Ye, X., Zhang, S.: A Hybrid Method For Robust Car Plate Character Recognition. Engineering Applications of Artificial Intelligence 18(8), 963–972 (2005)
20. Juntanasub, R., Sureerattanan, N.: Car License Plate Recognition through Hausdorff Distance Technique. In: 17th IEEE International Conference on Tools with Artificial Intelligence, pp. 647–651 (2005)
21. Fukumi, M., Takeuchi, Y., Fukumoto, H., Mitsura, Y., Khalid, M.: Neural Network Based Threshold Determination for Malaysia License Plate Character Recognition. In: Proceedings of 9th International Conference on Mechatronics Technology, vol. 1, pp. 1–5 (2005)
22. Wang, P., Zhang, W.: Research And Realization Of Improved Pattern Matching In License Plate Recognition. In: International Symposium on Intelligent Information Technology Application Workshops (2008)
23. Kulkarni, P., Khatri, A., Banga, P., Shah, K.: Automatic Number Plate Recognition (ANPR) System For Indian Conditions. In: Proceeding of 19th International Conference on Radioelektronika, pp. 111–114 (2009)
24. Williams, P.G., Kirby, H.R., Montgomery, F.O., Boyle, R.D.: Evaluation Of Video-Recognition Equipment For Number-Plate Matching. In: Proceedings of Second International Conference on Road Traffic Monitoring, pp. 89–93 (1989)
25. Parisi, R., Claudio, E.D.D., Lucarelli, G., Orlandi, G.: Car Plate Recognition By Neural Networks And Image Processing. Proceeding of IEEE International Symposium on Circuits and Systems 3, 195–198 (1998)
26. Al Faqheri, W., Mashohor, S.: A Real-Time Malaysian Automatic License Plate Recognition (M-ALPR) Using Hybrid Fuzzy. International Journal of Computer Science and Network Security 9(2), 333–340 (2009)
27. Nijhuis, J.A.G., Brugge, M.H.T., Helmholt, K.A., Pluim, J.P.W., Spaanenburg, L., Venema, R.S., Westenberg, M.A.: Car License Plate Recognition With Neural Networks And Fuzzy Logic. Proceeding of IEEE International Conference of Neural Networks 5, 2232–2236 (1995)
28. Freeman, H.: Computer Processing of Line-Drawing Images. ACM Computing Surveys 6(1), 57–97 (1974)
29. Madhvanath, S., Kim, G., Govindaraju, V.: Chaincode Contour Processing for Handwritten Word Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(9) (1999)
30. Cha, S., Shin, Y., Srihari, S.N.: Approximate Stroke Sequence String Matching Algorithm For Character Recognition And Analysis. In: Proceedings of the Fifth International Conference on Document Analysis and Recognition, pp. 53–56 (1999)
31. Jahne, B.: Digital Image Processing, 6th edn. Springer, New York (2005)
32. Seul, et al: Practical Algorithms for Image Analysis: Description, Examples and Code. Cambridge University Press, USA (1999)
33. McAndrew, A.: Introduction to Digital Image Processing With Matlab. Thomson Course Technology, USA (2004)
34. Jusoh, N.: Application of Freeman Chain Codes: An Alternative Recognition Technique for Malaysian Car Plates. International Journal of Computer Science and Network Security 9(11), 222–227 (2009)

Improving the Robustness of ISB Watermarking Techniques by Repetition of the Embedding

Akram M. Zeki[1], Azizah A. Manaf[2], and Shayma'a S. Mahmod[3]

[1] Department of Information System,
Kulliyyah of Information & Communication Technology,
International Islamic University Malaysia, Malaysia
akramzeki@iiu.edu.my
[2] Advanced Informatics School (AIS),
University Technology Malaysia, Malaysia
azizah07@ic.utm.my
[3] Department of Electrical and Computer Engineering,
Kulliyyah of Engineering,
International Islamic University Malaysia, Malaysia
shay_sinan@yahoo.co.uk

Abstract. Digital watermarking is a direct embedding of additional information into the original content or host image, this study is overcome the problems existing in the classic LSB method by adapting the method to intermediate significant bits (ISB), which improve the robustness and maintain the quality of the image. Enhancing the proposed method has been done by repeating the watermark data certain number of times (3, 5, 7, and 9 times) in order to improve the robustness of the watermarking technique, correspondingly, a majority criterion is used in the watermark detecting procedure, which makes the algorithm more robust, especially to the geometric transform attacks.

Keywords: Watermarking, Robustness, ISB, LSB.

1 Introduction

In general, watermark can be embedded in spatial domain or in transform domain of an image. The spatial domain is a domain in which an image is represented by the intensities at given points in space. This is the most common representation for image data. In the spatial domain approach, the pixel value of an image is modified to embed the watermark information [1]. Many studies [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] have used these spatial domain techniques.

In the transform domain approach, some sorts of transform are applied to the original image first. The watermark is embedded by modifying the transform domain coefficients. The applied transform may be Discrete Cosine Transform (DCT) [14] [15] [16], Discrete Fourier Transform (DFT) [17], or Discrete Wavelet Transform (DWT) [18].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 592–599, 2011.

2 Bit-Plane Model

A bit-plane of digital images is a set of bits having the same position in the respective binary numbers. To penetrate an image, the grey-scale of each pixel is decomposed into its 8 different bits; the first bit-plane contains the set of the most significant bits and the 8th bit-plane contains the least significant bits. This simple LSB embedding approach is easy for computation, and a large amount of data can be embedded without great quality loss. The more LSBs are used for embedding, the more distorted result will be produced. Not all pixels in an image can tolerate equal amounts of changes without causing notice to an observer. The largest number of the LSBs, whose grey values can be changed without producing a perceptible artifact in each pixel, is different.

The next step after selecting one bit-plane for embedding is to find the ranges of the chosen bit-plane, the length of the range L is 2k-1 (L = the maximum value of each range – the minimum value of the range + 1) and the number of ranges in each bit-plane is 256 / L. It can be noticed that in each range, the bit changes between 0 and 1, as shown in Figure 3 below.

3 Bit-Plane Model

In this paper, the ISB method founded by (Akram and Azizah, 2009) will be implemented and repeated 3, 5, 7 and 9 times in order to improve the robustness. In the proposed technique a tradeoff between the image quality and robustness has to be reached and the robustness of the system will be improvement by repeating the embedding. Figure 1 show the structure of the proposed scheme.

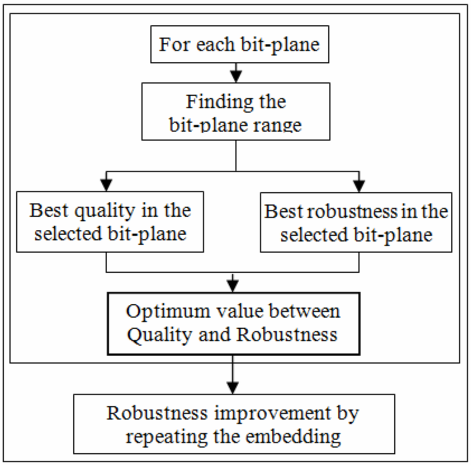

For each bit-plane
Finding the bit-plane range
Best quality in the selected bit-plane
Best robustness in the selected bit-plane
Optimum value between Quality and Robustness
Robustness improvement by repeating the embedding

Fig. 1. A flow chart of the proposed approach based on bit-plane model

3.1 Embedding Model

The embedding model is based on the ISB method [19], this method was found the best embedding status that can survive against different types of attacks and at the same time keeping minimum image distortion (the threshold value) where the best pixel value in between the middle and the edge of the range of bit-plane model, assume that the bias value is at least the distance from the position of the watermarked pixel to the edge of the range (which is more close to the original pixel). That means if the distance from the pixel to the edge of the range is greater than the bias value, then the position of the pixel will not change. While if the distance from the pixel to the edge of the range is smaller than the bias value, then the position of the pixel will change to be as far as the bias value.

The best robustness can be obtained when the bias value is maximum, (in the middle of ranges) while the worst one when the bias value is minimum (in the edges of ranges). Regarding the quality, the best image quality when the bias value is minimum, while the worst one when the bias value is maximum. And the best embedding status was addressed when the bias value is 6 [19].

3.2 Repetition of the Embedding

In this section, robustness is improved by repeating the embedded bits. The first step of the watermark generation is to repeat each bit of the hidden information for a certain number of times. Correspondingly, a majority criterion is used in the watermark detecting procedure, which makes the algorithm more robust, especially to the geometric transform attacks and noise attack [20].

The watermark is then encoded by repeating the original signal R times, in a block section, known s the block section (R, 1). For example, in case three repetitions are done (R=3), the image is partitioned into blocks with the size of 3 pixels, as shown in Figure 2 below.

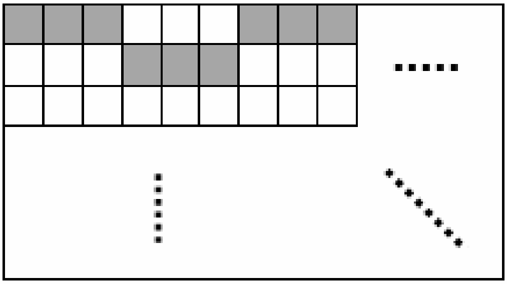
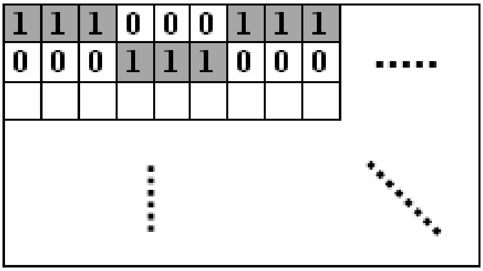


Fig. 2. Partitioned the image into blocks in the size of 3 pixels to repeat the embedding within each block

During the embedding, the watermarked object is considered as a long sequence of bits; if the watermark bits contain the following bits: (101010), every bit is embedded 3 times, as shown in Figure 3 below.



1	1	1	0	0	0	1	1	1
0	0	0	1	1	1	0	0	0

Fig. 3. Repeating the embedding within the blocks

In the decoding process, the majority elements of the block section are used to reconstruct the original signal. For example, R=3 in the binary signal and the (000) represents 0, the (111) represents 1; while (010, 100 or 001) represent 0 and (101, 110 or 011) represents 1, i.e. the reconstructed signal becomes ‘0’ if the number of ‘0’s is 2 or more in the block section; otherwise, it is ‘1’ [22].

In other words, if the value of the extracted bit is repeated (1 + number of repetition / 2) times or more, (i.e. R` ≥ 1 + R/2) the value is then selected as an extracted bit, i.e. if "1" is extracted in most of the pixels in the block, the bit with value "1" is considered for reconstructing the watermark. If "0" is extracted in most of the pixels in the block, the bit with value "0" is considered for reconstructing the watermark. For 5 repetitions, if the value of the bit is redundant (R`) 3 times or more, the value is selected as an extracted bit. As for 7 repetitions, if the value of the bit is redundant (R`) 4 times or more, the value is selected as an extracted bit. For 9 repetitions, if the value of the bit is redundant (R`) 5 times or more, the value is selected as an extracted bit [23].

As for a higher number of repetitions, it can give better robustness because the maximum number of repeating watermarks embedding R is shown in Equation 1.

$$R = floor\ R' \tag{1}$$

Where the floor is a function whose value is the largest integer less than or equal to R', and R' is the size of host image / the size of watermark object. If the number of repetition increases, the capacity of embedding is decreased, as shown in Equations 2 and 3.

$$New\ Capacity\ (C`) = Old\ Capacity\ (C) \,/\, Size\ of\ the\ block \tag{2}$$

$$New\ Capacity = \frac{\text{Total numberof bytes of data hiding}}{\text{Total numberof bytes of coverimage} \times \text{Size of the block}} \tag{3}$$

3.3 Implementation and Experimental Results

In this study, grey scale image (logo) contains 90 × 90 pixels as shown in Figure 4 will be embedded within three host images containing 256 × 256 pixels as shown in Figure 5. To improve the security of the system, the watermark object is encrypted using Random Pixel Manipulation Technique [21]. In this technique, a key is chosen.

This key is a string which can be effectively manipulated to obtain a random number sequence. This sequence is then used to 'scramble' the hidden data.

The repetition of the embedding will be done 3, 5, 7 and 9 times in the 4th bit-plane of the host images with the bias value = 6. The idea of repeating the embedding was to increase the robustness of the watermark system against all types of attacks, specifically against the geometric transform attacks.

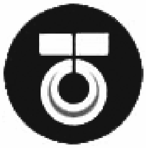
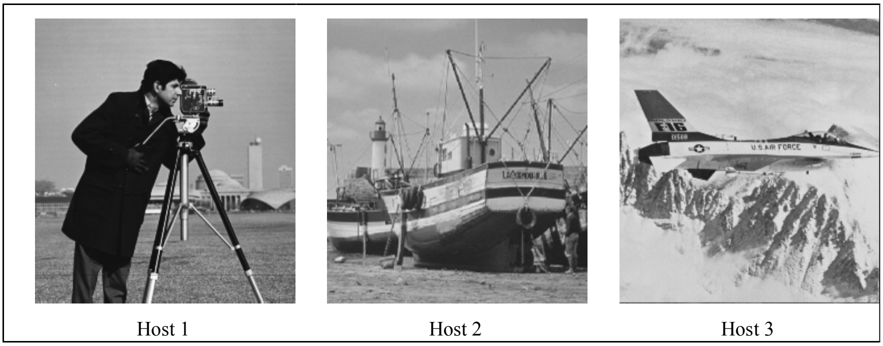


Fig. 4. Grey scale logo with 90 × 90 pixels.



Host 1 Host 2 Host 3

Fig. 5. Grey scale host image with 256 × 256 pixels.

To study the proposed method, under different image processing operations (Attacks), the following attacks will be applied to the image: Lossy compression with 85% compression level, Blurring, Gaussian filter, Wiener filter, Speckle noise, and geometric transform attacks (Rotation and Scaling).

During the extracting stage the encrypted logos have been extracted and they are decrypting to the original images by the same key has been used during encryption stage.

The watermark image was extracted from Hosts 1 to 3, and the normalized cross correlation (NCC) was measured for every embedding in different sizes of blocks, as shown in Tables 1 - 3, respectively.

Table 1. The NCC of the 4th Bit Plane at Bias value =6 for host 1

Repeating	JPEG	Blurring	Gaussian	Wiener	Speckle	Rotation	Scaling
1	0.931264	0.910986	0.919141	0.847613	0.932271	0.799517	0.809247
3	0.951085	0.968978	0.935721	0.864331	0.977016	0.852578	0.863577
5	0.97888	0.978954	0.963583	0.891514	0.989924	0.888657	0.894251
7	0.985841	0.986548	0.978995	0.94775	0.993589	0.931555	0.948953
9	0.998977	0.99039	0.981728	0.999882	0.995276	0.998842	0.993604

Table 2. The NCC of the 4th Bit Plane at Bias value =6 for host 2

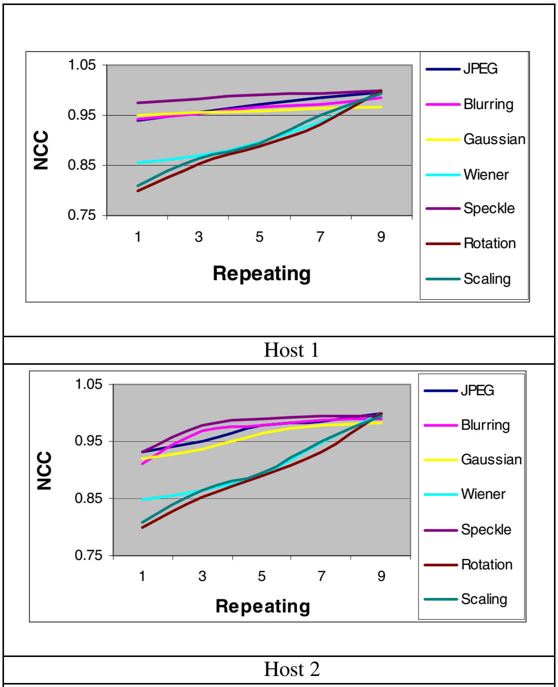
Repeating	JPEG	Blurring	Gaussian	Wiener	Speckle	Rotation	Scaling
1	0.917325	0.899574	0.911418	0.871249	0.886982	0.799517	0.809229
3	0.947652	0.946014	0.951353	0.890363	0.936686	0.852578	0.863583
5	0.978366	0.964745	0.978446	0.911319	0.96874	0.888657	0.894242
7	0.983007	0.978428	0.98172	0.954946	0.983741	0.931555	0.948957
9	0.990667	0.98278	0.985976	0.997542	0.990698	0.998842	0.993601

Table 3. The NCC of the 4th Bit Plane at Bias value =6 for host 3.

Repeating	JPEG	Blurring	Gaussian	Wiener	Speckle	Rotation	Scaling
1	0.90519	0.863423	0.873039	0.855802	0.817691	0.799517	0.8093
3	0.942476	0.905038	0.910462	0.886317	0.871951	0.852578	0.863536
5	0.967077	0.934157	0.940517	0.898179	0.893755	0.888657	0.894255
7	0.984243	0.954411	0.95745	0.939749	0.929244	0.931555	0.948957
9	0.993528	0.959585	0.963423	0.999382	0.948403	0.998842	0.993604

The above tables show that the value of the NCC increases with the increasing of embedding times for all the attacks. The result of NCC was found to be very close to 1, after nine repetitions.

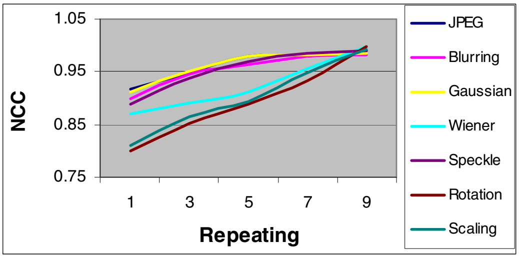
Figure 6 below presents the results gathered for the NCC of the different attacks, after (3, 5, 7, and 9) times repeating of the embedding.



Host 1

Host 2

Fig. 6. The NCC value for the proposed method



Host 3

Fig. 6. (*continued*)

The above figures show different graphs for the three host images which illustrate the relation between the number of repeating and the NCC values for all the attacks. At the same time, these figures also show that the value of the NCC increased when the embedding number for all the attacks was increased, including the geometric transform attacks (Rotation and Scaling), which were not improved when the method based on only one pixel was used. The result of the NCC for all the attacks was found to be very close to 1 after nine reparations.

4 Conclusion

Digital watermarking technique based on intermediate significant bit (ISB) has been implemented by this paper. This technique can survive against different types of attacks and at the same time keep the quality of the image. The embedding data has been repeated certain number of times. The idea of repeating the embedding was to increase the robustness of the watermark system against all types of attacks, specifically against the geometric transform attacks. In the present study, the capacity of embedding was found to be decreased by increasing the number of repeating times. The repetition was done 3, 5, 7 and 9 times, by embedding the watermark image, in all host images within the 4th bit-plane with the bias value = 6. The results show that the value of the NCC increases with the increase of embedding times for all attacks. The result of NCC was found to be very close to 1, after nine repetitions.

References

1. Chen, P.C.: On the Study of Watermarking Application in WWW – Modelling, Performance Analysis, and Applications of Digital Image Watermarking Systems. Ph.D. Thesis, Monash University (1999)
2. Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. Pattern Recognition, 469–474 (March 2004)
3. Chang, C.C., Hsiao, J.Y., Chan, C.S.: Finding optimal least-significant-bit substitution in image hiding by dynamic programming strategy. Pattern Recognition 36(7), 1583–1595 (2003)

4. Chang, C.C., Tseng, H.W.: A Steganographic method for digital images using side match. Pattern Recognition Letters 25, 1431–1437 (2004)
5. Lin, C.C., Tsai, W.H.: Secret image sharing with steganography and authentication. Journal of Systems and Software 73, 405–414 (2004)
6. Lou, D.C., Liu, J.L.: Steganographic method for secure communications. Computers and Security 21, 449–460 (2002)
7. Marvel, L.M., Boncelet, C.G., Retter, C.T.: Spread spectrum image steganography. IEEE Transactions on Image Processing 8, 1075–1083 (1999)
8. Thien, C.C., Lin, J.C.: A simple and high-hiding capacity method for hiding digit-by-digit data in images based on modulus function. Pattern Recognition 36(12), 2875–2881 (2003)
9. Wang, R.Z., Lin, C.F., Lin, J.C.: Image hiding by optimal lsb substitution and genetic algorithm. Pattern Recognition 34(3), 671–683 (2001)
10. Wu, D.C., Tsai, W.H.: A steganographic method for images by pixel-value differencing. Pattern Recognition Letters 24(9-10), 1613–1626 (2003)
11. Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pixel-value differencing and LSB replacement methods. IEE Proceedings of Visual Image Signal Process 152, 611–615 (2005)
12. Yu, Y.H., Chang, C.C., Hu, Y.C.: Hiding secret data in images via predictive coding. Pattern Recognition 38, 691–705 (2005)
13. Zhang, X., Wang, S.: Steganography using multiple-base notational system and human vision sensitivity. IEEE Signal Processing Letters 12, 67–70 (2005)
14. Barni, M., Bartolini, F., Cappellini, V.: A DCT-domain System for Robust Image Watermarking. Signal Processing (Special Issue on Watermarking) 66(3), 357–372 (1998)
15. Hsu, C.T., Wu, J.L.: DCT-Based Watermarking for Video. IEEE Transactions on Consumer Electronics 44(1), 206–215 (1998)
16. Langelaar, G.C., Lagendijk, R.L.: Optimal Differential Energy Watermarking of DCT Encoded Images and Video. IEEE Transactions on Image Processing 10(1), 148–158 (2001)
17. O'Ruanaidh, J., Pun, T.: Rotation, Scale and Translation Invariant Digital Image Watermarking. In: Proceedings of IEEE Int. Conf. Image Processing., vol. 1, pp. 536–538 (1997)
18. Voloshynovskiy, S., Deguillaume, F., Pereira, S., Pun, T.: Optimal Adaptive Diversity Watermarking with Channel State Estimation. In: Proceedings of SPIE: Security and Watermarking of Multimedia Contents III, vol. 4314(74) (2001)
19. Zeki, A.M., Manaf, A.A.: A Novel Digital Watermarking Technique Based on ISB (Intermediate Significant Bit). In: Akram, M. (ed.) ICACEM 2009 - International Conference on Applied Computing and Engineering Mathematics. WCSET 2009. World Congress on Science, Engineering and Technology, Penang. Malaysia, February 25-27 (2009)
20. Niu, X.: A Survey of Digital Vector Map Watermarking. International Journal of Innovative Computing, Information and Control 2(6), 1301–1316 (2006)
21. Venkatraman, S., Abraham, A., Paprzycki, M.: Significance of Steganography on Data Security. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004), IEEE Computer Society Press, Los Alamitos (2004)
22. Hsieh, C.T., Lu, Y.L., Luo, C.P., Kuo, F.J.: A Study of Enhancing the Robustness of Watermark. In: ISMSE Conference, pp. 325–327 (2000)
23. Ohbuchi, R., Ueda, H., Endoh, S.: Robust Watermarking of Vector Digital Maps. In: Proceedings of the IEEE International Conference on Multimedia and Expo 2002 (ICME 2002), Lausanne, Switzerland, August 26-29 (2002)

Geometrically Invariant Watermarking Based on Local Generalized Orthogonalized Complex Moments

Hai Tao[1], Jasni Mohamad Zain[1], Mohammad Masroor Ahmed[1], and Ahmed Abdalla[2]

[1] Faculty of Computer Systems and Software Eng.,
University Malaysia Pahang, Malaysia
[2] Faculty of Electrical and Electronic Engineering,
University Malaysia Pahang, Malaysia
taotao27@gmail.com, jasni@ump.edu.my,
masroorahmed@gmail.com, waal85@yahoo.com

Abstract. This paper proposes a novel geometrically invariant watermarking scheme based on Hessian-Laplace detector and Bessel–Fourier moments. Firstly, Hessian-Laplace detector is adopted to extract feature points. Then, non-overlapped disks centered at feature points are generated. These disks are normalized for the invariance to rotation, scaling and translation distortions. Finally, the watermark is embedded in magnitudes of Bessel–Fourier moments of each disk via dither modulation to realize the robustness to common image processing operations and de-synchronization attacks. Simulation results demonstrate the proposed watermarking procedure has the superior and remarkable performance in imperceptibility and robustness to various attacks.

Keywords: Hessian-Laplace, Bessel–Fourier moments, Invariant watermarking.

1 Introduction

Digital watermarking [1] is the process of embedding watermark into a multimedia product. The embedded data can later be extracted or detected from the watermarked product, for protecting digital content copyright and ensuring tamper-resistance. Nowadays, there is an unprecedented development in the image watermarking field. Simultaneously, attacks against image watermarking systems have become more sophisticated [2]. In general, these attacks can be categorized into common image processing operations such as median filtering, sharpening, noise adding, and JPEG compression etc, and de-synchronization attacks such as rotation, scaling, translation, random bend attack, and cropping etc. While the common image processing operations reduce watermark energy, de-synchronization attacks induce synchronization errors between the original and the extracted watermark during the detection process. Most of the previous watermarking schemes are robust to common image processing operations, but show severe problems to de-synchronization attacks.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 600–611, 2011.

Fortunately, several approaches [2-4] against the de-synchronization attacks have been developed in recent years. In [3], this paper puts forward invariant transforms apply to of a geometrically attacked watermarking approach. The Fourier–Mellin transform can be used to produce an RST invariant image watermarking algorithm for the first time. The spread spectrum is used to implement a watermark generation and the watermark insertion and extraction are implemented in a domain that is invariant to geometric attacks. However, it comes up against the implementation difficulties that are the obstruction and hindrance of the further research in this area. In [4], it proposes a robust image watermarking scheme based on Zernike moments. The proposed approach computes locally Zernike moments and modifies them to embed watermarks, achieving resistant to local geometric attacks. Moreover, the salient region parameters, which consist of an invariant centroid and a salient scale, and transmit them to the decoder, are extracted to deal with scaling attacks. Although there are techniques that efficiently deal with local geometric attacks, resisting both signal processing attacks and geometric attacks still remain to be a challenging problem to experts.

In this paper, we first propose a geometrically invariant digital image watermarking technique to construct watermark synchronization using local Bessel–Fourier Moments and Hessian-Laplace feature detectors. Some important improvement properties of local structure for the detection of salient feature points are represented. The normalized square regions centered at selected feature detectors are avoidable traditional many redundant points generation and correlated complex computation, meanwhile, survive general signal processing operations and geometric transformations. Subsequently, the watermark is embedded individually into each normalized square patch using the invariance of Bessel–Fourier moments. The comparison of the presented approach with previous algorithms in terms of robustness and imperceptibility is also provided. The experimental results show that the proposed scheme has the remarkable performance in resistant to common image processing operations survival the de-synchronization attacks.

2 Preliminaries

Bessel–Fourier Moments and Hessian-Laplace detector play important roles in achieving the proposed watermark synchronization technique to deal with geometric attacks or distortions. In this section, they will be introduced in detail.

2.1 Bessel–Fourier Moments

In [5], Bin X. *et.al.* introduced a new set of Bessel–Fourier moments which are the generalized orthogonalized complex moments based on the Bessel function of the first kind, for images analysis and reconstruction purposes, by using a set of complex polynomials $\mathrm{B_{mn}}$ (x, y), which form a complete orthogonal set over the interior of the unit disk $\mathrm{x^2 + y^2 = 1}$. These polynomials in polar coordinates have the form

$$\mathrm{B_{nm}} = \frac{1}{2\pi a_n}\int_0^{2\pi}\int_0^1 f(r,\theta)\, J_v(\rho_n r)\exp(-jm\theta)\, r dr d\theta \tag{1}$$

where $j = \sqrt{-1}$ $f(r, y)$is the image defined in the polar coordinate system, and n (the order of the Bessel radial transform) is a non-negative integer, $m = 0, +1, +2, ...$(the circular harmonic order), r is the length of the vector from the origin $(\bar{x}, \bar{y})$ to the pixel (x, y) and θ is the angle between vector r and x axis in counter-clockwise direction. $a_n = [J_{v+1}(\rho_n)]^2/2$ is the normalization constant. And $J_v(\rho_n r)$ is the Bessel function of the first kind that the definition is as follows

$$J_v(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(v+k+1)}\left(\frac{x}{2}\right)^{v+2k} = \frac{\left(\frac{x}{2}\right)^v}{\Gamma(v+1)} F_1\left(v+1, -\left(\frac{x}{2}\right)^2\right) \tag{2}$$

where v is a real constant, $\Gamma(\bullet)$ is the gamma function, F_1 is the generalized hypergeometric function. The Bessel function is the solution of the Bessel's equation

$$\frac{d^2y}{dx^2} + \frac{1}{x}\frac{dy}{dx} + \left(1 - \frac{v^2}{x^2}\right)y = 0 \tag{3}$$

The polynomials of (1) are orthogonal and satisfy the orthognality principle

$$\int_0^1 r\, J_v(\rho_n r) J_v(\rho_k r) dr = a_n \delta_{nk} \tag{4}$$

where δ_{nk} is the Kronecker delta and $\delta_{\alpha\beta} = 1$ for $\alpha = \beta$ and $\delta_{\alpha\beta} = 1$ otherwise, $r \in [0,1]$ is the size of the objects that can be encountered in a particular application. Hence the basis functions $J_v(\rho_n r)\exp(-jm\theta)$ of the Bessel–Fourier moments are orthogonal over the interior of the unit circle.

$$\int_0^{2\pi} \int_0^1 [J_v(\rho_n r)\exp(-jp\theta)]^* J_v(\rho_m r)\exp(-jq\theta)\, r dr d\theta = 2\pi a_n \delta_{nm} \delta_{pq} \tag{5}$$

Suppose that one knows all moments B_{nm} of $f(x, y)$ up to a given order n_{max}. It is desired to reconstruct a discrete function $\bar{f}(r, \theta)$, whose moments exactly match those of f (x, y) up to the given order n_{max}. The orthogonal Bessel–Fourier moments are the coefficients of the image expansion into the orthogonal polynomials (1)over the unit disk, as it can be seen in the following reconstruction equation for solving Eq.(4)

$$\bar{f}(r, \theta) = \sum_{n=1}^{n_{max}} \sum_{m}^{m_{max}} B_{nm} J_v(\rho_n r)\exp(-jm\theta) \tag{6}$$

Note that as n_{max} approaches infinity and $\bar{f}(r, \theta)$ will approach $f(x, y)$. The basis function $J_v(\rho_n r)\exp(-jm\theta)$ of Bessel–Fourier moments defined in Eq.(1) can be expressed as complex polynomials in $(x + jy)$ and $(x - jy)$:

$$J_v(\rho_n r)\exp(-jm\theta) = \sum_k b_{n,k}\, r^{v+2k}\exp(-jm\theta) = \sum_k b_{n,k}(x + jy)^p(x - jy)^q \tag{7}$$

where $b_{n,k} = \frac{(-1)^k}{k!\Gamma(v+k+1)}\left(\frac{\rho_n}{2}\right)^{v+2k}$, $p = \frac{v+2k-m}{2}$, $q = \frac{v+2k+m}{2}$.

The Bessel–Fourier moments can be expressed as linear combination of complex moments:

$$B_{nm} = \frac{1}{2\pi a_n} \sum_k b_{n,k}\, C_{pq} \tag{8}$$

where the complex moments are defined by

$$C_{pq} = \iint_{-\infty}^{+\infty} f(x, y)(x + jy)^p(x - jy)^q dxdy \tag{9}$$

The orders p and q are nonnegative integers in Eq.(9).In Eq. (8)the complex moments are modified so that the orders p and q are real valued and can be negative, the Bessel–Fourier moments are therefore the orthogonal complex moments. When both $p + q = v + 2k$ and $p - q = m$ are integers and $v + 2k = p + q \geq 0$, the modified complex moments are convergent and integrable. The Bessel–Fourier moments therefore may be computed in the Cartesian coordinate system according to Eqs. (8) and (9).

Bessel–Fourier moments have been shown to be superior to others in the term of feature representation capabilities and robustness. The degree n of $J_v(\rho_n r)$ in the Bessel–Fourier moments depend on representing an image can be lower than that of an expression using Zernike moments and orthogonal Fourier–Mellin moments. The lower degree, results in insensitive to variation and the presence of noise. This gives an advantage to Bessel–Fourier moments over others moments in the case that it is able to achieve has minimal information redundancy and better noise robustness.

2.2 Hessian-Laplace Detector

It has been verified that feature points extracted by Hessian-Laplace regions are invariant to rotation, scale translation and projective transformations. The Hessian-Laplace detector is based on the local autocorrelation matrix of an image; where the local autocorrelation matrix measures the local changes of the image with patches shifted by a small amount in different directions. Detected points are localized in space and scale, at the local maxima of the Hessian determinant and the local maxima of the Laplacian-of-Gaussian, respectively. It is noticed that Harris-Laplace detectors localize points at local scale-space maxima of the difference-of-Gaussian in the construction of local invariant regions. However, Hessian-Laplace acquires a more accurate scale localization and feature selection in scale-space. Compared with scale-invariant feature transform (SIFT), this method is more efficient in the term of robustness to scaling, shearing and rotation attacks. To obtain the invariant scale transformations and accurate detectors, a set of images are calculated by reliable Hessian-Laplace detector at different resolution levels. Then, an approach to automatic scale selection is applied to feature points decision.

Let $g(X, \sigma_D)$ denote a 2D Gaussian kernel with mean zero and standard deviation σ_D. Gaussian smoothing function $g(X, \sigma_D)$ is defined as

$$g(X, \sigma_D) = \frac{1}{2\pi\sigma_D} e^{-\frac{\|X\|^2}{2\sigma_D}} \tag{10}$$

Let f(x) denote the original image brightness at a spatial vector X. The uniform Gaussian scale-space representation $L(X, \sigma_D)$ is defined by convolution operator over $X \in R^2$. $L(X, \sigma_D)$ is defined as

$$L(X, \sigma_D) = g(X, \sigma_D) \otimes f(X) \tag{11}$$

The directional derivative I of a multivariate differentiable function along a given vector X at a given point $(x \text{ or } y)$ represents intuitively the instantaneous rate of change of the function. For Hessian-Laplace detector, a Taylor expansion truncated to the scale-normalized second order moment matrix is defined as

$$H(X) = \sigma_D^2 g(\sigma_1) \otimes \begin{bmatrix} \frac{\partial^2 I}{\partial x^2}(X, \sigma_D) & \frac{\partial^2 I}{\partial x \, \partial y}(X, \sigma_D) \\ \frac{\partial^2 I}{\partial x \, \partial y}(X, \sigma_D) & \frac{\partial^2 I}{\partial y^2}(X, \sigma_D) \end{bmatrix} \tag{12}$$

The matrix H represents the distributions of the second directional derivatives in a local neighborhood of X . The feature detection procedure is similar to the performance of a Laplacian operator. However, the second derivative is very small in one particular orientation if the function penalizes long structures based on the determinant of the Hessian matrix. The local derivatives are computed by smoothing Gaussian window with scale σ_D. Then, the derivatives averaging is performed over the area of the image patch in the neighborhood of X with a Gaussian window of integration scale σ_1. The matrix eigenvalues represent two principal signal changes in the neighborhood of X.

This property makes extracting corner or junction points possible, at which curvatures are significant in orthogonal directions. To decrease the complexity of explicit eigen problem, the definitions of matrix determinant and trace are introduced. The strength measure matrix C of Harris corner can be computed as

$$C(X, \sigma_1, \sigma_D) = \det(H) - \tau \cdot \mathrm{trace}^2(H) \tag{13}$$

where τ is a predefined constant. Feature points are detected by the scale space maxima of the strength measure.

The general idea for automatic scale selection is proposed is study the evolution properties over scales of normalized differential detectors. Specially, local extrema over scales are likely to correspond to interesting image structures.

The maximum uniformity between the local image structures and the feature detection function is reflection of the characteristic scales. Laplacian-of- Gaussians (LoG) is used for locating the characteristic scale. It is defined as

$$|\mathrm{LoG}(X, \sigma_1)| = \sigma_1^2 \left| \frac{\partial^2 I}{\partial x^2}(X, \sigma_1) + \frac{\partial^2 I}{\partial y^2}(X, \sigma_1) \right| \tag{14}$$

where $\frac{\partial^2 I}{\partial x^2}$ and $\frac{\partial^2 I}{\partial y^2}$ are second order partial derivatives with respect to the directions of x and y, respectively. For each candidate point, an iterative operator is applied to detect the scale and the location of feature points. The extreme over scale of the LoG are used to select the scale of feature points.

2.3 Local Feature Region Construction

Local feature regions (LFR) are generated by grouping the feature points into geometrically invariant subsets of the host image which reflect the important semantics of the image. It is significant for the appropriate selection of scale-space feature points to solve the problem of geometrical synchronization. Some strong feature points are selected to be centers of the circles since the weak feature points will easily disappear when image contents are modified. In the feature-point set of the image, the relative positions of all the pixels to the centers are allowed to undergo geometric deformations. The distribution of feature points also is an important index to affect the performance of watermark synchronization. If the distances are so small between adjacent features points that the patches overlap serious in large areas and if only a small number of patches are available thank to the broad distances watermark strength will decrease. The radius of the LFR is chosen in accordance with the repeatability. For constructing the disks to embedding and extracting watermark, feature points as the centre and feature scale are employed and the radius of these disks is defined

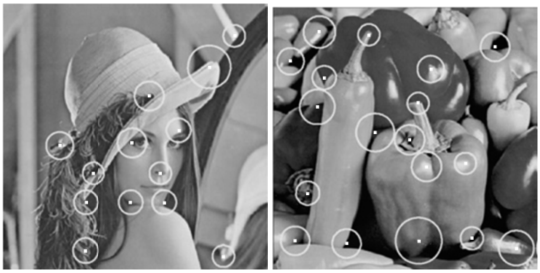
$$R = \tau[\sigma] \tag{15}$$

where [•] is rounding operator, σ is the characteristic scale and τ is a magnification factor for controlling the size of the circular characteristic region. Generally speaking, a small feature scale of the selected points means a low repeatability, and a large feature scale results in serious overlapping among them. As a result, the scale of feature points is limited 5 to 9. The value of τ is increased while the capacity of the embedded watermark will be improved. However, the robustness of watermarking scheme would decrease. Hence, there is a compromise between capacity and robustness. The range of R is limited by

$$[\sigma] \le R \le \frac{\min(M,N)}{2} \tag{16}$$

where M and N are the height and width of the original image, respectively.

Once the disks are overlapped with each other, the disk with a large number of feature points is selected for avoiding the interference between any two disks and enhancing imperceptible watermark in highly texture domain. In Fig 1, it is shown that the final disks are selected by Hessian-Laplace detectors.



(a) (b)

Fig. 1. The final disks selection by Hessian-Laplace detectors (a) Lena and (b) Peppers

3 The Proposed Watermarking Scheme

In this section, to improve the robustness of transmitted information, the same copy of the chosen watermark should be embedded in all channels. Based on the above ideas, a novel feature-based image watermarking algorithm resist to de-synchronization attacks is presented, in which the feature points detection and generalized orthogonalized complex moments theory is utilized. A random sequence with the size

L (digital watermark) $W = \{w_i, i = 1, \ldots, L\}$ is generated by the secret key K_1. And the popular images "Lena" and "Peppers" are tested with the size 512×512.

3.1 Watermark Embedding

For watermark embedding, LFRs according to the process mentioned in Section 2 are selected, wherein the selected regions are suitable for watermark insertion. Second, these regions are transformed into circular patches by normalization technique. Finally, the watermark pattern is embedded repeatedly in all normalized square patches using the invariance of Bessel–Fourier moments. The detailed algorithm is given as follows:

Step 1: The steady image feature points, denoted as $\mathrm{P_j}(\mathrm{j} = 1, \ldots, \mathrm{n})$ are extracted by using Hessian-Laplace detector according to the presented selection criterion. The disks are generated with the centers at these feature points for the purpose of watermark insertion. Since the radius of each disk is in direct proportion to its corresponding characteristic scale, the disk can be covariant to the image content changes, such as scaling. And the local feature regions (LFRs), denoted as $\mathrm{O_k}(\mathrm{k} = 1, \ldots, \mathrm{m})$, are constructed adaptively in accordance with the feature scale-space theory.

Step 2: The image normalization technique, developed as an elegant recognition preprocessing algorithm, will be exploited in the proposed image watermarking, since image normalization obtained from a series of geometric transformations can produce a RST-invariant output, i.e. it transforms the distorted input into its corresponding normalized pattern form such that it is invariant to rotation, scaling and translation. To improve the robustness against local de-synchronization attacks and image processing distortions, image normalization is implemented on the local feature region for achieving the standard size because the target of image normalization is to project different deformed views of the same image regions to its canonical size and dominant orientation.

So in our scheme, the watermark sequence is embedded individually into each normalized disk to resist de-synchronization. Thus, each disk can be considered as an individual spatial domain for the watermark insertion. Under various affine transforms and signal processing distortions, watermark information can be retrieved correctly in only several disks. It is difficult to implement directly image normalization in the rounded patch centered at extracted feature points. For conquering the problem, the zero-padding operator is exploited for mapping The LFR into the block with size $2\mathrm{R} \times 2\mathrm{R}$. Consequently, the image normalization procedure performs profitably to each square region and the normalized LFR is obtained.

Step 3: Several low order Bessel–Fourier moments are calculated in the LFR. It is difficult to modify Bessel–Fourier moments of the LFR for carrying the random sequence due to severe implementation difficult. Moreover, in the reconstruction procedure, if the order of moments is higher than 3, the computation leads to heavy complication and time consuming. Hence, the invariant properties of Bessel–Fourier moments of input images have to be compromised to certain degrees allowing for geometric error of normalized disk transformations, interpolation error of rotation and scaling and approximation error of discrete Bessel–Fourier polynomial integration. As a result, the Bessel–Fourier moments are denoted as $\Omega = \{\mathrm{B}_{\mathrm{nm}}, \mathrm{m} + \mathrm{n} < \mathrm{T}, \mathrm{m}\,\mathrm{and}\,\mathrm{n} \neq 0\}$ whose order are less than or equal to T, and watermarks should be embedded into the intensity of the host image using Bessel–Fourier moments to trade off invisibility and robustness. Dither modulation, in which the watermark information modulates a dithered signal and the original image is quantized with an associated dithered quantization step, has considerable performance advantages over previous frameworks in terms of the feasibility for the modification of B_{mn} to implement watermark insertion. The binary watermark sequence is denoted as $W = \{w_i, i = 1, \ldots, L\}$, and $w_i \in \{0,1\}$. A secret key K_1 to randomly select L Bessel–Fourier moments form Ω is applied to a form Bessel–Fourier moments vector $B = (B_{p_1q_1}, \ldots B_{p_Lq_L})$. The magnitude of $B_{p_iq_i}$ is quantized for carrying a watermark bit w_i with length L. After the dither modulation, a new vector $\bar{B} = (\bar{B}_{p_1q_1}, \ldots \bar{B}_{p_Lq_L})$ is produced by

$$\left|\bar{B}_{p_iq_i}\right| = \left[\frac{B_{p_iq_i} - d_i(b_i)}{\Delta}\right]\Delta + d_i(w_i) \qquad i = 1, \ldots L \tag{17}$$

where $[\bullet]$ is rounding operation, Δ is quantization step and $d_i(\bullet)$ is the dither function for the ith quantization step function satisfying $d_i(1) = \frac{\Delta}{2} + d_i(0)$. The dither vector $(d_0(0), d_1(0), \ldots, d_L(0))$, which follows uniform distribution over $[0, \Delta]$, is generated by secret key K_2. Thus, the modified Bessel–Fourier moments can be expressed as

$$\bar{B}_{p_iq_i} = \frac{\left|\bar{B}_{p_iq_i}\right|}{\left|B_{p_iq_i}\right|} B_{p_iq_i} \qquad i = 1, \ldots L \tag{18}$$

Step 4: For each watermarked inscribed square patch, there are two parts. One part is constructed the patch without the selection of Bessel–Fourier moments, which is

$$\mathrm{f}_{\mathrm{unchanged}}(\mathrm{x}, \mathrm{y}) = \mathrm{f}(\mathrm{x}, \mathrm{y}) - \mathrm{f}_{\mathrm{T}}(\mathrm{x}, \mathrm{y}) \tag{19}$$

where $\mathrm{f}(\mathrm{x}, \mathrm{y})$ is denoted as the original patch and $\mathrm{f}_{\mathrm{T}}(\mathrm{x}, \mathrm{y})$ will be used to reconstruct patch for inserting watermark by the selected Bessel–Fourier moments.

The other is the patch $\mathrm{f}_{\bar{\mathrm{T}}}(\mathrm{x}, \mathrm{y})$, which is reconstructed by the quantizing Bessel–Fourier moments. Consequently, a watermarked inscribed square patch will be acquired by combining the two parts which is expressed by

$$\bar{\mathrm{f}}(\mathrm{x}, \mathrm{y}) = \mathrm{f}_{\bar{\mathrm{T}}}(\mathrm{x}, \mathrm{y}) + \mathrm{f}_{\mathrm{rest}}(\mathrm{x}, \mathrm{y}) \tag{20}$$

The watermarked image can be obtained by replacing of the original inscribed square patches with watermarked ones.

Step 5: After watermark insertion, the zero-removing operator is applied to the square patch in order to using invertible transformation of circle area. Note that the zero-padding and zero-removing procedures result in energy loss, but the effectiveness is so small that it affects hardly watermark extraction.

3.2 Watermark Extraction

The procedure for watermark detection is illustrated as follows,

Steps 1: These first three steps are similar to Steps 1-3 in the watermark embedding procedure.

Step 2: With the same secret key K_1, L relevant Bessel–Fourier moments will be selected for watermark extraction, which is denoted as $B' = ({B'}_{p_1q_1}, \ldots {B'}_{p_Lq_L})$. First, the same key K_2 is used to reproduce the same two dithered vector $(d_0(0), d_1(0), \ldots, d_L(0))$ and $(d_0(1), d_1(1), \ldots d_L(1))$. Then, according to the Eq.(17), the magnitude of each ${B'}_{p_iq_i}$ is quantized with the two corresponding dithered vectors, respectively,

$$\left|\overline{B'}_{p_iq_i}\right| = \left[\frac{{B'}_{p_iq_i} - d_i(j)}{\Delta}\right]\Delta + d_i(j) \qquad i = 1, \ldots L \qquad j = 0,1 \tag{21}$$

Finally, the distances between $\left|\overline{B'}_{p_iq_i}\right|$ are estimated with its two quantized versions by the minimum distance operator defined by (22), the watermark bit embedded will be obtained in $\left|B_{p_iq_i}\right|$

$$w_i' = \operatorname{argmin}_{j\in\{0,1\}}\left(\left|\overline{B'}_{p_iq_i}\right|_j - \left|\overline{B'}_{p_iq_i}\right|\right)^2 \tag{22}$$

4 Experimental Results

In this section, some numerical experiments and simulations are presented to evaluate the performance of the proposed watermarking scheme. In all experiments, the 8-bit grayscale image "Lena" and "Peppers" are tested. Watermark sequence is regarded as the owner's signature which is used to verify image copyrights. In order to implement the proposed watermarking scheme, the experimental parameters should be determined. In the watermark detection scheme, false-alarm probability and false-positive probability are considered to analyze the detector's thresholds. False-alarm probability is the probability of detecting a watermark in the disk which is not present. Moreover, false-positive probability is the probability of not detecting a watermark in a watermarked disk. There is a conflict between these two probabilities. A trade-off between them is decided for selecting the watermark detection threshold (T_1=140).

4.1 Imperceptibility

The imperceptibility determines to which extent the embedding process has altered the perceptual image quality. Image quality is usually measured using the peak signal-to-noise rate (PSNR) value between the original image and watermarked image

$$\mathrm{PSNR}(I, I') = 10\log\left(\frac{255^2 \times M \times N}{\sum_{i=1}^{M}\sum_{j=1}^{N}[f(x_i,y_j) - f'(x_i,y_j)]^2}\right) \tag{23}$$

where $f(x_i, y_j)$ is the original image and $f'(x_i, y_j)$ is the watermarked version, both with dimensions $M \times N$. The PSNR of a watermarked image is determined by two main factors. On the one hand, given a fixed number of bits to be embedded, the PSNR is determined by quantization step size Δ. A larger quantization step size leads to a stronger watermark, but results in a lower PSNR, and vice versa. On the other hand, given fixed quantization step size or watermark strength, the number of bits to be embedded decides the PSNR of the watermarked image. The more bits embedded the lower the value of PSNR, and vice versa. Furthermore, the lower order Bessel–Fourier Moments is perceptually significant components of the image, and the higher order Bessel–Fourier Moments has poor robustness. It is the same as saying that the maximum moment order T also has an effect on PSNR. A larger T will improve the quality of the image patch reconstruction, but will increase the computational cost. Therefore, the maximum moment order for reconstruction is set to 20. In our experiments, we set the quantization step Δ 18 and the watermark length L 224. Thus, the overall resulting PSNR is great than 46dB. Fig. 2 shows the original images, the watermarked images and the residuals between the original and watermarked images. As shown in Fig. 2 (a) and (b), it is clear that the embedded watermarks are perceptually invisible.

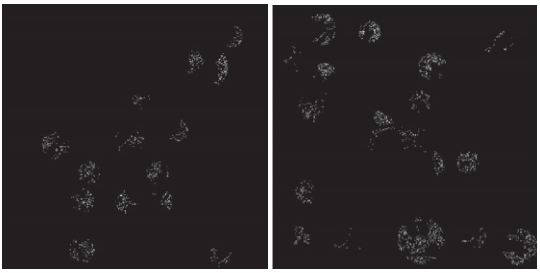

(a) (b)

Fig. 2. the residuals between the original and watermarked images (a) Lena (PSNR=48.27) and (b) Peppers (PSNR=47.81)

4.2 Robustness

Apart from the invisibility test, most of attacks listed in Stirmark4.0 are applied to the test data set to evaluate the robustness of the proposed scheme. Tables 1 and 2 show the simulation results of the proposed scheme in comparison with representative schemes [4] under the common image processing operations and the geometric distortions for Lena and Peppers images. The values in main table unites indicate the detection ratio, which refers to the ratio of the number of disks where watermarks are successfully detected from attacked images to the number of the original watermarked disks. From Tables 1 and 2,it is shown that the proposed image watermarking is not only robust against common image processing operations such as Media filter, Media filter, and JPEG compression etc, but also robust against the geometric distortions such as rotation, translation, scaling and shearing. Furthermore, it has the superior and remarkable performance compared with other scheme.

Table 1. The performance of common image processing

Attacks	Lena		Peppers	
	Proposed scheme	**[4] scheme**	**Proposed scheme**	**[4] scheme**
JPEG 70	15/17	8/10	16/18	9/11
JPEG 50	13/17	7/10	16/18	8/11
JPEG 30	9/17	5/10	12/18	6/11
Media filter 3×3	11/17	4/10	14/18	5/11
Gaussian filter 3×3	8/17	3/10	9/18	4/11
Sharpening 3×3	16/17	8/10	17/18	9/11
Media filter 3×3+ JPEG 90	10/17	4/10	14/18	5/11
Gaussian filter 3×3+ JPEG 90	10/17	3/10	9/18	4/11

Table 2. The performance of geometric distortions

Attacks	Lena		Peppers	
	Proposed scheme	**[4] scheme**	**Proposed scheme**	**[4] scheme**
Cropping 5%	10/17	7/10	9/18	5/11
Cropping 10%	9/17	6/10	7/18	4/11
Scaling 90%	7/17	5/10	7/18	4/11
Scaling 130%	10/17	7/10	10/18	5/11
Shearing 1%	8/17	6/10	8/18	5/11
Rotation 5°	5/17	3/10	7/18	4/11
Rotation 30°	5/17	3/10	6/18	2/11
Random bending	6/17	1/10	4/18	0/11
Translating (x-10 and y-10)	12/17	1/10	14/18	2/11

5 Conclusions

De-synchronization attack is commonly comprehended as one of the most complicated attacks to resist, due to minor geometric distortions can confuse most watermarking algorithms and hence causes incorrect watermark detection. It is a challenging task to design a blind image watermarking algorithm resistance against de-synchronization attacks. In this paper, it proposed a novel geometrically invariant watermarking scheme based on Harris-Laplace detector and local Bessel–Fourier moments with minor visual distortion, accurate retrieval capability and reasonable resistance against watermark de-synchronization attacks. Firstly, Hessian-Laplace detector is adapted to extract feature points. Then, non-overlapped disks centered at feature points are generated. These disks are normalized for the invariance to rotation, scaling and translation distortions. Finally, the watermark is embedded in magnitudes of Bessel–Fourier moments of each disk via dither modulation to realize the robustness to common image processing operations and de-synchronization attacks. In the simulation results, it demonstrates that that the proposed image watermarking is not only robust against common image processing operations such as Media filter, Media filter, and JPEG compression etc, but also robust against the geometric distortions such as rotation, translation, scaling, shearing.

References

1. Cox, I.J., Miller, M.L.: The first 50 years of electronic watermarking. In: EURASIP J. Appl. Signal Processing, pp. 126–132 (2002)
2. Zheng, D., Wang, S., Zhao, J.: RST Invariant Image Watermarking Algorithm With Mathematical Modeling and Analysis of the Watermarking Processes. IEEE Transactions on Image Processing, 1055–1068 (2009)
3. O'Ruanaidh, J., Pun, T.: Rotation, scale, and translation invariant digital image watermarking. Signal Processing, 303–317 (1998)
4. Singhal, N., Lee, Y.Y., Kim, C.S., Lee, S.U.: Robust image watermarking using local Zernike moments. Journal of Visual Communication and Image Representation, 408–419 (2009)
5. Bin, X., Ma, J.F., Xuan, W.: Image analysis by Bessel–Fourier moments. Pattern Recognition, 2620–2629 (2010)
6. Zheng, D., Liu, Y., Zhao, J., Saddik, A.E.: A survey of RST invariant image watermarking algorithm. ACM Comp. Surveys, 1–91 (2007)

Hierarchical Groups with Low Complexity Block Compensation for Reconstruction Video Sequences

Ta-Te Lu[1], Tsung-Hsuan Tsai[2], and Jia-Yuan Wu[1]

[1] Department of Computer Science & Information Engineering, Ching Yun University, Chung-Li, Taiwan
[2] Department of Electric Engineering, Ching Yun University, Chung-Li, Taiwan
{ttlu,m9852003}@cyu.edu.tw, tsunghsuantsai@gmail.com

Abstract. The amounts of video streaming are transmitting to cause network delay in transmission by limited network bandwidth. Thus, the high video quality retrieval in client is still a challenge problem. The motivation of this paper is to reduce the total computing time in reconstruction and raise reconstruction video quality. In this paper, we propose hierarchical groups with low complexity block compensation method using eigenvalues and local variance statistics in each group of picture (GOP), which is regarded as the feature parameters of each video frame between low-resolution and high-resolution sequences. The index selection mechanism is then applied to those blocks to compensate for blocks that have variances larger than a threshold to reduce the transmission amounts. The results of index numbers from index selection are regarded as reconstruction information. Simulation results show the percentage of non-compensation blocks are nearly 36% saving for Foreman and 48% saving for News, respectively.

Keywords: Block compensation, eigenvalues, variance.

1 Introduction

The Internet is a communication infrastructure that interconnects the global community of end users and content servers. In many applications, e, g, IP-TV, IP-surveillance, Live sports, are transmitting the video streaming to receiver ends over the Internet [1]. However, the amounts of video streaming cause network delay in transmission by limited network bandwidth. Network congestion will affect receiver video quality and limit the amounts of video in end-user. Thus, the high video quality retrieval in client is still a challenge problem.

Super-resolution (SR) is a technique to obtain a high-resolution (HR) frame from a sequence of low-resolution (LR) frames. Thus, SR is an efficient method to improve the visual quality for clients. SR reconstruction technology has many applications, e.g. medical, remote sensing, video surveillance and video conference. The SR technique has the advantage of an existing low-resolution system that can be still used.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 612–622, 2011.

The super-resolution method was first proposed by Tsai and Huang [2]. The algorithm exploits the correlation between Continuous Fourier transform (CFT) of HR frames and the discrete Fourier transform (DFT) of LR frames in the frequency domain. Traditional SR methods, e.g. projection onto convex sets (POCS) [3], maximum a posteriori (MAP) [4]-[5], non-uniform interpolation [6], iterative [7] estimate one HR frame from a set of LR frames, while the total computational cost of the traditional SR methods is still very high.

In recent years, SR researchers focus on computational efficiency improvement, fast reconstruction or robust SR algorithms [8]-[13]. Anantrasirichai and Canagarajah [8] used a weighting map to decrease inaccuracy of motion estimation and a quantisation noise model for low bitrate in the super-resolution estimator. Hung and Siu [9] propose a translational motion compensation model via frequency classification for video super-resolution system. Banerjee [10] propose critical-based sampling function to apply higher sampling for high motion and lower sampling for edge regions. Zibetti and Mayer [11] proposed a class super-resolution algorithm to exploit the correlation among the frames of video sequence. The weakness of this method is that the computational cost and the numbers of iteration still high. Katartzis and Petrou [12] used Bayesian estimation method for SR reconstruction, while the estimation algorithm is too complexity for computing. Martins et al. [13] used Markov random fields for SR reconstruction and the Iterated Conditional Models for computing the maximum a posteriori conditional probability, while the algorithm is a weakness for blurring distortion. These SR methods still waste too much computing time for video reconstruction.

Therefore, the motivation of this paper is to reduce the total computing time in reconstruction and raise reconstruction video quality. To achieve this, we use eigenvalue to obtain the features in each Group of Pictures (GOP) and variance to estimate blocks' differences between high-resolution (HR) frame and low-resolution (LR) frame. Then, a low complexity block compensation mechanism is applied for video compensation to achieve high visual quality at the receiver end.

The paper is organized as follows: Section 2 presents the blocks diagram of features detection for video encoder. Section 3 presents the proposed of SR reconstruction structure. Simulation results are given in Section 4. Finally, the conclusions are drawn in Section 5.

2 The Blocks Diagram of Features Detection for Video Encoder

The blocks diagram of video encoder with the features detections algorithm, blocks' compensations and index selections are illustrated in Fig. 1. The video coding basically includes three major parts – quantization, motion estimation, and motion compensation. First, DCT transform is used for intra picture correlation reduction. Second, motion estimation and motion compensation technique are used to reduce the temporal redundancy. The feature detection mechanism identifies each GOP's characteristic after the motion compensation, and blocks' compensation is used to compensate values between high-resolution (HR) frame and low-resolution (LR) frame.

The index selection mechanism is applied to those blocks to compensate for blocks that have variances larger than a threshold to reduce the transmission amounts. The results of index numbers from index selection are regarded as SR reconstruction information. The details of the features detections, blocks' compensation and index selections are described separately as follows.

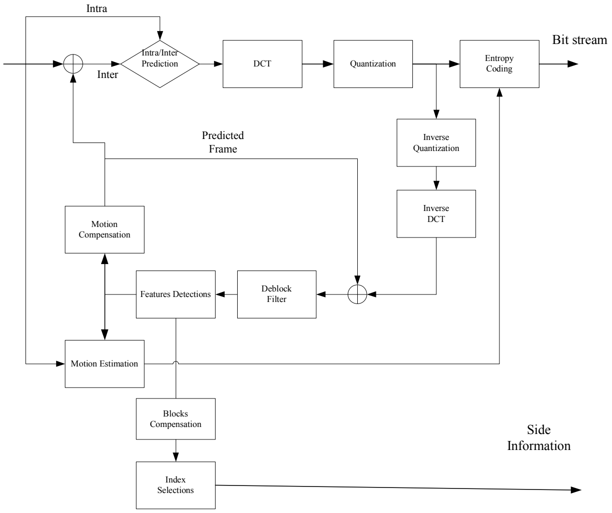


Fig. 1. The blocks diagram of video encoder with the features detections

2.1 Features Detection

The feature detection mechanism identifies each GOP's characteristic after the motion compensation. We denote video sequence with *n* GOPs, each GOP having *m* frames. Each frame has a matrix A with $N \times N$ coefficients, the autocorrelation matrix R_{NxN} is calculated from a matrix A and A^T

$$R_{NxN} = E\left\lfloor AA^T \right\rfloor, \tag{1}$$

where A^T with $N \times N$ coefficients,, and then the eigenvalues $\lambda = \{\lambda_1, \lambda_2, \ldots \lambda_N\}$ is calculated from

$$R_{N \times N} = E[AA^T] = \lambda\, v, \tag{2}$$

where $v = \{v_1, v_2, \ldots, v_N\}$ represent eigenvectors of $E[AA^T]$ associated with λ.

Frame feature F_k in *k*-th frame is calculated as

$$F_k = \frac{1}{r}\sum_{i=1}^{N} \lambda_i \ , \tag{3}$$

where F_k represents the mean of all eigenvalues in k-th frame. The total GOP features denote as $\{G_1, G_2, \ldots, G_n\}$, where G_j is the j-th GOP with m frames $\{F_1, F_2, \ldots, F_m\}$that is calculated as

$$G_j = \frac{1}{m}\sum_{i=1}^{m} F_i \tag{4}$$

2.2 Blocks' Compensation

Variances are used to obtain the features of each video frame between low-resolution and high-resolution sequences. We denote the HR and LR video sequences have m frames, each frame is partitioned into k non-overlapping blocks, each block having p pixels, and then the variance between HR and LR video frames are estimated to obtain the blocks' compensation values as the feature in each video frame. The blocks' variances are denoted as $\sigma = \{\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2\}$, and then the j-th blocks' variance σ_j^2 is calculated as

$$\sigma_j^2 = \sum_{i=1}^{p} \left| (x_i - \eta)^2 - (y_i - \eta)^2 \right|, \tag{5}$$

where x_i is the LR pixel value, y_i is the HR pixel value, $\eta = \frac{\eta_x + \eta_y}{2}$ is the average result of the LR mean η_x and the HR mean η_y. The blocks' variances from $\{\sigma_{\min}^2, \ldots, \sigma_{\max}^2\}$, $\sigma_{\min}^2$ represents that the results $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2\}$are reordered coefficients have high correlation between HR block and LR block, $\sigma_{\max}^2$ represents that the correlation of the block coefficients in HR and LR is very low.

$$\begin{cases} F_b = 1, & \sigma_{\max}^2 - \sigma_{\min}^2 \geq T_0 \\ F_b = 0, & \sigma_{\max}^2 - \sigma_{\min}^2 < T_0 \end{cases} \tag{6}$$

If the variance result larger than threshold T_0, then the block set as a compensation block. Otherwise, the block set as a non-compensation block. If the block is a non-compensation block, then the flag F_b is set to '0' for the non-compensation block. Otherwise, the flag F_b is set to'1' for the compensation block. To reduce transmission overhead, all pixels in compensation blocks indicate one index number in next section.

2.3 Index Selection

The transmission amounts of the features are very large for video sequences reconstruction. To reduce the transmission amounts, the index selection mechanism is applied to those pixels in the compensation blocks F_b=1. Otherwise, those pixels set zero in the non-compensation blocks F_b=0.

Each compensation block has p pixels, and then the differences of pixels between HR and LR video frames are estimated to obtain the blocks' distortion values.

$$D^l_{i,j} = HR^l_{i,j} - LR^l_{i,j}, \tag{7}$$

where $l \in \{1, 2, \ldots, m\}$, $i \in \{1, 2, \ldots, k\}$, $j \in \{1, 2, \ldots, p\}$. The maximum estimation difference value $D^l_{\max}$ and the minimum estimation difference value $D^l_{\min}$ in the l-th frame are determined by (7). The range $\left[D^l_{\min}, D^l_{\max}\right]$ will be divided into t intervals and the interval size Δ is an integer,

$$D^l_i = D^l_{\min} + \Delta \times n_i, \tag{8}$$

where $l \in \{1, 2, \ldots, m\}$, $n_i \in \{1, 2, \ldots, t\}$, $D^l_i \in \{D^l_{\min} + \Delta, D^l_{\min} + 2\Delta, \ldots, D^l_{\max}\}$. The compensation value

$$D^l_{k_i, p_i} = D^l_{\min} + (n_i - \frac{1}{2}) \times \Delta \tag{9}$$

for p_i-th coefficient in k_i-th compensation block, which

$$D^l_{\min} + (n_i - 1)_{p_i} \times \Delta \le D^l_{k_i, p_i} < D^l_{\min} + (n_i)_{p_i} \times \Delta .$$

3 Reconstruction Process

The steps of reconstruction process are described as follows:

Step 1 **:** LR video streaming and side information are received for reconstruction;

Step 2 **:** Video decoder is applied to get the LR video sequence from LR video streaming;

Step 3 : The same procedure is used as in section II to get the features of LR video sequence;

Step 4 : HR reconstruction video sequence is from blocks' compensation. If the flag is '1'for the k_i-th block in the l-th frame, the reconstruction p_i-th coefficient $\tilde{R}^l_{ki,pi}$ is compensated in the block as

$$\tilde{R}^{l}_{ki,pi} = LR^{l}_{ki,pi} + D^{l}_{\min} + (n_i - \frac{1}{2})_{p_i} \times \Delta. \qquad (10)$$

Otherwise, the flag is ‘0’, then the compensation value can be ignored that the reconstruction p_i-th coefficient $\tilde{R}^{l}_{ki,pi}$ is $LR^{l}_{ki,pi}$.

4 Simulation Results

Simulations are performed on various video-coding algorithms, including H.264/AVC and our proposed hierarchical groups with low complexity block compensation methods at the same bit-rate. Class B sequences, e.g. Foreman and News, are sampled at frame rate 5 fps, GOP=10 and encoded at bit-rate 30 kbps. All video sequences are in the QCIF (176x144) format. The experiments were conducted using Matlab with Intel CPU (1.66GHz), 1 GByte memory. The threshold T0 = 8 and $\Delta = \pm 10$ are chosen for the simulation test. Fig. 2 and Fig. 3 show the reconstruction results from low-resolution based on H.264/AVC to high-resolution based on hierarchical groups with low complexity block compensation method for Foreman and News sequences at 30 kbps, respectively. In the condition of the same bit rate including the overhead, on average, hierarchical groups with low complexity block compensation mechanism achieves 10 dB for Foreman and 4 dB for News from low-resolution (LR), respectively. Fig. 4 and Fig. 5 show the 190 th, 288-th reconstruction results for Foreman and the 153-th, 245-th reconstruction results for News, respectively. Both the perceptual quality and the peak signal-to-noise ratio (PSNR) of the proposed algorithm are better than LR (only with H.264) at 30 kbps. These results reveal that the proposed algorithm offers a more efficient way to decrease the percentage of compensation blocks about 30~48% and enhance video quality about 4~10 dB at the receiver end.

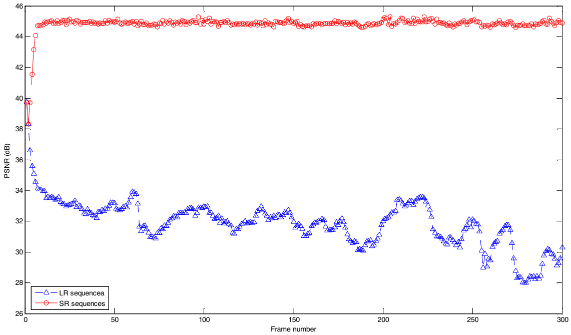


Fig. 2. PSNR results of LR (H.264/AVC) and the proposed algorithm for Foreman sequence at 30 kbps

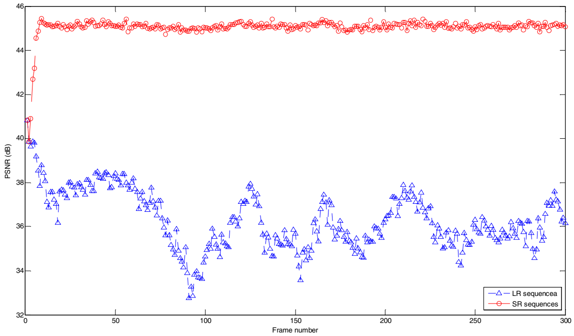
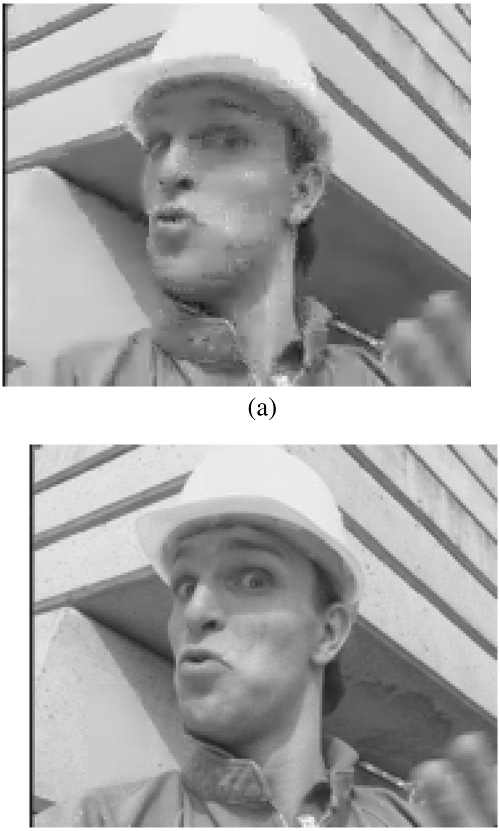


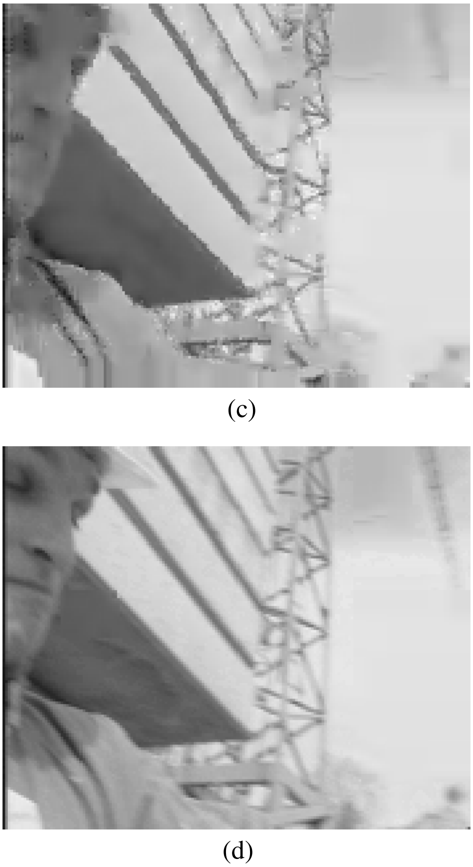
Fig. 3. PSNR results of LR (H.264/AVC) and the proposed algorithm for News sequence at 30 kbps



(a)

(b)

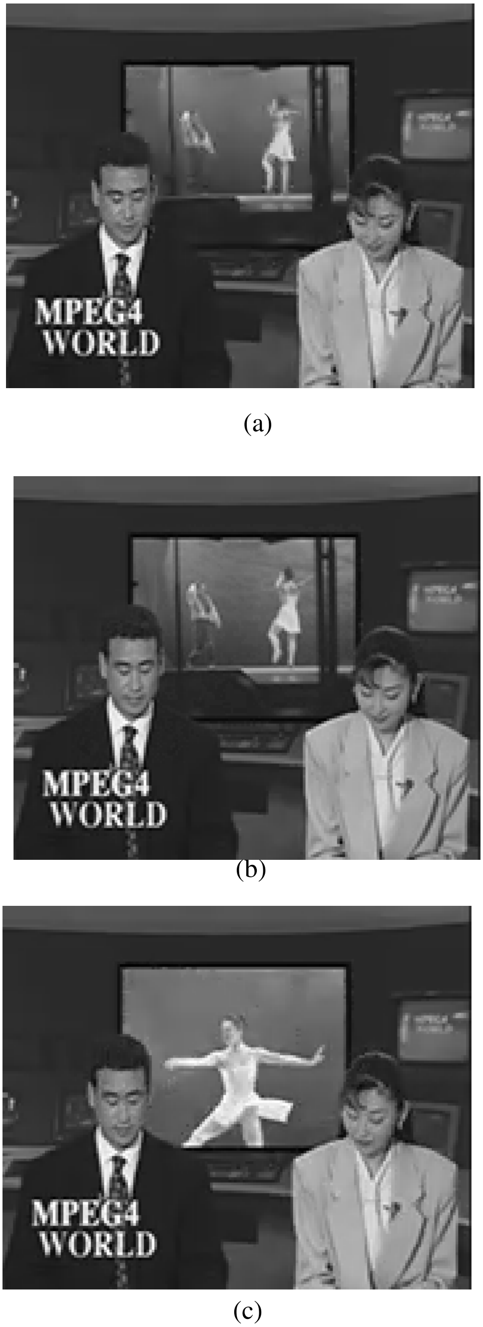
Fig. 4. Reconstruction 19-th frame for Foreman at 30 kbps (a) LR (only with H.264), PSNR=30.46 dB (b) low complexity block compensation mechanism, PSNR=43.88 dB; 288-th frame (c) LR (only with H.264), PSNR=28.46 dB (d) low complexity block compensation mechanism, PSNR=42.40 dB



(c)

(d)

Fig. 4. (*continued*)

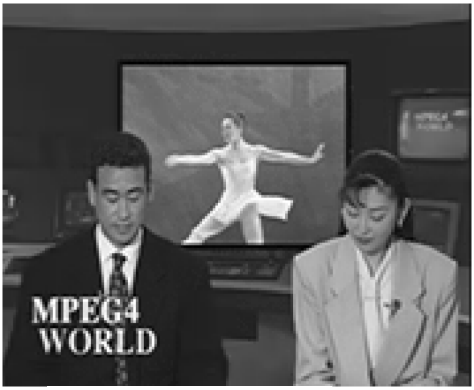


(a)

(b)

(c)

Fig. 5. Reconstruction 153-th frame for News at 30 kbps (a) LR(only with H.264), PSNR=33.57 dB (b) low complexity block compensation mechanism, PSNR=42.22 dB; 245-th (c) LR(only with H.264),PSNR=35.39 dB (d) low complexity block compensation mechanism, PSNR=43.59 dB



(d)

Fig. 5. (*continued*)

5 Conclusions

This work presents hierarchical groups with low complexity block compensation mechanism that is appropriate for lower resolution compensation to achieve high visual quality at the receiver end. The approach is particularly suitable for coding sequences at very low-bit rates. Simulation results show that the proposed method decreases the percentage of compensation blocks about 30~48% and enhances video quality about 4~10 dB at the receiver end. Furthermore, this block compensation mechanism can be applied to other video coding schemes.

Acknowledgments. The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC 99-2221-E-231-031.

References

1. Wiegand, T., Sullian, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264 video coding standard. IEEE Trans. Circuits Syst. Video Technol. 13, 560–576 (2003)
2. Tsai, R.Y., Huang, T.S.: Multi-frame image restoration and registration. Adv. Comput. Vis. Image Process. 1, 317–339 (1984)
3. Gevrekci, M., Gunturk, B.K., Altunbasak, Y.: POCS-Based Restoration of Bayer-Sampled Image Sequences. In: IEEE International Conf. Acoustics, Speech and Signal Processing, 2007, vol. 1, pp. 753–756 (2007)
4. Shen, H., Zhang, L., Huang, B., Li, P.: A MAP Approach for Joint Motion Estimation, Segmentation, and Super Resolution. IEEE Trans. Image Processing. 16, 479–490 (2007)
5. Chantas, G.K., Galatsanos, N.P., Woods, N.A.: Super-Resolution Based on Fast Registration and Maximum a Posteriori Reconstruction. IEEE Trans. Image Processing. 16, 1821–1830 (2007)

6. Panagiotopoulou, A., Anastassopoulos, V.: Super-resolution image reconstruction employing Kriging interpolation technique. In: 14th International Workshop 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, pp. 144–147 (June 2007)
7. Bannore, V., Swierkowski, L.: Fast Iterative Super-Resolution for Image Sequences. In: 9th Biennial Conference of the Australian Pattern Recognition Society Digital Image Computing Techniques and Applications, pp. 286–293 (December 2007)
8. Anantrasirichai, N., Canagarajah, C.N.: Spatiotemporal Super-Resolution for Lowbitrate H.264 Video. In: IEEE International Conf. Image Processing, ICIP 2010, pp. 2809–2812 (September 2010)
9. Hung, K.W., Siu, W.C.: New Motion Compensation Model via Frequency Classification for Fast Video Super-resolution. In: IEEE International Conf. Image Processing, ICIP 2009, pp. 1193–1197 (November 2009)
10. Banerjee, S.: Low-Power Content-Based Video Acquisition for Super-resolution Enhancement. IEEE Trans. on Multimedia 11, 455–464 (2009)
11. Zibetti, M.V., Mayer, J.: A Robust and Computationally Efficient Simultaneous Super-Resolution Scheme for Image Sequences. IEEE Trans. Circuits Syst. Video Technol. 17, 1288–1300 (2007)
12. Katartzis, A., Petrou, M.: Robust Bayesian Estimation and Normalized Convolution for Super-resolution Image Reconstruction. In: IEEE International Conf. Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–7 (June 2007)
13. Martins, A.L.D., Homem, M.R.P., Mascarenhas, N.D.A.: Super-Resolution Image Reconstruction using the ICM Algorithm. In: IEEE International Conf. Image Processing, ICIP 2007, vol. 4, pp. 205–208 (October 2007)

Gait Classification by Support Vector Machine

Hu Ng[1], Hau-Lee Tong[1], Wooi-Haw Tan[2], and Junaidi Abdullah[1]

[1] Faculty of Information Technology, Multimedia University
Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia
[2] Faculty of Engineering, Multimedia University
Jalan Multimedia, 63100 Cyberjaya, Selangor, Malaysia
{nghu,hltong,twhaw,junaidi.abdullah}@mmu.edu.my

Abstract. This paper presents a simple model-free gait extraction approach for human identification by using Support Vector Machine. The proposed approach consists of three parts: extraction of human gait features from enhanced human silhouette, smoothing process on extracted gait features and classification by Support Vector Machine (SVM). The gait features extracted are height, width, crotch height, step-size of the human silhouette and joint trajectories. To improve the classification performance, two of these extracted gait features are smoothened before the classification process in order to alleviate the effect of outliers. The proposed approach has been applied on SOTON covariate database, which is comprised of eleven subjects walking bidirectional in a controlled indoor environment with thirteen different covariate factors that vary in terms of apparel, walking speed, shoe types and carrying objects. From the experimental results, it can be concluded that the proposed approach is effective in human identification from a distance.

Keywords: gait classification, support vector machine, covariate factors.

1 Introduction

Human identification based on biometrics is to distinguish individuals based on their physical and/or behavioural characteristics such as face, fingerprint, gait, iris and spoken voice. Biometrics are getting significant and widely acceptable today because they are unique and one will not lose or forget them over time. Gait is a complex locomotion pattern which involves synchronized movements of body parts, joints and the interaction among them [1]. Basically, every individual has his/her own walking pattern. Thus, it can be considered as a unique feature for biometric. In 1973, psychological research from Johannson [2] has proved that human can easily recognize walking friends based on the light markers that are attached to them. Even since then, much research has been carried out on gait analysis and it has been proven that gait can be used to identify people. As a result, gait has become one of the latest promising biometrics.

Gait offers the ability to identify people at a distance when other biometrics are obscured. Furthermore, it does not require any intervention from the user and can be

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 623–636, 2011.

captured by hidden cameras or synchronised closed-circuit television (CCTV) cameras, which do not require any direct contact with feature capturing device unlike other biometrics. This is also motivated by the increasing number of CCTV cameras that have been installed in many major cities, in order to monitor and prevent crime by identifying the criminal or suspect.

The performance of gait as biometric can be affected by covariate factors, such as light illumination during video capturing, imperfect object segmentation from background, changes in the subject appearance, nature of ground and different camera viewing angle with respect to the subjects. This paper aims to concentrate on extracting gait features regardless of covariate factors such as apparel, carrying objects, shoe types and walking speed of subject.

The remainder of this paper is organised as follows. Section 2 describes the previous works on gait analysis and its classifying tools. Section 3 discusses the proposed system overview. Section 4 provides the details of gait features extraction and SVM model. Section 5 illustrates the experimental set up and Section 6 presents the experiment results. Finally, Section 7 concludes the paper presentation.

2 Previous Works

Basically, gait analysis can be divided into two major categories, namely model-based approach and model-free approach. Model-based approach generally models the human body structure or motion and extracts the features to match them to the model components. It incorporates knowledge of the human shape and dynamics of human gait into an extraction process. The gait dynamics are extracted directly by determining joint positions from model components, rather than inferring dynamics from other measures (such as movement of other objects). Thus, the effect of background noise can be eliminated. Research examples of this approach are static body parameters [3], thigh joint trajectories [4], dual oscillator [5], articulated model [6], 2D stick figure [7] and elliptic Fourier descriptors [8].

The advantages of this approach are the ability to derive dynamic gait features directly from model parameters. It is free from background noise as well as the effect of different subject's apparel or camera shooting viewpoint. However, it creates many parameters from extracted gait features and hence resulting in a complex model. Due to that reason, the computational time, date storage and cost are extremely high due to its complex searching and matching procedures.

Conversely, model-free approach generally differentiates the whole motion pattern of the human body by a concise representation such as silhouette without considering the underlying structure. Normally, its parameters are obtained from the static gait features like centroid, width and height of the silhouette. Research examples of this approach are self similarity Eigen gait [1], key frames analysis [9], spatial-temporal distribution characterization [10], kinematic features [11], unwrapped silhouette [12], higher order correlation [13], video oscillations [14] and gait sequences [15].

The advantages of this approach are speedy processing, low computational cost and small data storage. However, the performance of this approach is highly affected by the background noise and the changes of the subject's apparel.

Gait classification can be defined as human identification based on the variation and characteristics of a subject walking motion. From the gait classification research record, classifier that have been applied are: K-nearest neighbor (KNN) [1, 4, 5, 6, 8, 9, 12, 14], back-propagation neural network algorithm [7], Baseline algorithm [10], Genetic algorithm [11], Fisher discriminant analysis [13], Hausdorff distance [15] and Support Vector Machine (SVM) [15]. Most of the gait classification applied KNN. This is mainly due to its simplicity and ability to handle large data set.

Since gait includes both the physical appearance of body and dynamics of human walking stance [16], this paper presents a model-free silhouette based approach to extract the static gait features (height and width, step size) and dynamic gait features (joint trajectories). This concept of joint trajectory calculation is found faster in process and less complicated than the model-based method like 2D stick figure by Yoo et al. [7], articulated model by Wagg et al. [6] and elliptic Fourier descriptors by Bouchrika et.al. [8]. As there are only a few studies on gait classification using the covariate database [8], [10] [15], this study aims to evaluate the recognition rate of the walking subjects with different covariate factors by using SVM classifier. It has been chosen due to its low sensitivity to data dimensionality.

3 Overview of the System

This paper presents a model-free silhouette based approach to extract the gait features by dividing human silhouette into six body segments and applying Hough transform to obtain the joint trajectories.

For the gait feature extraction, morphological opening is first applied to reduce background noise on the original images downloaded from SOTON covariate database provided by University of Southampton [17]. Each of the human silhouettes is then measured for its width and height. Next, each of the enhanced human silhouettes is divided into six body segments based on anatomical knowledge [18]. Morphological skeleton is later applied to obtain the skeleton of each body segment. The joint trajectories are obtained after applying Hough transform on the skeletons. Step-size, which is the distance between the bottom of both feet. Crotch height, which is the distance between the subject's crotch and the floor, is also determined. The dimension of the human silhouette, step-size, crotch height and two joint trajectories from the body segments are then used as the gait features for classification.

To mitigate the effect of outliers, both thigh trajectory and crotch height are smoothened by Gaussian filter before their average values are applied in the classification process. Even though the smoothing process reduces the peak value of the data, it is not affecting the uniqueness of gait features of the subject. In addition, it also reduces the outlier of the data. This is proven as better gait recognition results have been obtained as compared to the technique without smoothing. Fig. 1 summarizes the process flow of the proposed approach.

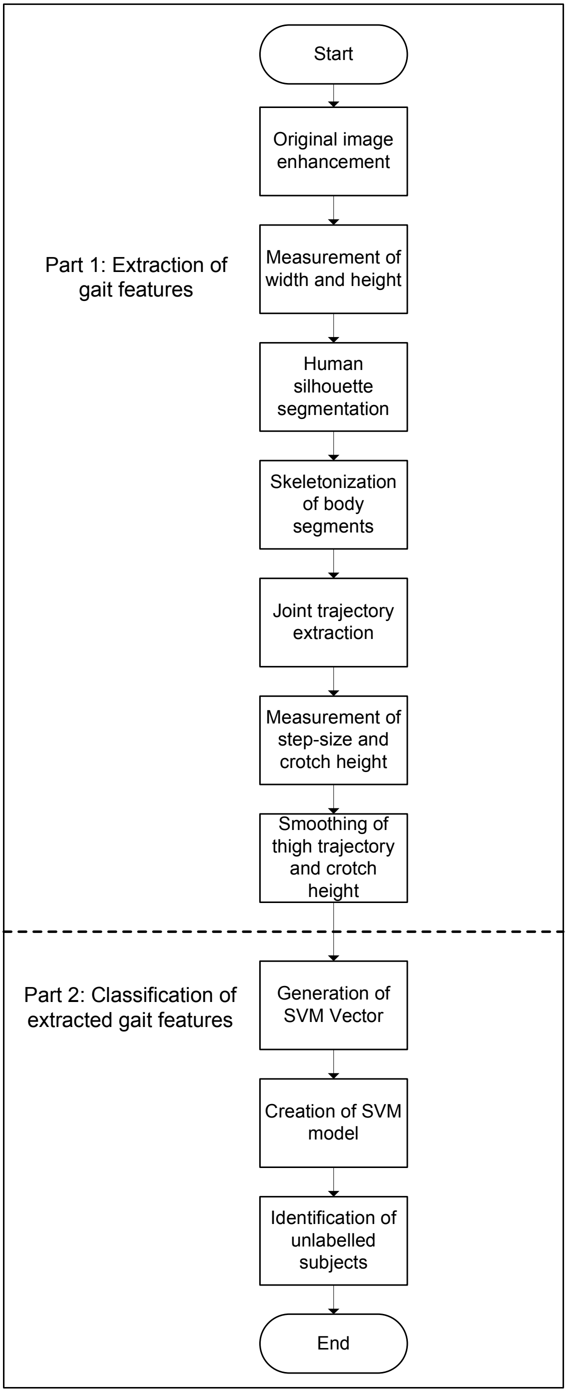
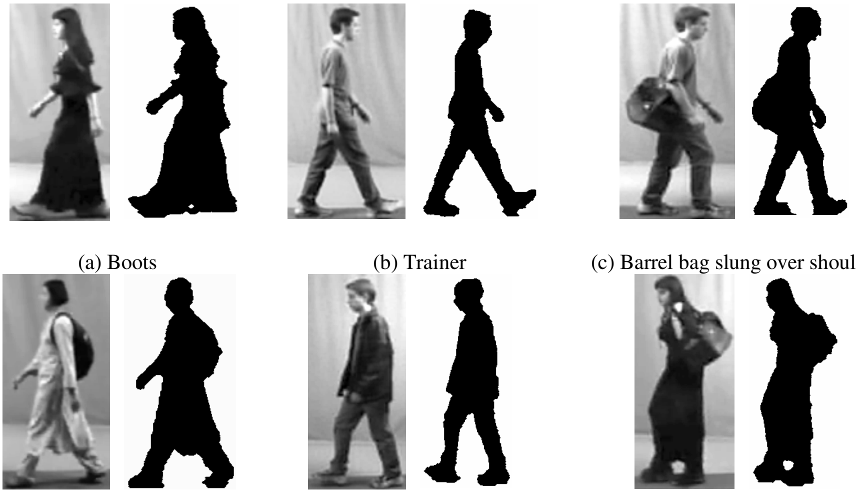

Start
Original image enhancement
Measurement of width and height
Human silhouette segmentation
Skeletonization of body segments
Joint trajectory extraction
Measurement of step-size and crotch height
Smoothing of thigh trajectory and crotch height
Part 1: Extraction of gait features
Part 2: Classification of extracted gait features
Generation of SVM Vector
Creation of SVM model
Identification of unlabelled subjects
End

Fig. 1. Flow chart of the proposed approach

4 Extracting the Gait Features

The original human silhouette images are obtained from the well-known SOTON covariate database [17]. It consists of eleven walking subjects walking bidirectional on an indoor track, with a green chroma-key backdrop. The video was captured by CANON camcorders with 25 frames per second. Background subtraction approach has been applied to segment out the subject from the background. The generated silhouette images have the resolution of 720 x 576 (width x height) pixels.

The database consists of nine males and three females. All the males and one of the females are wearing trousers, whereas another two females are wearing attires that cover the legs, full blouse and Indian cloth. This created the occlusion of knee, thigh, and hip joints on their images. Each subject was captured wearing a variety of shoe types, apparel and carrying various objects. They were also recorded walking at different speed. This database was used to evaluate the recognition rate of the walking subjects with different covariate factors. The examples of subjects with different apparel, shoe types and carrying various objects can be found in Fig. 2.



(a) Boots (b) Trainer (c) Barrel bag slung over shoulder

(d) Rucksack (e) Rain coat (f) Barrel bag carried by hand

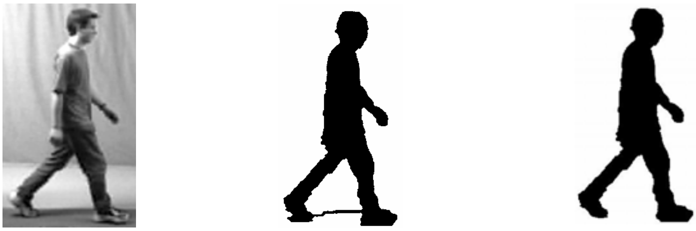
Fig. 2. Examples of subjects with different apparel, shoe types and carrying various objects. Left: original image. Right: silhouette image

4.1 Original Image Enhancement

In most of the human silhouette images, shadow is found especially near to the feet. It appears as part of the subject body in the human silhouette image as shown in Figure 2. The presence of the artifact affects the gait feature extraction and the measurement of joint trajectories. The problem can be reduced by applying a morphological opening operation with a 7×7 diamond shape structuring element, as denoted by

$$A \circ B = (A \ominus B) \oplus B) \qquad (1)$$

where A is the image, B is the structuring element, $\ominus$ represents morphological erosion and $\oplus$ represents morphological dilation. The opening first performs erosion, followed by dilation. Fig. 3 shows the original and enhanced images.



(a) Original video image (b) Original silhouette image (c) Enhanced silhouette image

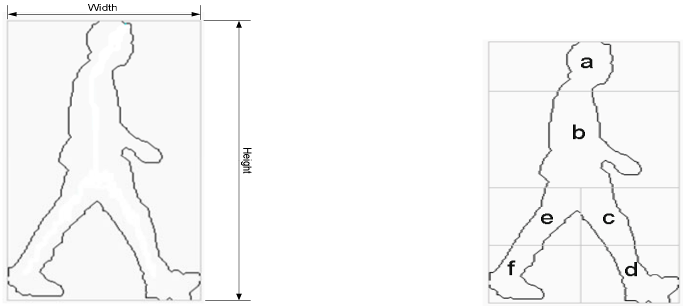
Fig. 3. Original and enhanced images after morphological opening

4.2 Measurement of Width and Height

The width and height of the subject from each frame during the walking sequences are measured from the bounding box of the enhanced human silhouette, as shown in Fig. 4(a). These two features will be used for gait analysis in the later stages.

4.3 Dividing Enhanced Human Silhouette

Next, the enhanced human silhouette is divided into six body segments based on anatomical knowledge [18] as shown in Fig. 4(b). Where *a* represents head and neck, *b* represents torso, *c* represents right hip and thigh, *d* represents right lower leg and foot, *e* represents left hip and thigh and *f* represents left lower leg and foot.



(a) (b)

Fig. 4. (a) Width and height of human silhouette. (b) Six body segments.

4.4 Skeletonization of Body Segments

To reduce the segments to a simpler representation, morphological skeleton is used to construct the skeleton from all the body segments. Skeletonization involves consecutive erosions and opening operations on the image until the set differences between the two operations are zero. The operations are given as:

Erosion	Opening	Set differences	
$A\ominus kB$	$(A\ominus kB)\circ B$	$(A\ominus kB)-((A\ominus kB))\circ B$	(2)

where A is an image, B is the structuring element and k is from zero to infinity. Fig. 5 shows an example of skeleton of the body segments from a series of frames.

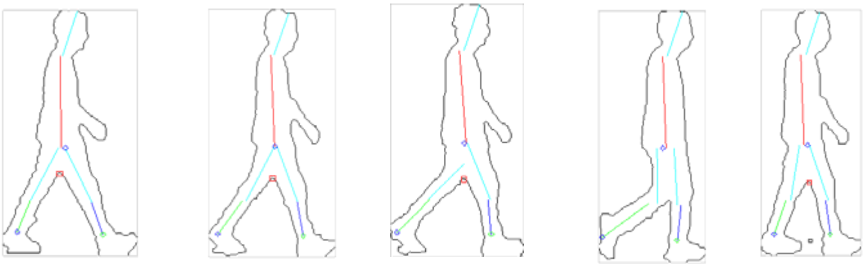


Fig. 5. Example of skeleton of the body segments from a series of frames

4.5 Joint Trajectory Extraction

To extract the joint trajectory for each body segment, Hough transform is applied on the skeleton. Hough transform maps pixels in the image space to straight lines in the parameter space. The skeleton in each body segment, which is the most probable straight line, is indicated by the highest intensity point in the parameter space.

4.6 Measurement of Step-Size and Crotch Height

To obtain the step-size of each walking frame, Euclidean distance between the bottom ends of both feet are measured. To obtain the crotch height, the distance between the subject's crotch and the floor is measured. If the crotch height is lower than the knee height, it will be deduced to 0. Fig. 6 shows all the gait features extracted from a human enhanced silhouette, where θ_3 is the thigh trajectory, calculated as:

$$\theta_3 = \theta_2 - \theta_1 \tag{3}$$

where θ_2 is the front knee trajectory and θ_1 is the back knee trajectory.

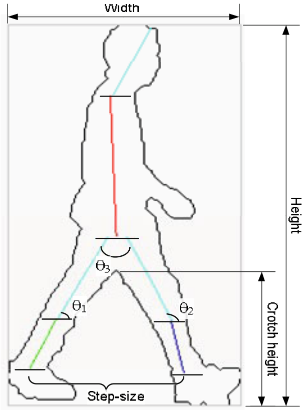


Fig. 6. The entire extracted gait features

4.7 Smoothing Technique

It can be observed from the collected gait features that the changes in crotch height and thigh trajectory are sinusoidal over time. However, the curves for crotch height and thigh trajectory are uneven due to the presence of outliers. Fig. 7 shows an example of the original thigh trajectory over time. Therefore, smoothing is required to reduce the effect of these outliers.

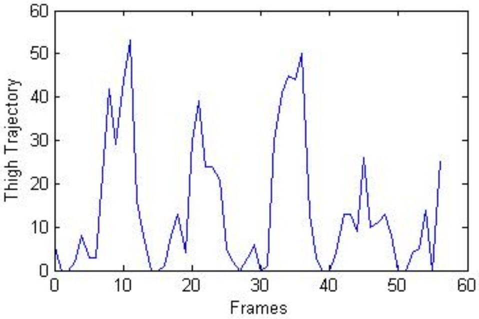


Fig. 7. Changes in original thigh trajectory over time

Therefore, Gaussian filter has been applied to overcome those outliers. It is designed to give no overshoot to a step function input while minimizing the rise and fall time. It generates a bell-shaped curve after the smoothing process. The operation is shown as:

$$y_j' = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_j-\mu)^2}{2\sigma^2}} \tag{4}$$

where parameters μ and σ^2 are the mean and the variance, and y is the raw data. Fig. 8 shows the smoothed thigh trajectory by using Gaussian filter with σ of 1.4.

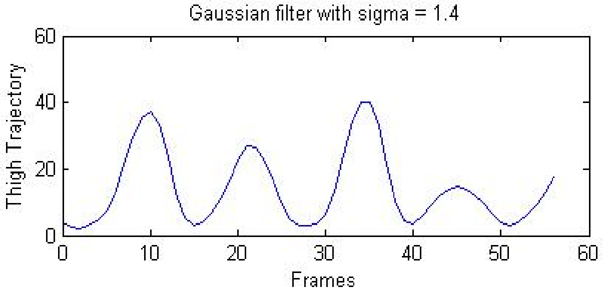


Fig. 8. Smoothed thigh trajectory over time by using Gaussian filer with σ of 1.4

4.8 SVM Classifier

After feature extraction, multiclass SVM is employed for classification. For the SVM technique used in this paper, we refer to the description by C.J.C. Burges [19] and implement the SVM experiments by LIBSVM package [20].

SVM is based on structural risk minimization principle, which optimizes the training data to create machine learning model. Given a training set of instance-label pairs: (x_i, y_i) , i = 1, 2…, l where $x_i \in R^n$ and $y_i \in \{1,-1\}^i$ (n and l denote the space dimensions and size of training set). In this case if x belongs to positive category then y_i =1; if x belongs to negative category then y_i = - 1.

Basically SVM is a classifier that focuses on finding an optimal hyperplane to separate different classes by solving the following quadratic optimization problem:

$$\text{Minimise: } \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{N} \xi_i \tag{5}$$

Subject to $y_i(w \bullet x_i) + b \geq 1 - \xi_i, \xi_i \geq 0$. Where, w, b and ξ_i denote weight vector for learned decision hyperplane, model bias and slack variable. Parameter C is penalty factor which keeps the balance of classification accuracy. For our approach, C is to be fixed as with the value of one.

SVM classifies the test instance, X based on the following decision function:

$$f(x) = \text{sign} \left(\sum_{x_i \in sv} \alpha_i y_i K(x_i, x) + b \right) \tag{6}$$

where *sv*, α_i and $K(x_i, x)$ represent support vectors, Lagrange multipliers and kernel function. For this paper, we applied linear kernel SVM, therefore $K(x_i, x)$ is equal to $x_i \bullet x$.

5 Experimental Set Up

The experiment was carried out for eleven subjects walking parallel to a static camera, with thirteen covariate factors. Each subject was captured wearing a variety of footwear (flip flops, bare feet, socks, boots, own shoes and trainers), clothes (normal or with rain coat) and carrying various objects (barrel bag slung over shoulder or carried by hand, and rucksack). They were also recorded walking at different speeds (slow, fast and normal speed).

For each subject, there are approximately twenty sets of walking sequences, which are from left to right and vice-versa on normal track. In total, there are 2722 walking sequences that are used for training and testing process.

In order to obtain the optimized results, four gait features were adopted. Firstly, maximum thigh trajectory, $\theta_3^{\max}$ was determined from all the thigh trajectories collected during a walking sequence. When $\theta_3^{\max}$ was located, the corresponding values for the step-size, *S* and width, *w* and height, *h* were determined as well. To improve the recognition rate, additional features were used. These features were the average of the local maxima detected for width (A^W), smoothed crotch height (A^{HS}) and smoothed thigh trajectory (A^{TS}). Therefore, one SVM feature vector will be generated from each walking sequence with the extracted seven features. In total, 2722 feature vectors were created and used for SVM classification.

6 Experiment Results

For the classification, two-folds cross validation was applied by partitioning all the feature vectors into two disjoint subsets. By performing the cross validation, each disjoint subsets will be used in training and testing. This is to ensure that every feature vector will be trained and tested in order to evaluate the results appropriately. The results obtained from the two-folds cross validation are then normalized to produce a sole recognition rate.

To evaluate the experimental results by smoothing process, different sigma values (σ) of Gaussian filter have been used during the smoothing process. The overall results are summarized in Table 1. From Table 1, the highest recognition rate (83.21%) is obtained with σ = 1.6, which outperforms the recognition rate without smoothing by 2.68%. This has brought an additional 73 vectors to be correctly classified. Hence, it is prove that the smoothing process for crotch height and thigh trajectory has significant contribution.

Table 1. Recognition rates with different sigma values of Gaussian filter

Gaussian filter, sigma value (σ)	Total correct classified vectors	Recognition rate (%)
No smoothing	2192	80.53
0.8	2247	82.55
0.9	2248	82.59
1.0	2253	82.77
1.1	2251	82.70
1.2	2245	82.48
1.3	2230	81.93
1.4	2239	82.26
1.5	2250	82.66
1.6	2265	83.21
1.7	2263	83.14
2.0	2226	81.78

The extended analysis for the result of σ = 1.6 was conducted by using the Cumulative Match Scores (CMS) as initiated by Philips et. al. [21]. CMS shows the probability of correct identification versus relative rank *k*. It depicts the recognition rate for the correct subject with respect to top *n* matches. By referring to Fig 9, it can be observed that most of the subjects are correctly classified at rank 2, which achieved 95% performance. The recognition rate reached 100% at the rank 6, which mean all the subjects can be correctly identified within top six matching.

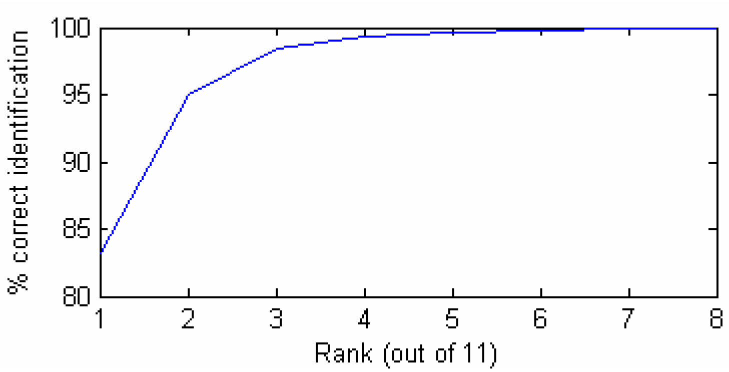


Fig. 9. Cumulative Match Scores of top 8 matches for all subjects

The results of CMS across the subjects are also determined, as shown in Table 2. The recognition rate for each subject was found to be around 95% at rank 3.

For comparison with other approaches that use the same SOTON covariate database, our approach performed better than the recognition rate of 73.4% as reported by Bouchrika et al. [22]. Even though the recognition rate achieved by

Pratheepan et. al. is 86%, we believed that our approach is better after comparing the number of subjects , the number of covariate factors and the number of walking sequences that were used for training and testing. Table 3 summarizes the comparison.

Table 2. Cumulative Match Scores of top 3 matches across the subjects

Subject	Rank 1 (%)	Rank 2 (%)	Rank 3 (%)
1	88.28	97.07	98.75
2	88.80	94.40	97.20
3	69.02	87.45	94.90
4	79.22	94.37	97.84
5	97.61	99.60	99.60
6	95.98	98.00	99.60
7	73.55	99.28	99.40
8	60.16	92.03	99.60
9	86.40	94.40	97.60
10	95.69	99.60	99.60
11	95.80	95.80	95.80

Table 3. Benchmark with other approaches on SOTON covariate database

	Bouchrika et. al. [22]	Pratheepan et. al. [15]	Our approach
Recognition rate (%)	73.4	86	83.2
Number of subjects	Ten	Ten	Eleven**
Number of covariate factors	Eleven	Four #	Thirteen
Number of walking sequences	440	180	2722

** Subjects include one female wearing long blouse

# Without covariate factors on shoe types and walking speed.

7 Conclusion

A novel model-free approach for extracting the gait features from enhanced human silhouette image has been developed. The gait features are extracted from human silhouette by determining the skeleton from body segments. The joint trajectories are obtained after applying Hough transform on the skeleton. Both crotch height and thigh trajectory had been smoothened before their average values were applied for classification by SVM. The results show that the proposed method can perform well regardless of walking speed, carrying objects and apparel of subject. Future development includes experiments on other gait databases and the application of various classification algorithms.

Acknowledgment

The authors would like to thank Professor Mark Nixon, School of Electronics and Computer Science, University of Southampton, United Kingdoms for his kindness in providing the SOTON covariate database for use in this work.

References

1. BenAbdelkader, C., Cutler, R., Nanda, H.: L. Davis, L.: Eigen Gait: Motion-based Recognition of People Using Image Self-similarity. In: Proc. of International Conference Audio and Video-Based Person Authentication, pp. 284–294 (2001)
2. Johannson, G.: Visual Perception of Biological Motion and A Model For Its Analysis. Perception and Psychophysics, 201–211 (1973)
3. Bobick, A.F., Johnson, A.Y.: Gait Recognition Using Static, Activity-specific Parameters. In: Proc. of IEEE Computer Vision and Pattern Recognition, I, vol. 1, pp. 423–430 (2001)
4. Cunado, D., Nixon, M.S., Carter, J.N.: Automatic Extraction and Description of Human Gait Models for Recognition Purposes. Computer and Vision Image Understanding 90(1), 1–41 (2003)
5. Yam, C., Nixon, M.S., Carter, J.N.: Automated Person Recognition by Walking and Running via Model-Based Approaches. Pattern Recognition 37(5), 1057–1072 (2004)
6. Wagg, K., Nixon, M.S.: On Automated Model-Based Extraction and Analysis of Gait. In: Proc. of 6th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 11–16 (2004)
7. Yoo, J.-H., Nixon, M.S., Harris, C.J.: Extracting Human Gait Signatures by Body Segment Properties. In: Fifth IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 35–39 (2002)
8. Bouchrika, I., Nixon, M.S.: Model-based Features Extraction for Gait Analysis and Recognition. In: Proc. of Mirage: Computer and Vision / Computer Graphics Collaboration Techniques and Applications, pp. 150–160 (2007)
9. Collin, R., Gross, R., Shi, J.: Silhouette-based Human Identification from Body Shape and Gait. In: Proc. of Fifth IEEE International Conference, pp. 366–371 (2002)
10. Phillips, P.J., Sarkar, S., Robledo, I., Grother, P., Bowyer, K.: The Gait Identification Challenge Problem: Dataset and Baseline Algorithm. In: Proc. of 16th International Conference Pattern Recognition, vol. I, pp. 385–389 (2002)
11. Bhanu, B., Han, J.: Human Recognition on Combining Kinematic and Stationary Features. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, Springer, Heidelberg (2003)
12. Wang, L., Tan, T.N., Hu, W.M., Ning, H.Z.: Automatic Gait Recognition Based on Statistical Shape Analysis. IEEE Transactions on Image Processing 12(9), 1120–1131 (2003)

13. Kobayashi, T., Otsu, N.: Action and Simultaneous Multiple-Person Identification Using Cubic Higher-order Local Auto-Correlation. In: Proceedings 17th International Conference on Pattern Recognition (2004)
14. Boyd, J.E.: Synchronization of Oscillations for Machine Perception of Gaits. Computer Vision and Image Understanding 96, 35–59 (2004)
15. Pratheepan, Y., Condell, J.V., Prasad, G.: Individual Identification Using Gait Sequences under Different Covariate Factors. In: Fritz, M., Schiele, B., Piater, J.H. (eds.) ICVS 2009. LNCS, vol. 5815, pp. 84–93. Springer, Heidelberg (2009)
16. Lee, L., Grimson, W.: Gait Analysis for Recognition and Classification. In: Proc. International Conference Automatic Face and Gesture Recognition, pp. 155–162 (2002)
17. Shutler, J.D., Grant, M.G., Nixon, M.S., Carter, J.N.: On A Large Sequence-based Human Gait Database. In: Proc. of 4th International Conference on Recent Advances in Soft Computing, Nottingham, UK, pp. 66–71 (2002)
18. Dempster, W.T., Gaughran, G.R.L.: Properties of Body Segments Based on Size and Weight. American Journal of Anatomy 120, 33–54 (1967)
19. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
20. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), Software, `http://www.csie.ntu.edu.tw/~cjlin/libsvm`
21. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.: The FERET Evaluation Methodology For Face-recognition Algorithms. IEEE Transaction on Pattern Analysis and Machine Intelligence 22(10), 1090–1104 (2000)
22. Bouchrika, I., Nixon, M.S.: Exploratory Factor Analysis of Gait Recognition. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition (2008)

Selecting Significant Features for Authorship Invarianceness in Writer Identification

Azah Kamilah Muda, Satrya Fajri Pratama, Yun-Huoy Choo,
and Noor Azilah Muda

Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka. Hang Tuah Jaya,
76100 Durian Tunggal, Melaka, Malaysia
azah@utem.edu.my, rascove@yahoo.com, huoy@utem.edu.my,
azilah@utem.edu.my,

Abstract. Handwriting is individualistic. The uniqueness of shape and style of handwriting can be used to identify the significant features in authenticating the author of writing. Acquiring these significant features leads to an important research in Writer Identification domain where to find the unique features of individual which also known as Individuality of Handwriting. It relates to invarianceness of authorship where invarianceness between features for intra-class (same writer) is lower than inter-class (different writer). This paper discusses and reports the exploration of significant features for invarianceness of authorship from global shape features by using feature selection technique. The promising results show that the proposed method is worth to receive further exploration in identifying the handwritten authorship.

Keywords: feature selection, authorship invarianceness, significant features.

1 Introduction

Feature selection has become the focus of research area for a long time. The purpose of feature selection is to obtain the most minimal sized subset of features [1]. Practical experience has shown that if there is too much irrelevant and redundant information present, the performance of a classifier might be degraded. Removing these irrelevant and redundant features can improve the classification accuracy.

The three popular methods of feature selection are filter method, wrapper method, and embedded method has been presented in [2]. Filter method assesses the relevance of features [3], wrapper method uses an induction algorithm [4], while embedded method do the selection process inside the induction algorithm [5]. Studies have shown that there are no techniques more superior compared to others [6].

Writer Identification (WI) can be included as a particular kind of dynamic biometric in pattern recognition for forensic application. WI distinguishes writers based on the shape or individual style of writing while ignoring the meaning of the word or character written. The shape and style of writing are different from one person to another. Even for one person, they are different in times. However,

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 637–651, 2011.

everyone has their own style of writing and it is individualistic. It must be unique feature that can be generalized as significant individual features through the handwriting shape.

Many previous works on WI problem has been tried to be solved based on the image processing and pattern recognition technique [7], [8], [9], [10], [11] and involved feature extraction task. Many approaches have been proposed to extract the features for WI. Mostly, features are extracted from the handwriting focus on rigid characteristics of the shape such as [7], [9], [11], [12], [13], [14], [15], [16], [17] except by [18] and [19], focus on global features.

The main issue in WI is how to acquire the features that reflect the author of handwriting. Thus, it is an open question whether the extracted features are optimal or near-optimal to identify the author. Extracted features may include many garbage features. Such features are not only useless in classification, but sometimes degrade the performance of a classifier designed on a basis of a finite number of training samples [20]. The features may not be independent of each other or even redundant. Moreover, there may be features that do not provide any useful information for the task of writer identification [21]. Therefore, feature extraction and selection of the significant features are very important in order to identify the writer, moreover to improve the classification accuracy.

Thus, this paper focuses on identifying the significant features of word shape by using the proposed feature selection technique prior the identification task. The remainder of the paper is structured as follows. In next section, an overview of individuality of handwriting is given. Global feature representation by United Moment Invariant is described in Section 3. Section 4 provides an overview of feature selection techniques, followed by the proposed approach to identify the significant features in Section 5. Finally, conclusion and future work is drawn in Section 6.

2 Authorship Invarianceness

Handwriting is individual to personal. Handwriting has long been considered individualistic and writer individuality rests on the hypothesis that each individual has consistent handwriting [10], [18], [23], [24], [25]. The relation of character, shape and the style of writing are different from one to another.

Handwriting analysis consists of two categories, which are handwriting recognition and handwriting identification. Handwriting recognition deals with the contents conveyed by the handwritten word, while handwriting identification tries to differentiate handwritings to determine the author. There are two tasks in identifying the writer of handwriting, namely identification and verification. Identification task determines the writer of handwriting from many known writers, while verification task determines whether one document and another is written by the same writer.

The challenge in WI is how to acquire the features that reflect the author for these variety styles of handwriting [7], [9], [12], [13], [15], [24], either for one writer or many writers. These features are required to classify in order to identify the variance between features for same writer is lower than different writer which known as Authorship Invarianceness. Among these features are exists the significant individual features which directly unique to those individual. Figure 1 shows that each person has its individuality styles of writing. The shape is slightly different for the same writer and quite difference for different writers.

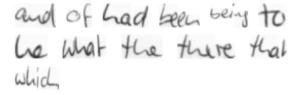
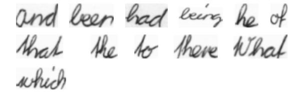
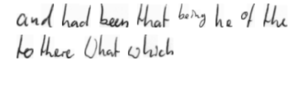
Writer 1	Writer 2	Writer 3
		

Fig. 1. Different Word for Different Writer

3 Global Features Representation

In pattern recognition problem, there are many shape representations or description techniques have been explored in order to extract the features from the image. Generally it can be classified into two different approaches when dealing with handwritten word problem, which are analytic (local / structural approach) and holistic (global approach) [26], [27]. For the each approach, it is divided into two method, which are region-based (whole region shape) methods and contour-based (contour only) methods. Holistic approach represent shape as a whole, meanwhile analytic approach represents image in sections. In this work, holistic approach of United Moment Invariant (UMI) is chosen due to the requirement of cursive word is needed to extract as one single indivisible entity. This moment function of UMI is applied in feature extraction task.

The choice of using holistic approach is not only based on the holistic advantages, but also due to its capability of using word in showing the individuality for writer identification problem as mentioned in [18] and holds immense promise for realizing near-human performance [28]. The holistic features and matching schemes must be coarse enough to be stable across exemplars of the same class such as a variety of writing styles [29]. This is aligning with this work where to extract the unique global features from word shape in order to identify the writer.

Global features extracted with this holistic approach are invariant with respect to all different writing styles [29]. Words in general may be cursive, minor touching discrete, purely discrete, one or two characters are isolated and others are discrete or mixture of these style and it still as one word. Global technique in holistic approach will extract all of these styles for one word as one whole shape. Shape is an important representation of visual image of an object. It is a very powerful feature when it is used in similarity search. Unlike color and texture features, the shape of an object is strongly tied to the object functionality and identity [30]. Furthermore, the use of holistic approach is shown to be very effective in lexicon reduction [31], moreover to increase the accuracy of classification.

3.1 United Moment Invariant Function

Moment Function has been used in diverse fields ranging from mechanics and statistics to pattern recognition and image understanding [32]. The use of moments in image analysis and pattern recognition was inspired by [33] and [34]. [33] first presented a set of seven-tuplet moments that invariant to position, size, and orientation of the image shape. However, there are many research have been done to prove that there were some drawback in the original work by [33] in terms of invariant such as [35], [36], [37], [38], [39], and [40]. All of these researchers proposed their method of moment and tested on feature extraction phase to represents the image. A good shape descriptor should be able to find perceptually similar shape where it is usually means rotated, translated, scaled and affined transformed shapes. Furthermore, it can tolerate with human beings in comparing the image shapes. Therefore, [41] derived United Moment Invariants (UMI) based on basic scaling transformation by [33] that can be applied in all conditions with a good set of discriminate shapes features. Moreover, UMI never been tested in WI domain. With the capability of UMI as a good description of image shape, this work is explored its capability of image representation in WI domain.

[41] proposed UMI with mathematically related to GMI by [33] by considering (1) as normalized central moments:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\frac{p+q+2}{2}}}, \quad p+q = 2,3,\ldots. \tag{1}$$

and (2) in discrete form. Central and normalized central moments are given as:

$$\mu'_{pq} = \rho^{p+q}\mu_{pq}, \qquad \eta'_{pq} = \rho^{p+q}\eta_{pq} = \frac{\rho^{p+q}}{\mu_{00}^{\frac{p+q+2}{2}}}\mu_{pq}. \tag{2}$$

and improved moment invariant by [43] is given as:

$$\eta'_{pq} = \frac{\mu_{pq}}{\mu_{00}^{p+q+1}}. \tag{3}$$

(1) to (3) have the factor μ_{pq}. Eight feature vector derived by [40] are listed below:

$$\theta_1 = \frac{\sqrt{\phi_2}}{\phi_1}. \quad \theta_2 = \frac{\phi_6}{\phi_1\phi_4}.$$
$$\theta_3 = \frac{\sqrt{\phi_5}}{\phi_4}. \quad \theta_4 = \frac{\phi_5}{\phi_3\phi_4}.$$
$$\theta_5 = \frac{\phi_1\phi_6}{\phi_2\phi_3}. \quad \theta_6 = \left(\phi_1 + \sqrt{\phi_2}\right)\frac{\phi_3}{\phi_6}.$$
$$\theta_7 = \frac{\phi_1\phi_5}{\phi_3\phi_6}. \quad \theta_8 = \frac{\phi_3+\phi_4}{\sqrt{\phi_5}}. \tag{4}$$

where ϕ_i are Hu's moment invariants.

4 Feature Selection

Feature selection has become an active research area for decades, and has been proven in both theory and practice [44]. The main objective of feature selection is to select the minimally sized subset of features as long as the classification accuracy does not significantly decreased and the result of the selected features class distribution is as close as possible to original class distribution [1]. In contrast to other dimensionality reduction methods like those based on projection or compression, feature selection methods do not alter the original representation of the variables, but merely select a subset of them. Thus, they preserve the original semantics of the variables. However, the advantages of feature selection methods come at a certain price, as the search for a subset of relevant features introduces an additional layer of complexity in the modeling task [2]. In this work, feature selection is explored in order to find the most significant features which by is the unique features of individual's writing. The unique features a mainly contribute to the concept of Authorship Invarianceness in WI.

There are three general methods of feature selection which are filter method, wrapper method, and embedded method [45]. Filter method assesses the relevance of features by looking only at the intrinsic properties of the data. A feature relevance score is calculated, and low-scoring features are removed [3]. Simultaneously, wrapper method uses an induction algorithm to estimate the merit of feature subsets. It explores the space of features subsets to optimize the induction algorithm that uses the subset for classification [4]. On the other hand, in embedded method, the selection process is done inside the induction algorithm itself, being far less computationally intensive compared with wrapper methods [5]. However, the focus of this paper is to explore the use of wrapper methods. Wrapper strategies for feature selection use an induction algorithm to estimate the merit of feature subsets. The rationale for wrapper methods is that the induction method that will ultimately use the feature subset should provide a better estimate of accuracy than a separate measure that has an entirely different inductive bias [3].

The wrapper method is computationally demanding, but often is more accurate. A wrapper algorithm explores the space of features subsets to optimize the induction algorithm that uses the subset for classification. These methods based on penalization face a combinatorial challenge when the set of variables has no specific order and when the search must be done over its subsets since many problems related to feature extraction have been shown to be NP-hard [4]. Advantages of wrapper methods include the interaction between feature subset search and model selection, and the ability to take into account feature dependencies. A common drawback of these methods is that they have a higher risk of over-fitting than filter methods and are very computationally intensive, especially if building the classifier has a high computational cost [2]. There are several wrapper techniques, however only two techniques will be discussed here. These techniques are Sequential Forward Selection and Sequential Forward Floating Selection. Figure 2 depicts wrapper feature selection method.

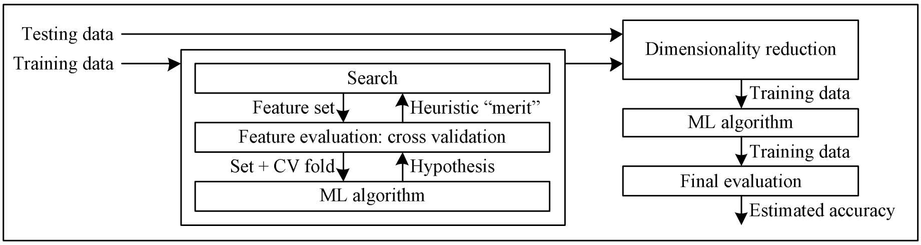


Fig. 2. Wrapper Feature Selection [3]

Sequential Forward Selection (SFS) is introduced by [46] which proposed the best subset of features Y_0 that is initialized as the empty set. The feature x^+ that gives the highest correct classification rate $J(Y_k + x^+)$ is added to Y_k at the each step along with the features which already included in Y_k. The process continues until the correct classification rate given by Y_k and each of the features not yet selected does not increase. SFS performs best when the optimal subset has a small number of features. When the search is near the empty set, a large number of states can be potentially evaluated, and towards the full set, the region examined by SFS is narrower since most of the features have already been selected. The algorithm of SFS is shown as below:

```
1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \text{argmax}_{x^+ \notin Y_k}[J(Y_k + x^+)]$
3. Update $Y_{k+1} = Y_k + x^+; k = k + 1$
4. Go to step 2
```

However, this method suffers from the nesting effect. This means that a feature that is included in some step of the iterative process cannot be excluded in a later step. Thus, the results are sub-optimal. Therefore, the Sequential Forward Floating Selection (SFFS) method was introduced by [47] to deal with the nesting problem. In SFFS, Y_0 is initialized as the empty set and in each step a new subset is generated first by adding a feature x^+, but after that features x^- is searched for to be eliminated from Y_k until the correct classification rate $J(Y_k - x^-)$ decreases. The iterations continue until no new variable can be added because the recognition rate $J(Y_k + x^+)$ does not increase. The algorithm is as below.

```
1. Start with the empty set $Y_0 = \{\emptyset\}$
2. Select the next best feature $x^+ = \text{argmax}_{x^+ \notin Y_k}[J(Y_k + x^+)]$
3. Update $Y_{k+1} = Y_k + x^+; k = k + 1$
4. Remove the worst feature $x^- = \text{argmax}_{x^- \in Y_k}[J(Y_k - x^-)]$
5. If $J(Y_k - x^-) > J(Y_k)$
     Update $Y_{k+1} = Y_k - x^-; k = k + 1$
     Go to 3
   Else
     Go to 2
```

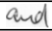
5 Proposed Approach

The experiments described in this paper are executed using the IAM database [48]. Various types of word images from IAM database are extracted using UMI to represent the image into feature vector. The selection of significant features using the wrapper methods are performed prior the identification task. The selected features which produce highest accuracy from the identification task are identified as the optimal significant features for WI in this work and also known as unique features of individual's writing.

5.1 Extracting Features

Feature extraction is a process of converting input object into feature vectors. The extracted features are in real value and unique for each word. A set of moments computed from digital image using UMI represents global characteristics of an image shape, and provides a lot of information about different types of geometrical features of the image [49]. Different types of words from IAM database such as 'the', 'and', 'where' and others have been extracted from one author. In this paper, a total number of 4400 instances are extracted to be used for the experiments, and are randomly divided into five different datasets to form training and testing dataset. Table 1 is the example of feature invariant of words using UMI with eight features vector for the each image.

Table 1. Example of Feature Invariant

Image	Feature 1	………	Feature 7	Feature 8
	0.217716	………..	1.56976	1.82758
	0.115774	………..	0.0552545	0.499824
	0.369492	…………	0.124305	0.580407

Extracted features can be divided into micro and macro feature classes which are local and global features. Local features denote the constituent parts of objects and the relationships, meanwhile global features describing properties of the whole object [50]. Good features are those satisfying two requirements which are small intra-class invariance and large inter-class invariance [51]. This can be defined as invarianceness of authorship in WI.

Invarianceness of authorship in WI shows the similarity error for intra-class (same-writer) is small compared to inter-class (different-writers) for the same words or different words. This is due to the individual features of handwriting's style which has been proof in many researchers such as [18], [24], and [52]. Related to this paper, the objective is to make contributions towards this scientific validation using the proposed techniques for selecting the significant features in order to proof the authorship of invarianceness in WI. The uniqueness of this work is to find the significant feature which actually is the unique features of individual's writing. The invarianceness of authorship relates to individuality of handwriting with the unique features of individual's writing. The highest accuracy of selected features proofs the invarianceness of authorship for intra-class is lower than inter-class where each individual's writing contains the unique styles of handwriting that is different with other individual. To achieve this, the process of selecting significant features is carried out using the proposed wrapper method before identification task.

5.2 Selecting Significant Features

Two commonly used wrapper method discussed earlier will be used to determine the significant features. These feature selection techniques will be using Modified Immune Classifier (MIC) [53] as their classifier. Every experiment has been performed using ten-fold cross-validation. These feature selection techniques will be executed five times to ensure the performance is stable and accurate.

In order to justify the quality of feature subset produced by each method, other state-of-the-art feature selection techniques are also used, which are Correlation-based Feature Selection (CFS) [3], Consistency-based Feature Selection, also known as Las Vegas Filter (LVF) [54], and Fast Correlation-based Filter (FCBF) [55]. Other classifiers are also being used for SFS to further validate the result, which are Naïve Bayes [56] and Random Forest [57] classifier. These feature selection techniques are provided in WEKA [58]. Justification of these feature selection techniques has been presented in [22]. Table 2 is the result of selection for each feature invariant data set.

Table 2. Experimental Results on Feature Selection

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
SFS (using MIC)	Execution #1	f2, f3, f6, f8	f2, f3, f4, f6, f8	f1, f3, f6, f7, f8	f3, f6, f8	f1, f2, f3, f5, f6, f7	**f3, f6**
	Execution #2	f1, f3, f4, f6, f8	f1, f3, f4, f5, f6, f8	f1, f3, f4, f6, f8	f1, f2, f3, f6	f1, f3, f6	**f3, f6**
	Execution #3	f2, f3, f4, f5, f6, f8	f1, f3, f6, f7, f8	f1, f3, f6, f8	f2, f3, f6, f7, f8	f3, f4, f5, f6, f7, f8	**f3, f6, f8**
	Execution #4	f2, f3, f6, f8	f1, f3, f4, f5, f6, f8	f1, f2, f3, f4, f5, f6	f1, f3, f4, f5, f6	f1, f3, f6	**f3, f6**
	Execution #5	f3, f6, f7, f8	f1, f2, f3, f6	f2, f3, f4, f5, f6, f7, f8	f1, f3, f6, f8	f2, f3, f6, f8	**f3, f6**
	Intersection	**f3, f6, f8**	**f3, f6**	**f3, f6**	**f3, f6**	**f3, f6**	**f3, f6**
SFFS (using MIC)	Execution #1	f1, f3, f6	f2, f3, f4, f6	f1, f3, f4, f5, f6, f8	f1, f3, f6, f8	f3, f4, f6, f7, f8	**f3, f6**
	Execution #2	f1, f3, f5, f6, f7, f8	f3, f4, f6, f7, f8	f1, f2, f3, f4, f6, f8	f1, f3, f4, f5, f6, f8	f2, f3, f5, f6	**f3, f6**
	Execution #3	f2, f3, f4, f5, f6, f7, f8	f2, f3, f5, f6, f8	f1, f2, f3, f6, f7, f8	f2, f3, f6, f8	f2, f3, f6, f8	**f3, f6, f8**
	Execution #4	f3, f4, f6, f8	f1, f2, f3, f6	f3, f6, f7, f8	f3, f4, f6, f8	f3, f6, f8	**f3, f6**
	Execution #5	f2, f3, f4, f5, f6, f7	f1, f2, f3, f6, f7, f8	f2, f3, f4, f5, f6, f8	f1, f3, f6, f8	f3, f6, f7, f8	**f3, f6**
	Intersection	**f3, f6**	**f3, f6**	**f3, f6, f8**	**f3, f6, f8**	**f3, f6**	**f3, f6**

Table 2. (*continued*)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
CFS	Execution #1	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #2	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #3	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #4	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Execution #5	f1, f2, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	f1, f3, f4, f5, f6, f7	f1, f3, f5, f7, f8	f1, f3, f4, f5, f6, f7	**f1, f3, f5, f7**
	Intersection	**f1, f2, f3, f5, f7, f8**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f5, f7, f8**	**f1, f3, f4, f5, f6, f7**	**f1, f3, f5, f7**
LVF	Execution #1	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #2	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #3	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #4	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Execution #5	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	f2, f3, f4, f6	**f2, f3, f4, f6**
	Intersection	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**	**f2, f3, f4, f6**
FCBF	Execution #1	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #2	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #3	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #4	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Execution #5	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	f1, f2, f3, f4, f5, f6, f7, f8	**f1, f2, f3, f4, f5, f6, f7, f8**
	Intersection	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**	**f1, f2, f3, f4, f5, f6, f7, f8**
SFS (using Naïve Bayes)	Execution #1	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #2	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**

Table 2. (*continued*)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Intersection
	Execution #3	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #4	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Execution #5	f1, f2, f3, f5, f8	f3, f4	f3	f1, f3, f4	f3	**f3**
	Intersection	**f1, f2, f3, f5, f8**	**f3, f4**	**f3**	**f1, f3, f4**	**f3**	**f3**
SFS (using Random Forest)	Execution #1	f3	f3	f3	f3	f3	**f3**
	Execution #2	f3	f3	f3	f3	f3	**f3**
	Execution #3	f3	f3	f3	f3	f3	**f3**
	Execution #4	f3	f3	f3	f3	f3	**f3**
	Execution #5	f3	f3	f3	f3	f3	**f3**
	Intersection	**f3**	**f3**	**f3**	**f3**	**f3**	**f3**

Based on the feature selection results, it is shown that these feature selection techniques yield different subset with different size. It is shown that SFS with Naïve Bayes in set C and E and Random Forest in all set select only one feature. These two are not capable to reduce the number of features partially due to the nature of the data itself. It is also known prone to over-fitting to some datasets and cannot handle large numbers of irrelevant features, thus it is not capable to reduce the number of features,

On the other hand, FCBF is shown to unable reduce the number of features, this is because this feature selection technique is more suitable when handling high-dimensional data, because it analyze the correlation between features, which is feature relevancy and feature redundancy. Thus, these methods will perform poorly when they failed to find the correlation between features, or they overestimate the correlation between features. In other domain of pattern recognition, the results obtained from FCBF and SFS with Naïve Bayes and Random Forest can be considered as suboptimal result, however in this WI domain, these feature selection techniques is still considered to achieve the purpose of the experiment. This is because the purpose of feature selection in WI is not only to reduce the number of features; instead it is to determine the most significant features (unique features). Thus, FCBF considers all features are significant, while SFS with Naïve Bayes and Random Forest consider that the selected features are the most significant feature.

On the contrary, the rest of the techniques (SFS and SFFS with MIC, CFS, and LVF) are able to identify the significant features. It is also worth mentioning that although these feature selection techniques yield different features, they seem to always include the third feature (f3) in their results. Therefore, it can be concluded that the third feature (f3) is the most significant feature, and it is chosen as significant unique feature in order to proof the invarianceness of authorship in this work.

5.3 Identifying the Authorship Using Significant Features

The selected significant features from every feature selection techniques must be justified and validated through identification performance. In order to justify the quality of feature subset produced by each method, the feature subsets are tested against classification, which uses MIC as the classifier. Table 3 is the result of identification accuracy for each feature subset.

Table 3. Experimental Results on Identification Accuracy (%)

Method	Execution	Set A	Set B	Set C	Set D	Set E	Average
SFS (using MIC)	Execution #1	97.40	97.18	96.92	96.14	96.94	**96.92**
	Execution #2	97.29	97.77	96.01	96.47	95.80	**96.67**
	Execution #3	97.63	97.30	95.78	96.80	97.05	**96.91**
	Execution #4	97.40	97.77	97.26	96.80	95.80	**97.01**
	Execution #5	97.51	96.59	97.38	96.14	96.49	**96.82**
	Average	**97.45**	**97.32**	**96.67**	**96.47**	**96.42**	**96.87**
SFFS (using MIC)	Execution #1	96.95	96.71	97.04	96.14	96.49	**96.66**
	Execution #2	97.40	97.18	96.58	97.13	96.94	**97.05**
	Execution #3	94.35	97.41	97.04	96.03	96.49	**96.26**
	Execution #4	97.06	96.59	96.58	96.14	96.03	**96.48**
	Execution #5	97.51	97.18	97.04	96.14	96.60	**96.89**
	Average	**96.66**	**97.02**	**96.85**	**96.32**	**96.51**	**96.67**
CFS	Execution #1	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #2	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #3	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #4	94.24	97.18	97.18	94.01	97.18	**95.95**
	Execution #5	94.24	97.18	97.18	94.01	97.18	**95.95**
	Average	**94.24**	**97.18**	**97.18**	**94.01**	**97.18**	**95.95**
LVF	Execution #1	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #2	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #3	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #4	97.40	97.40	97.40	97.40	97.40	**97.40**
	Execution #5	97.40	97.40	97.40	97.40	97.40	**97.40**
	Average	**97.40**	**97.40**	**97.40**	**97.40**	**97.40**	**97.40**
FCBF	Execution #1	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #2	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #3	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #4	97.74	97.74	97.74	97.74	97.74	**97.74**
	Execution #5	97.74	97.74	97.74	97.74	97.74	**97.74**
	Average	**97.74**	**97.74**	**97.74**	**97.74**	**97.74**	**97.74**
SFS (using Naïve Bayes)	Execution #1	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #2	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #3	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #4	93.79	88.47	80.23	92.09	80.23	**86.96**
	Execution #5	93.79	88.47	80.23	92.09	80.23	**86.96**
	Average	**93.79**	**88.47**	**80.23**	**92.09**	**80.23**	**86.96**
SFS (using Random Forest)	Execution #1	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #2	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #3	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #4	80.23	80.23	80.23	80.23	80.23	**80.23**
	Execution #5	80.23	80.23	80.23	80.23	80.23	**80.23**
	Average	**80.23**	**80.23**	**80.23**	**80.23**	**80.23**	**80.23**

Based on the results, the accuracy is at its highest when the number of features is between 4-7 features. It is shown that FCBF produces the best accuracy (97.74%) and equal with the original dataset performance (97.74%). However, the number of features produced by FCBF is equal with the actual set (8 features). Meaning that, FCBF needs all features to produce the best performance. The second best accuracy is produced by LVF (97.40%). The results of LVF are shown to be stable, regardless of dataset and the number of execution. This is because the nature of the data that is consistent allows LVF to perform well.

On the other hand, both SFS with MIC (96.87%) and SFFS with MIC (96.67%) with lower number of features still can obtain almost similar performance, although it is slightly lower than original dataset (97.74%). These feature selection technique outperform some other techniques (CFS, SFS using Naïve Bayes, and SFS using Random Forest). This is due to the behavior of these techniques which can specifically identify the unique features in dataset, therefore it is resulting the highest performance. Besides that, the wrapper technique is able to recognize importance of each feature in every iteration. However, due to the nature of both Naïve Bayes and Random Forest, the performance of SFS is deteriorating (86.96% and 80.23%).

These techniques are both capable to identify the most significant features and at the same time they validate the invarianceness of authorship concept where the invariance between features for intra-class is lower than inter-class. As a normal practice in pattern recognition, it can be achieved by calculating the invariance for intra-class and inter-class using Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - r_i| \ . \quad (5)$$

The result in Table 4 shows that the invarianceness of authorship is proven where the invarianceness between features using selected features for intra-class (same author) is smaller compared to inter-class (different author). This conforms the significant features is relate to invarianceness of authorship on WI.

Table 4. Identification Accuracy Results (%)

Various words	1 writer	10 writers	20 writers
20 words	0.278666	0.295112	0.524758
40 words	0.289052	0.295236	0.512279
60 words	0.282408	0.293509	0.527289
80 words	0.270236	0.3018	0.520221
100 words	0.281886	0.355219	0.544051

It is also shown that CFS is also capable to obtain good result (95.95%), although it is not as good as LVF, SFS and SFFS with MIC. Although FCBF is the enhancement of CFS, it is shown that CFS is still better than FCBF in some dataset. This is because FCBF determines the correlation between features faster than CFS, which may causing the technique to overestimate the correlation between features, thus causing it to select all the features.

6 Conclusion and Future Work

The exploration of significant unique features relates to authorship invarianceness has been presented in this paper. A scientific validation has been provided as evidence of significant features can be used to proof the authorship invarianceness in WI. In future works, the selected unique features will be further explored with other classifier to confirm these features can be used as optimized features with higher accuracy. An improved sequential forward selection will also be developed to better adapt the nature of the data, and thus increase the performance.

References

1. Dash, M., Liu, H.: Feature Selection for Classification. J. Intelligent Data Analysis 1, 131–156 (1997)
2. Saeys, Y., Inza, I., Larranaga, P.: A Review of Feature Selection Techniques in Bioinformatics. J. Bioinformatics 23, 2507–2517 (2007)
3. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis, University of Waikato (1999)
4. Gadat, S., Younes, L.: A Stochastic Algorithm for Feature Selection in Pattern Recognition. J. Machine Learning Research 8, 509–547 (2007)
5. Portinale, L., Saitta, L.: Feature Selection: State of the Art. In: Feature Selection, pp. 1–22. Universita del Piemonte Orientale, Alessandria (2002)
6. Refaeilzadeh, P., Tang, L., Liu, H.: On Comparison of Feature Selection Algorithms. In: Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II, pp. 34–39. AAAI Press, Vancouver (2007)
7. Schlapbach, A., Bunke, H.: Off-line Handwriting Identification Using HMM Based Recognizers. In: Proc. 17th Int. Conf. on Pattern Recognition, pp. 654–658. IEEE Press, Washington (2004)
8. Bensefia, A., Nosary, A., Paquet, T., Heutte, L.: Writer Identification by Writer's Invariants. In: Eighth Intl. Workshop on Frontiers in Handwriting Recognition, pp. 274–279. IEEE Press, Washington (2002)
9. Shen, C., Ruan, X.-G., Mao, T.-L.: Writer Identification Using Gabor Wavelet. In: Proceedings of the 4th World Congress on Intelligent Control and Automation, vol. 3, pp. 2061–2064. IEEE Press, Washington (2002)
10. Srihari, S.N., Cha, S.-H., Lee, S.: Establishing Handwriting Individuality Using Pattern Recognition Techniques. In: Sixth Intl. Conference on Document Analysis and Recognition, pp. 1195–1204. IEEE Press, Washington (2001)
11. Said, H.E.S., Tan, T.N., Baker, K.D.: Writer Identification Based on Handwriting. Pattern Recognition 33, 149–160 (2000)
12. Bensefia, A., Paquet, T., Heutte, L.: A Writer Identification and Verification System. Pattern Recognition Letters 26, 2080–2092 (2005)
13. Yu, K., Wang, Y., Tan, T.: Writer Identification Using Dynamic Features. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 512–518. Springer, Heidelberg (2004)
14. Tapiador, M., Sigüenza, J.A.: Writer Identification Method Based on Forensic Knowledge. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 555–561. Springer, Heidelberg (2004)
15. He, Z.Y., Tang, Y.Y.: Chinese Handwriting-based Writer Identification by Texture Analysis. In: Proceedings of 2004 Intl. Conference on Machine Learning and Cybernetics, vol. 6, pp. 3488–3491. IEEE Press, Washington (2004)

16. Wirotius, M., Seropian, A., Vincent, N.: Writer Identification From Gray Level Distribution. In: Seventh Intl. Conference on Document Analysis and Recognition, pp. 1168–1172. IEEE Press, Washington (2003)
17. Marti, U.-V., Messerli, R., Bunke, H.: Writer Identification Using Text Line Based Features. In: Sixth Intl. Conference on Document Analysis and Recognition, pp. 101–105. IEEE Press, Washington (2001)
18. Bin, Z., Srihari, S.N.: Analysis of Handwriting Individuality Using Word Features. In: Seventh Intl. Conference on Document Analysis and Recognition, pp. 1142–1146. IEEE Press, Washington (2003)
19. Zois, E.N., Anastassopoulos, V.: Morphological Waveform Coding for Writer Identification. Pattern Recognition 33, 385–398 (2000)
20. Kudo, M., Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. Int J. Pattern Recognition 33, 25–41 (2000)
21. Schlapbach, A., Kilchherr, V., Bunke, H.: Improving Writer Identification by Means of Feature Selection and Extraction. In: Eight Intl. Conference on Document Analysis and Recognition, pp. 131–135. IEEE Press, Washington (2005)
22. Pratama, S.F., Muda, A.K., Choo, Y.-H.: Feature Selection Methods for Writer Identification: A Comparative Study. In: Proceedings of 2010 Intl. Conference on Computer and Computational Intelligence, pp. 234–239. IEEE Press, Washington (2010)
23. Srihari, S.N., Huang, C., Srinivasan, H., Shah, V.A.: Biometric and Forensic Aspects of Digital Document Processing. In: Chaudhuri, B.B. (ed.) Digital Document Processing, pp. 379–405. Springer, Heidelberg (2006)
24. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of Handwriting. J. Forensic Sciences 47, 1–17 (2002)
25. Zhu, Y., Tan, T., Wang, Y.: Biometric Personal Identification Based on Handwriting. In: Intl. Conference on Pattern Recognition, vol. 2, pp. 797–800. IEEE Press, Washington (2000)
26. Cajote, R.D., Guevara, R.C.L.: Global Word Shape Processing Using Polar-radii Graphs for Offline Handwriting Recognition. In: TENCON 2004 IEEE Region 10 Conference, vol. A, pp. 315–318. IEEE Press, Washington (2004)
27. Parisse, C.: Global Word Shape Processing in Off-line Recognition of Handwriting. IEEE Trans. on Pattern Analysis and Machine Intelligence 18, 460–464 (1996)
28. Madvanath, S., Govindaraju, V.: The Role of Holistic Paradigms in Handwritten Word Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 23, 149–164 (2001)
29. Steinherz, T., Rivlin, E., Intrator, N.: Offline Cursive Script Word Recognition – ASurvey. Intl. Journal on Document Analysis and Recognitio 2, 90–110 (1999)
30. Cheikh, F.A.: MUVIS: A System for Content-Based Image Retrieval. PhD Thesis, TampereUniversity of Technology (2004)
31. Vinciarelli, A.: A Survey on Off-line Cursive Word Recognition. Pattern Recognition 35(7), 1433–1446 (2002)
32. Liao, S.X.: Image Analysis by Moment. PhD Thesis, University of Manitoba (1993)
33. Hu, M.K.: Visual Pattern Recognition by Moment Invariants. IRE Transaction on Information Theory 8, 179–187 (1962)
34. Alt, F.L.: Digital Pattern Recognition by Moments. J. the ACM 9, 240–258 (1962)
35. Reiss, T.H.: The Revised Fundamental Theorem of Moment Invariants. IEEE Trans. on Pattern Analysis and Machine Intelligence 13, 830–834 (1991)
36. Belkasim, S.O., Shridhar, M., Ahmadi, M.: Pattern Recognition with Moment Invariants: A Comparative Study and New Results. Pattern Recognition 24, 1117–1138 (1991)
37. Pan, F., Keane, M.: A New Set of Moment Invariants for Handwritten Numeral Recognition. In: IEEE Intl. Conference on Image Processing, vol. 1, pp. 154–158. IEEE Press, Washington (1994)

38. Sivaramakrishna, R., Shashidhar, N.S.: Hu's Moment Invariant: How Invariant Are They under Skew and Perspective Transformations? In: Conference on Communications, Power and Computing, pp. 292–295. IEEE Press, Washington (1997)
39. Palaniappan, R., Raveendran, P., Omatu, S.: New Invariant Moments for Non-Uniformly Scaled Images. Pattern Analysis and Applications 3, 78–87 (2000)
40. Shamsuddin, S.M., Darus, M., Sulaiman, M.N.: Invarianceness of Higher Order Centralised Scaled-Invariants on Unconstrained Handwritten Digits. Intl.J. Inst. Maths. and Comp. Sciences 12, 1–9 (2001)
41. Yinan, S., Weijun, L., Yuechao, W.: United Moment Invariant for Shape Discrimination. In: IEEE Intl. Conference on Robotics, Intelligent Systems and Signal Processing, pp. 88–93. IEEE Press, Washington (2003)
42. Zhang, D.S., Lu, G.: Review of Shape Representation and Description Techniques. Pattern Recognition 37, 1–19 (2004)
43. Chen, C.-C.: Improved Moment Invariants for Shape Discrimination. Pattern Recognition 26, 683–686 (1993)
44. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. J. Machine Learning Research, 1205–1224 (2004)
45. Geng, X., Liu, T.-Y., Qin, T., Li, H.: Feature Selection for Ranking. In: 30th Annual Intl. ACM SIGIR Conference, pp. 407–414. ACM Press, Amsterdam (2007)
46. Whitney, A.W.: A Direct Method of Nonparametric Measurement Selection. IEEE Trans. in Computational, 1100–1103 (1971)
47. Pudil, P., Novovicova, J., Kittler, J.: Floating Search Methods in Feature Selection. Pattern Recognition Letter 15, 1119–1125 (1994)
48. Marti, U.-V., Bunke, H.: The IAM Database: An English Sentence Database for Off-line Handwriting Recognition. J. Document Analysis and Recognition 5, 39–46 (2002)
49. Balthrop, J., Forrest, S., Glickman, M.R.: Coverage and Generalization in An Artificial Immune System. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 3–10. Morgan Kaufmann, San Francisco (2002)
50. Palhang, M., Sowmya, A.: Feature Extraction: Issues, New Features, and Symbolic Representation. In: Huijsmans, D.P., Smeulders, A.W.M. (eds.) VISUAL 1999. LNCS, vol. 1614, pp. 418–427. Springer, Heidelberg (1999)
51. Khotanzad, A., Lu, J.H.: Classification of Invariant Image Representations Using a Neural Network. IEEE Trans. on Acoustics, Speech and Signal Processing 38, 1028–1038 (1990)
52. Liu, C.-L., Dai, R.-W., Liu, Y.-J.: Extracting Individual Features from Moments for Chinese Writer Identification. In: Proceedings of the Third Intl. Conference on Document Analysis and Recognition, vol. 1, pp. 438–441. IEEE Press, Washington (1995)
53. Muda, A.K.: Authorship Invarianceness for Writer Identification Using Invariant Discretization and Modified Immune Classifier. PhD Thesis, Universiti Teknologi Malaysia (2009)
54. Liu, H., Setiono, R.: A Probabilistic Approach to Feature Selection - A Filter Solution. In: Intl. Conference of Machine Learning, pp. 319–337. Morgan Kaufmann, Bari (1996)
55. Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: Proceedings of the Twentieth Intl. Conference on Machine Learning, pp. 856–863. ICM Press, Washington (2003)
56. Rish, I.: An Empirical Study of the Naive Bayes Classifier. In: Intl. Joint Conference on Artificial Intelligence, pp. 41–46. AAAI Press, Vancouver (2001)
57. Breiman, L.: Random Forests. J. Machine Learning, 5–32 (2001)
58. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. J. SIGKDD Explorations 11, 10–18 (2009)

Pixel Mask Analyzer (PMA): Noise Removal in Gland Detection for Colon Cancer Image

Mohd Yamin Ahmad[1], Yasmin Anum Mohd Yusof[2],
Siti Aishah Md. Ali[3], and Azlinah Mohamed[1]

[1] Faculty of Computer and Mathematical Sciences
University Technology MARA, 40450 Shah Alam Selangor Malaysia
[2] Department of Biochemistry, Faculty of Medicine, Universiti Kebangsaan Malaysia Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia
[3] Department of Pathology, UKM Medical Centre, Jalan Yaacob Latif, Bandar Tun Razak, Cheras, 56000 Kuala Lumpur,Malaysia

Abstract. It is well known that for any type of cancer, mortality rate can be reduced by early detection of the cancer. Detection of cancer markers from the time of taking blood from patients until microscopic examination of the biopsy tissues is not only laborious but time consuming. It is known that with the help of a computerized system, the diagnosis time can be shortened. In this paper, we proposed the method of gland edge detection and noise removal using Pixel Mask Analyzer (PMA) as image enhancement. PMA is based on kFill filter that uses 5x5 mask slide through entire 2D biopsies tissue images. Pixels within mask are analyzed on 8 sample images to remove noise. We compare the median filter and PMA and found that PMA gives better result on noise removal. After noise removal, Sobel and Canny edge detectors are applied on the images and we found that Canny provides better output for edge detection. The output of this detection is important for further analysis of colon cancer cells image detection.

Keywords: Colorectal Cancer, Gland Detection, Adenocarcinoma, Image Processing, Noise Removal, Pattern Recognition, Computer Aided Diagnosis.

1 Introduction

Colon Cancer (CRC) also known as colorectal cancer is the 3rd common cancer affecting men and woman worldwide [1]. Adenocarcinoma is the major type of CRC and its occurrence is more than 98% [2]. Mortality rate depends on its early detection and how fast treatment is given. In the early stage, most of the cancers including CRC are curable with surgery by removing the precancerous polyps or affected area.

The reason for the longer time taken for diagnosing CRC is the period of lab tests performed to confirm the presence of tumour cells. It can consume more than a week for confirming the CRC result and this contributes to further development of the cancer. Pathologists need to scrutinize a large number of microscope slides containing biopsy tissues manually which is a laborious process. Under the microscope,

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 652–659, 2011.

pathologists will identify the gland feature, looking for the abnormal cells and measuring the size of cell's nucleus. The shape and irregular hyper chromatic nuclei of the tissues will also be identified. We can obviously see that the procedures are time consuming, difficult and tedious which will lead to physical and visual fatigue of the pathologist.

Mortality rate from CRC has slightly decreased in the past two decades due to the result of early detection and effective modern treatment [2] such as Computer Aided Diagnosis (CAD). CAD is a procedure that assists doctors in scanning and diagnosing the medical images such as X-Ray and Ultrasound Images. However the process of CRC detection still requires manual analysis from expert pathologists. The unavailability of the experts may result in late detection of CRC.

In processing the image, many researchers have focused on detecting object from images. Tan et al. [3] proposed a solution to recover the content on the front side of the page from the interfering or noise images caused by the handwriting on the reverse side. The method used improved Canny edge detector with norm-orientation similarity constrains. Mohd Rahim et al. [4] proposed Hybrid Technique for edge detection which can detect river from aerial photo. This technique combines thresholding technique and Sobel edge detector to gain better result in detecting the river with higher clarity. In detecting colon gland feature, some of the details outside the gland will be removed. It is related to proposed method of removing salt and pepper noise using kFill base algorithm to remove noise outside the drawing line [5][6].

Prior to detection of CRC image, it is important to detect the gland. In this paper we focused on gland edge detection using PMA to remove noise with combination of Sobel and Canny edge detector. Section 2 of this paper addresses the problem, section 3 discusses the algorithm of gland edge detection, section 4 presents the simulation result and finally the conclusion is given in section.

2 Problem

A biopsy image may contain gland, mucosa, muscularis mucosae and other cells. The details inside the image that are not related to the detection criteria will produce noise. By applying Canny or Sobel edge detection directly to the image is insufficient to create acceptable output. Linking process is needed to connect edges to produce the desired output [4]. Even though Canny edge detector provides the optimal edge detector and can compromise between noise reduction, it is reported that canny is highly subjected to false edge detection [7].

A method known as median filter is usually used to remove noise. However it suffers the distortion of corners and thin line in the image. Meanwhile another possible method known as kFill filter can be used to remove noise in the image. kFill filter uses N x N mask slide through the entire image to remove noise. The size of N should be appropriately selected according to the known noise size. The noise removal process using kFill is by filling the core inside $(N\text{-}2)^2$ with opposite value. This method may produce shortening of one pixel wide graphical object from their end points [5]. In many cases, noises spots are frequently larger than one pixel and kFill algorithm will never fill the noise smaller than the core size [6]. Gland images in this study contain noises spots of different sizes and shapes produced by cells and other components inside and outside the gland. Therefore we proposed a modification of kFill filter to remove the noise.

Sometimes acquired images contain area that are not related to detection criteria and may consume processing time. Hence the images need to undergo enhancement process to eliminate unnecessary elements and reduce the work area only to the relevant region [8].

3 Algorithm

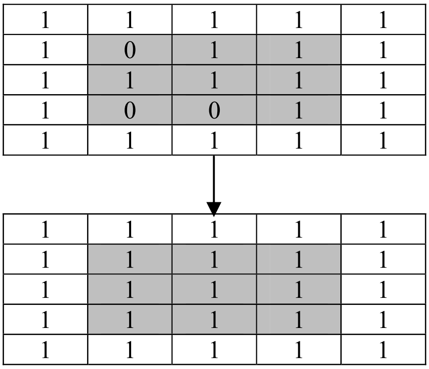
A set of full coloured colon gland images are obtained from a website [9]. The images are then cropped to the size of 720x540 pixels and saved as JPEG format. Using MATLAB, the images are converted into binary images with the midway between black and white of 0.5 [10].

After binarization process, some details in the image need to be removed. We proposed Pixel Mask Analyzer (PMA), a method based on kFill filter, which uses *N*x*N* pixel masking that slide through entire image to find and delete smaller pixel groups inside the mask. In this study we proposed the appropriate mask size *N* is equal to 5. The *c* variable is the core where total sum of white pixel in the core is equal to $(N\text{-}2)^2$ and variable *r* is the neighborhood pixels surrounding the core where *r* is equal to 4(*N*-1). All the contents inside the mask will be set to white which will convert the value of 0 to 1 when the following equation is met:

$$c < (N\text{-}2)^2 \ \wedge \ r{=}4(N\text{-}1) \tag{1}$$

Figure 1 illustrates the filling of core where the equation (1) is met. We then analysed the weak pixel groups from the neighbourhood of *N*x*N* centre pixel groups. The sum values of its neighbourhood $\mathrm{n_i}$ are calculated and if the sum of each individual neighbour is less than threshold value *T*, then the centre $\mathrm{n_c}$ is considered as noise. In this study *T* is set to $(1/3)N^2$ and $\mathrm{n_c}$ will be filled with white where the condition below is met:

$$\mathrm{n_1} < T \wedge \mathrm{n_2} < T \ldots \ \wedge \mathrm{n}_{4(N\text{-}1)} < T \tag{2}$$



1	1	1	1	1
1	0	1	1	1
1	1	1	1	1
1	0	0	1	1
1	1	1	1	1

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

Fig. 1. Removing smaller pixel Group

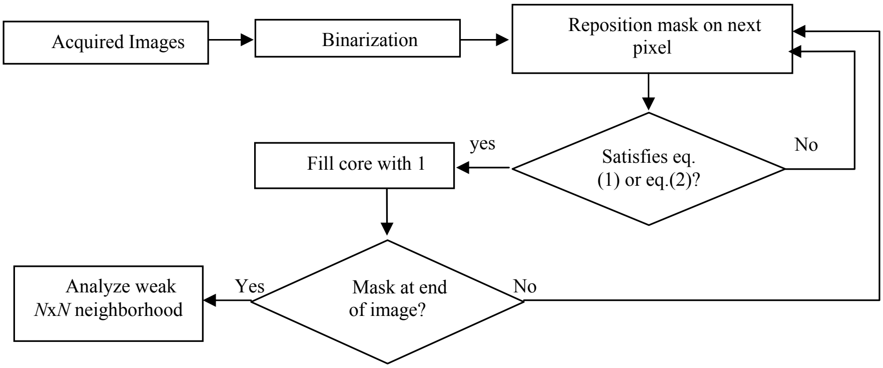

Acquired Images
Binarization
Reposition mask on next pixel
Satisfies eq. (1) or eq.(2)?
yes
No
Fill core with 1
Mask at end of image?
Yes
No
Analyze weak *N*x*N* neighborhood

Fig. 2. Process flow of proposed procedure

5x5 pixel (n_1)	5x5 pixel (n_2)	5x5 pixel (n_3)
5x5 pixel (n_8)	5x5 pixel (n_c)	5x5 pixel (n_4)
5x5 pixel (n_7)	5x5 pixel (n_6)	5x5 pixel (n_5)

Fig. 3. Neighbourhood pixel group

The process flow of proposed procedure is shown in Figure 2 while Figure 3 illustrates the filling of core where the equation (2) is met.

After the noise and details have been removed, a linking process is needed to connect the broken edges. We used bigger mask of 10x10 pixels that slide through the entire image. Pixel groups are analyzed in a clock-wise direction and broken edges are connected to the nearest group if the nearest group is less than 5 pixels.

									0
				0	0	0	0	0	0
			0	0	0	0	0	0	
			0	0	0	0	0	0	
			0	0				**0**	
			0					0	
		0						0	0
0	0	0					0	0	0
0	0						0	0	0
0	0							0	0

Fig. 4. Linking Process

After pixel group is connected through linking process, Canny and Sobel edge detectors are applied to the image. Areas that are completely covered by edge are then marked.

4 Simulation Result

The proposed algorithms are implemented using MATLAB performed using PC running on Windows 7 operating system. We tested on 8 biopsy images that consist of 2 normal images and 6 cancer images (adenocarcinoma). By applying binary conversion, we get the images as shown in Figure 5. Three complete glands in figure 5 are in oval shape where smaller dots inside and outside of the glands are considered as noise.

We tested the noise removal by applying median filter of 3x3 and 5x5 mask to the image. The result shows that bigger mask is better in dealing with noise removal as shown in Figure 6. However both images in Figure 6 suffered distortion at the corner of the gland. Bigger median filter mask also has the disadvantage of producing bigger noise and glands edges become wider. With the proposed noise removal algorithm we found that most of the noise spots outside and inside the glands are removed while still maintaining the gland structure as shown in Figure 7.

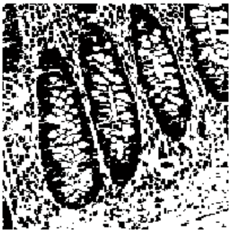
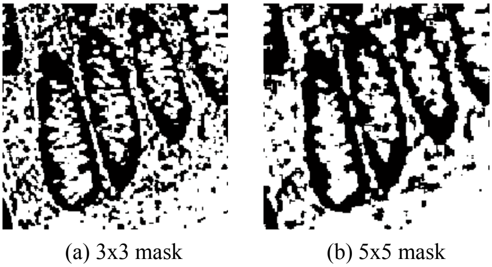


Fig. 5. Original binary image of gland



(a) 3x3 mask

(b) 5x5 mask

Fig. 6. Applying median filter

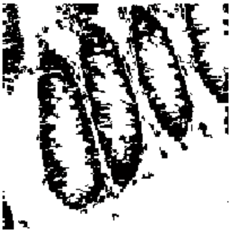
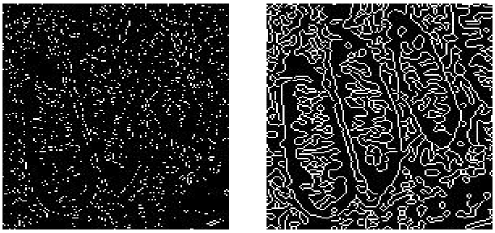


Fig. 7. Proposed image enhancement algorithm

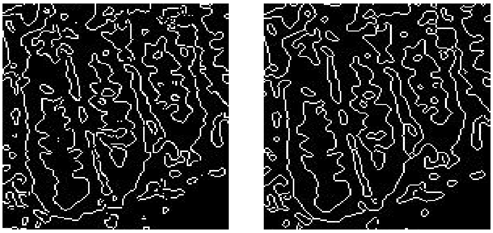
To detect the gland edge, Sobel and Canny algorithm are applied to the images. Figure 8 shows the image after applying Sobel and Canny edge detectors to the original binary image. With closer look, we found that almost no gland edges are detected with Sobel as shown in Figure 8(a). Canny offered almost perfect fine line surrounding the glands as shown in Figure 8(b). However without the noise removal, image is cluttered with too many details inside and outside the glands.



(a) Sobel (b) Canny

Fig. 8. Applying edge detector on original binary image

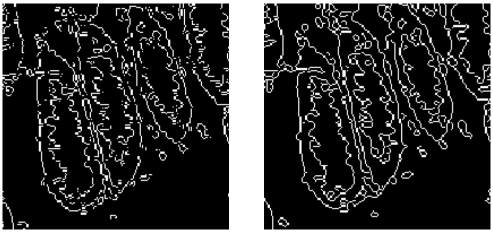
Figure 9 shows the image after applying Sobel and Canny edge detectors to the processed image with a 5x5 median filter. Both Figures 9(a) and 9(b) shows that this technique merges the gland edge and noise spots, which then alter the oval shape of the gland. This output should not be considered since it may result in false overall CRC detection.



(a) Sobel (b) Canny

Fig. 9. Applying edge detector on 5x5 median filter images

With the proposed PMA algorithm to remove noise, we found that gland shapes are maintained after applying Sobel and Canny edge detectors. However Sobel detector produces many discontinuous fine lines as shown in Figure 10(a) while Canny detector gives almost perfect gland edge detection as shown in Figure 10(b). Comparing the results from Figure 8 and 9, it is clear that the proposed algorithm has performed better in noise removal for gland detection of colon cancer images. With an exception that some noise spots still need to be removed, the result is more promising in general.



(a) Sobel (b) Canny

Fig. 10. Applying edge detector on proposed image enhancement algorithm

5 Conclusion

In this paper we demonstrate the new method of removing noise from colon gland image. Normally median filter can be used to remove noise. However the images suffer distortion on the round corner of the gland and the noise becomes larger. The proposed algorithm produces better noise removal while retaining the gland structure. Even though the noise is not completely removed, the result is more satisfactory in general.

In the future, we will conduct CRC image detection by looking at the gland feature. With the proposed algorithm, we can focus and limit the detection only to the gland area to improve overall CRC detection result.

Acknowledgments. We are indebted to Mr. Mohd Razif Shamsuddin for his technical support and help in carrying out this research.

References

1. Kerr, D.J., Young, M.A., Richard Hobbs, F.D.: ABC of colorectal cancer, pp. 1–15 (2001)
2. Meyerhardt, J., Saunders, M.: In: Skarin, A.T. (ed.) Colorectal Cancer, pp. 23–53 (2007)
3. Tan, C.L., Cao, R., Wang, Q., Shen, P.: Character extraction from interfering background - analysis of double-sided handwritten archival documents. In: Singh, S., Murshed, N., Kropatsch, W.G. (eds.) ICAPR 2001. LNCS, vol. 2013, p. 93. Springer, Heidelberg (2001)
4. Mohd Rahim, M.S., Nik Ismail, N.I., Shah Idris, M.A.: The Use of Hybrid Technique: Thresholding And Edge Detection For Identifying River From Aerial Photo. Journal Teknology 41(B) (2004)
5. Al-Khaffaf, H.S.M., Talib, A.Z., Salam, R.A.: Removing salt-and-pepper noise from binary images of engineering drawings. In: 19th International Conference on Pattern Recognition, Tampa, Florida, USA, vol. 1(6), pp. 1271–1274 (2008)
6. Chinnasarn, K., Rangsanseri, Y., Thitimajshima, P.: Removing salt-and-pepper noise in text/graphics images. In: The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, Chiangmai, pp. 459–462 (1998)
7. Jamil, N., Sembok, T.M.T.: Gradient-based edge detection of songket motifs. In: Sembok, T.M.T., Zaman, H.B., Chen, H., Urs, S.R., Myaeng, S.-H. (eds.) ICADL 2003. LNCS, vol. 2911, pp. 456–467. Springer, Heidelberg (2003)

8. Hern´andez-Cisneros, R.R., Terashima-Mar´ın, H.: Detection and Classification of Microcalcification Clusters in Mammograms Using Evolutionary Neural Networks, pp. 151–175 (2009)
9. Pathology Outline Website,
http://www.pathologyoutlines.com/colontumor.html
(visited at January 20, 2011)
10. Mathworks Inc.: Image Processing Toolbox^{TM} 6 User's Guide, pp. 17- 246 (2008)

A Review of Bio-inspired Algorithms as Image Processing Techniques

Noor Elaiza Abdul Khalid, Norharyati Md Ariff,
Saadiah Yahya, and Noorhayati Mohamed Noor

Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Tel.: +6012-5591341
sakurayati@yahoo.com

Abstract. This paper reviews 80 out of 130 bio-inspired Algorithm (BIAs) researches published in google scholar and IEEExplore between the periods of 1995 to 2010 used to solve image processing problems. BIAs has been successfully applied in many sciences, medical and engineering fields. The evolving, dynamic and meta-heuristics nature of BIAs makes it more robust, accurate and efficient in solving image processing problems. However finding the appropriate BIAs that matches the problem at hand is a tedious and difficult task. The BIAs investigated in this study are Genetic Algorithms, Evolutionary strategies, Genetic programming, memetic algorithms, swarm intelligence and artificial immune system. The publications are categorized by year of publication, by specific BIAs and by application. The statistics shows exponential increases in the application of BIAs to solve image processing problems and some algorithms have yet to be explored.

Keywords: Evolutionary Algorithm, Artificial Intelligent, Image Processing, Bioinspired Algorithm.

1 Introduction

The human eye and brain are biological systems that have the capability to adapt to changes, act sensors and have very reliable processor that are able to enhance, segment, register and recognize features of visions or images. Computer vision attempts to emulate these capabilities. Image enhancement, feature extraction, segmentation and registration plays a big role in analysing images in field such as medical[6][9][22][25][40], geosciences[28], remote sensing[1], facial extraction[19], biometrics[30][39] and many more.[42][47]. The main challenges in these type of application is to find the most suitable, accurate, faster and robust algorithm. As To date most algorithm have been design for customary use only thus there is a need to find a more adaptable and dynamic way.

Biological systems have natural capability to adapt to changes by learning, it's evolving, resiliant and robust. Metaphoring these system into algorithms introduced

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 660–673, 2011.

bio-inspired algorithms (BIAs) which has become popular in solving problems in virtually any area. Three main branches of BIAs are evolutionary algorithms (EAs), swarm intelligence (SI) and bacterial foraging algorithms (BFAs) which loosely bind the topics of connectionism, social behaviour and emergence based on the understanding of biological systems, genetic evolution, animal behaviours and bacterial foraging patterns. BIAs has drawn considerable research interests in the area of science and engineering of which in this case focus on the image processing area. However determining the suitable algorithm is a difficult task. The objective of this survey is to provide a better understanding of the application of bio-inspired algorithm as intelligent image technique.This paper overviews and compares three branches of Bio-inspired Algorithms (BIAs) applied in the area of image processing. Only four image processing area will be explored; image segmentation; feature extraction; image enhancement and image registration. Discussion includes the steps in each algorithm and their performance in terms of convergence, complexity, accuracy, speed and the optimum solutions.

2 Methodology

The source website of scientific paper related to BIAs applied in image processing are Google Scholar and IEEExplore. The reasons why these sources were chosen are, it provides simple way to broadly search for scholarly literature and helps to find relevant works across the world of scholarly research. Even though 130 papers between the period of 1995 to 2010 related to BIAs in various applications are found, only 80 papers appliy the various BIAs in the area of image processing. *Bio-inspired Algorithms (BIAs)* can be divided into three main type that is the evolutionary algorithms, swarm intelligence (SI) and bacteria foraging algorithm (BFAs).

Evolutionary Algorithm (EA) attempts to solve complex problems by mimicking Darwinian evolution where individuals in a population continuously compete with each other in the process of searching for optimal solutions [3]. As the history of the field suggests, there are many different variants of Evolutionary Algorithms (EAs). The common underlying idea behind all these techniques is the same: given a population of individuals, the environmental pressure causes natural selection (survival of the fittest) and this causes a rise in the fitness of the population. Given a quality function to be maximized, we can randomly create a set of candidate solutions. Based on this understanding, a family of EAs, known as the genetic algorithm (GA) [4] [8], evolutionary strategy (ES)[9], genetic programming (GP)[10], Selfish gene (SFGA)[11,12] and Memetic algorithm (MA) [13] have been developed. Members of the EA family share large number of common features and its population-based stochastic search algorithms perform best-to-survive criteria. Each algorithm commences by creating initial population of feasible solutions, and evolves iteratively from one generation to the next towards the best solution. Fitness-based selection takes place within the population. Diversity is introduced via mutations to uncover optimum solutions [8].

Selfish gene algorithm is a new member of EAs. It is focus on the fitness of the genes rather than the individuals. It does not have any crossover or mutation and its population store the genetic material which models the gene pool concept namely

“Virtual Population”. This population presents the number of individuals, and their specific identity represented by genome. Explicitly distinguished location in the genome is locus and the value is allele. The success of alleles is based on the frequency it appears in the virtual population. Generally evolution means, that organism which succeeds will increase its allele’s frequency at the expense of its children and on the other hand organism which fails will decreases its allele’s frequency [11,101]. The selfish gene algorithm has been successfully tested in several problems such as Automatic Test Pattern Generation for digital circuits [101], multiple knapsack problem [11] and optimization method [107].

Swarm Intelligence (SI) is designed based on collective behavior of decentralized, self-organized systems that occurs naturally or artificially. Particle swarm optimization (PSO) is a population based stochastic optimization technique invented by Eberhart and Kennedy in 1995, motivated by social behavior of fish schooling [86]. The system started with initializing random solution population called particles and searches for optimal solution by updating generations. These particles fly through the problem space following the current optimum particles. PSO have proven to be successfully applied in scientific and various other purposes [93].

In recent years, Bacterial Foraging Algorithms (BFAs) have emerged as another new branch of BIAs, which inherit the characteristics of bacterial foraging patterns such as chemo taxis, metabolism, reproduction and quorum sensing. This remarkable information processing biological system translates AIS [94]. AIS can be applied as Negative Selection Mechanism [94], and the clonal selection algorithm [91].

2.1 Bio-inspired Algorithm Matrix

Table 1 tabulates specifically nine algorithmic features in the Bio-inspired algorithm Matrix aligning the various BIAS algorithms.

Table 1. Bio-inspired Algorithm Matrix

Features	GA	SFGA	MA	ES	GP	PSO	AIS
Encoding	Binary string, char, vector	real number	real number	real number	Symbolic alphabet	Binary string, char, vector	Attribute strings
Population	Yes	Yes Virtual Pop.	Yes	Yes	Yes	Yes-using particles	Component concentration/ network
Selection	Random, Tournament, Roulette Wheel	Random	Random or using heuristic	Yes same as GA	Yes same as GA	No selection	Random
Crossover	Yes	No	Yes	Yes	Yes	No	No
Mutation	Yes	Yes	Yes	Yes	Yes	No	Yes
Fitness Function	Yes	Yes	Yes	Yes	Yes	Yes	Recognition /Object
Generation	Yes	No	Yes	Yes	Yes	No	Iteration
Locus	No	Yes	No	No	No	No	No
Alleles	No	Yes	No	No	No	No	No
Local search	No	No	Yes	No	No	No	No
No of steps	7	4	4	8	6	5	Many

Encoding is the stage where chromosomes are represented as binary strings, char and vectors in GA and PSO; list of real number in SFGA and MA; symbolic tree in

GP and attributes strings in AIS. BIAs parent selections stage can employ selection methods such as random, stochastic technique of roulette and tournament selection. Only MA uses heuristic search to select individuals from its population. Population size is very important for all BIAs algorithms variant, as limited population size produce low quality solutions [14]. The recombination or crossover stage is present in GA, MA, ES, GP and AIS. Among the types of crossover operation are single point, two point, uniform and arithmetic crossovers. Fitness function stage is important in all BIAs. It is a heuristic function that measures the performance of an individual chromosome. The fitness function establishes the basis for selecting chromosomes that will be mated during the reproduction in EAs and BFAs and best particles in PSO. The generation stage is used to repeat the process until it finds the most optimum values. Only SFGA and PSO do not include this stage.

Table 2 shows a comparison of five performance evaluation features of the bio-inspired algorithms that is the convergence rate, the algorithm complexity, the accuracy in finding solutions, the processing speed and the rate of achieving the optimal solutions.

Table 2. The performance evaluation matrix of the bio-inspired algorithms

Performance Features	**GA [98,100]**	**SFGA [11,12,101]**	**MA [13,99,104]**	**ES [9,102]**	**GP [10,103]**	**PSO [95,98]**	**AIS [91,94,106]**
Convergence	Difficult	Fast	Fast	difficult	difficult	Moderate	Difficult
complexity	Simple	Simple	simple	simple	difficult	Moderate	Difficult
Accuracy	Low	Reliable	reliable	moderate	low	Reliable	Low
Speed	Slow	Fast	Fast	slow	slow	Fast	Slow
Optimum sol.	Slow	Faster	faster	slow	slow	Fast	Slow

GA, ES, GP and AIS have difficulties in converging as the probability of making progress decreases rapidly as the minimum/maximum is approached. Thus, these algorithms are often hybridized with other techniques to improve their performance. [13] uses meta-heuristic population in GA and successfully resolved many optimization problems. However, premature convergence narrows down its ability to find many solutions. In the bid to reduce premature convergence possibility an algorithm that hybridized the classical GA with local search technique and named as Memetic [15], hybridized PSO and GA to counter the problems of early convergence in PSO and slow convergence in GA for global maximization.

The application of BIAs is becoming more and more popular in the area of image processing such as image segmentation, image enhancement, image restoration, image analysis, feature extraction, feature selection and face detection. Thus, this paper investigates the application of BIAs in the various area of image processing.

3 Results and Discussion

The results of the investigation are discussed in two main sections; the general statistical BIAs application and the detail of image processing area where each BIAs are applied.

3.1 Population of BIAs Applications between 1995-2010

Researches that apply BIAs in image processing began around the 1990s. The 130 BIAs publications used in this review are 1995 to 2010. Figure 1 shows the application of BIAs Algorithm by year. An obvious observation is that there is an increasing amount of research using BIAs. BIAs has become popular choice of algorithm in various applications.

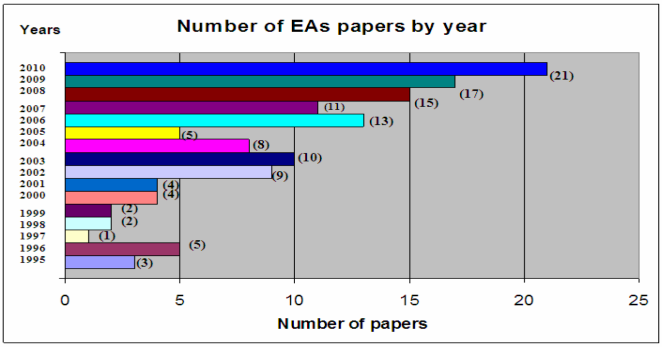


Fig. 1. BIAs publications by year

3.2 Statistics of BIAs in Image Processing

Out of 130 papers on BIAs algorithm found only 80 papers are applied in image processing. Most of the researches are centered in the area of image extraction (47%) and image segmentation (31%). Only a small number of researches are done in image enhancement (13%) and registration (9%). This can be seen in Figure 2.

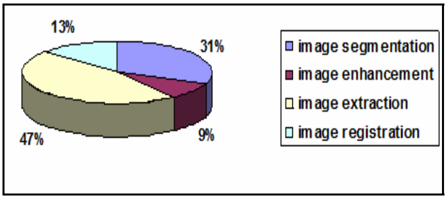


Fig. 2. A pie chart of statistical analysis of the Bio-inspired Algorithmn in Image processing techniques

The area of image extraction and feature extraction is extremely popular technique that used BIAs. The applications include the use of image extraction for remote sensing [16-17], development of robust active contour techniques which are suitable for the extraction of the head boundary [18], facial feature extraction will automatically extract features from various video images effectively [19] and memetic algorithm for intelligent feature extraction that will generate representation for isolated handwritten symbols [20].

Image segmentation comes second in popularity where most publications are either new application or invention related to image segmentation. It covers fields such as medical segmentation [21][9], texture segmentation [22][61][63], multi-objective segmentation [7], image color segmentation [44] and graph-based segmentation [96].

Most of the BIAs methods use in image registration is evolution strategies. It is useful for image registration because an invariant reference needs to be established within each source image. ES is capable of discovering transformations of larger scope [23]. Similar to application by Yuan X et al [9], ES is used in image registration via feature matching. The BIAs are enhanced by adding other features that can lower computational cost, high accuracy regardless of the magnitude of transformation required and reduce noise condition. In addition, various BIAs are used in 3D-feature based image registration [24][47]. Image registration has been applied to a broad range of situations from remote sensing to medical images or artificial vision [16-17], CAD systems [97] and other different techniques [24].

Finally BIAs increase robustness and efficiency in image enhancement [25] in applications such as automatic fingerprint identification system [26] and amplifying image contrast while removing noise [27]. A more detail view of BIAs applied in each image processing area is depicted as a spider diagram in Figure 3. The spider diagram is used as a visual technique [2] to model the various BIAs applied to image processing area.

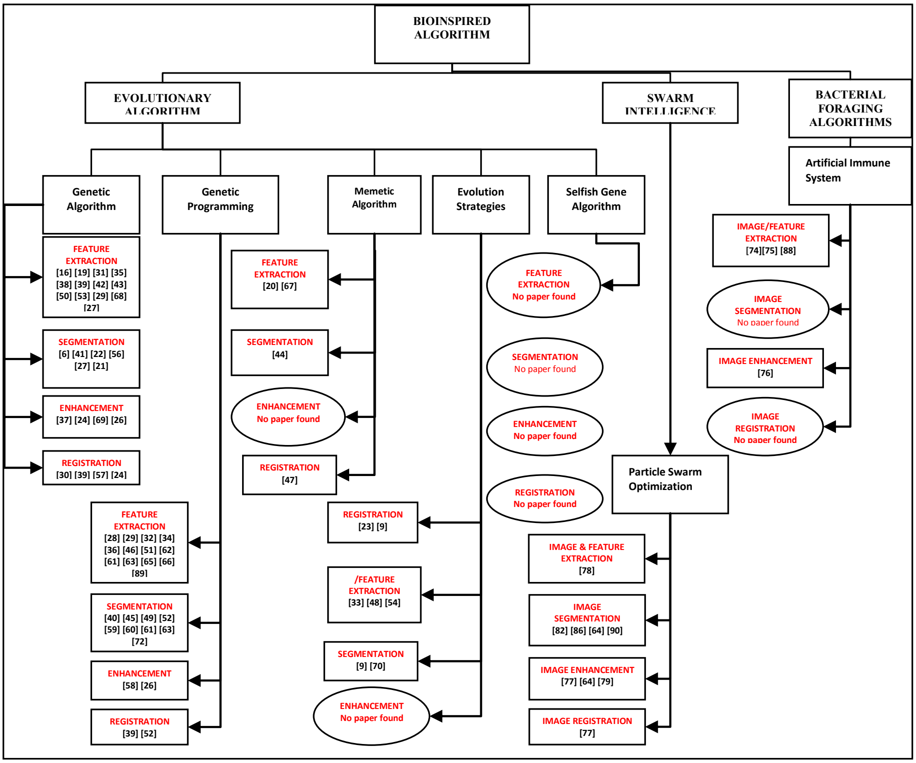

BIOINSPIRED ALGORITHM
EVOLUTIONARY ALGORITHM
SWARM INTELLIGENCE
BACTERIAL FORAGING ALGORITHMS
Artificial Immune System
Genetic Algorithm
Genetic Programming
Memetic Algorithm
Evolution Strategies
Selfish Gene Algorithm
FEATURE EXTRACTION [16] [19] [31] [35] [38] [39] [42] [43] [50] [53] [29] [68] [27]
SEGMENTATION [6] [41] [22] [56] [27] [21]
ENHANCEMENT [37] [24] [69] [26]
REGISTRATION [30] [39] [57] [24]
FEATURE EXTRACTION [28] [29] [32] [34] [36] [46] [51] [62] [61] [63] [65] [66] [89]
SEGMENTATION [40] [45] [49] [52] [59] [60] [61] [63] [72]
ENHANCEMENT [58] [26]
REGISTRATION [39] [52]
FEATURE EXTRACTION [20] [67]
SEGMENTATION [44]
ENHANCEMENT No paper found
REGISTRATION [47]
REGISTRATION [23] [9]
/FEATURE EXTRACTION [33] [48] [54]
SEGMENTATION [9] [70]
ENHANCEMENT No paper found
FEATURE EXTRACTION No paper found
SEGMENTATION No paper found
ENHANCEMENT No paper found
REGISTRATION No paper found
Particle Swarm Optimization
IMAGE & FEATURE EXTRACTION [78]
IMAGE SEGMENTATION [82] [86] [64] [90]
IMAGE ENHANCEMENT [77] [64] [79]
IMAGE REGISTRATION [77]
IMAGE/FEATURE EXTRACTION [74][75] [88]
IMAGE SEGMENTATION No paper found
IMAGE ENHANCEMENT [76]
IMAGE REGISTRATION No paper found

Fig. 3. A hierarchical diagram showing the branch of Bio-inspired Algorithm in Image Processing field

Table 3 shows the number of publications for each BIAs applied to the four image processing areas where a high number of publications are concentrated with the GA and GP. Only a small number of publications use MA, ES, PSO and AIS whereas SFGA has no publication.

Table 3. Number of publications for each BIAs sorted by the image processing technique

Bio-inspired algorithm	**GA**	**SFGA**	**MA**	**GP**	**ES**	**PSO**	**AIS**
Image application							
Image segmentation	9	0	1	9	2	4	0
Image Extraction / Feature extraction	13	0	2	14	3	1	2
Image Registration	4	0	1	2	2	1	0
Image Enhancement	4	0	0	2	0	3	1
Total	30	0	4	27	7	9	3

4 Conclusion and Recommendation

This paper presents a bird's of eye view of BIAs techniques in image processing by identifying and analyzing approximately papers related to image processing. Out of a total of 130 papers (1995 to 2010), only 80 papers are found using BIAs techniques in image processing. It is also found that since 1995 to 2010, an exponential trend is observed which involved the research of BIAs for image processing techniques. The image processing techniques that employed BIAs are for the area of image and feature extraction (47%), segmentation (31%), enhancement (13%) and registration (9%). The Hierarchical diagram shows the BIAs techniques involved in image processing are the Genetic Programming, Artificial Immune System, Evolution Strategies, Genetic Algorithm, Swarm Intelligence and Memetic Algorithm. New BIAs algorithm such as Selfish Gene Algorithm should be explored as there is no literature yet to be found. The algorithm should work on image processing because of its' significant smaller CPU time and better robustness with the changes of the algorithm parameters and good convergence [101].

References

1. Sbalzarini, I.F., Müller, S., Koumoutsakos, P.: Multiobjective optimization using evolutionary algorithms. In: Proceedings of the CTR Summer Program 2000, Center for Turbulence Research, Stanford University, Stanford (2000)
2. Howse, J., Stapleton, G., Taylor, J.: Spider Diagrams, London Mathematical Society. LMS J. Compt. Math. 8, 145–194 (2005) ISSN 1461-1570
3. Jones, G.: Genetic and evolutionary algorithms. In: von Rague, P. (ed.) Encyclopedia of Computational Chemistry. John Wiley and Sons, Chichester (1998)
4. Pignalberi, G., Cucchiara, R., Cinque, L., Levialdi, S.: Tuning range segmentation by genetic algorithm. EURASIP Journal Appl. Sig. Proc. 8, 780–790 (2003)
5. Lee, Z.J., Su, S.F., Lee, C.Y.: Efficiently solving general weapon-target assignment problem by genetic algorithms with greedy eugenics. IEEE Trans. Syst., Man Cybernet. Part B Cybernet. 33, 113–121 (2003)

6. Ghosh, P., Melanie, M.: Prostate segmentation on pelvic CT images using a genetic algorithm. In: Proceedings of the SPIE on Medical Imaging 2008: Image Processing, vol. 6914, pp. 691442–691442-8 (2008)
7. Talbi, H., Batouche, M., Draa, A.: A Quantum-Inspired Evolutionary Algorithm for Multiobjective Image Segmentation. World Academy of Science, Engineering and Technology 31 (2007)
8. Huang, C.F., Rocha, L.M.: A systematic study of genetic algorithms with genotype editing. In: Proc. of 2004 Genetic and Evolutionary Computation Conference, vol. 1, pp. 1233–1245 (2004)
9. Yuan, X., Zouridakis, G., Situ, N.: Automatic Segmentation of Skin Lesion Images Using Evolution Strategies. Preprint submitted to Elsevier (2008)
10. Poli, R., Langdon, W.B., McPhee, N.F.: A Field Guide to Genetic Programming. Lulu Enterprises, UK (2008) ISBN 978-1-4092-0073-4
11. Corno, F., Reorda, M.S., Squillero, G.: Exploiting the Selfish Gene Algorithm for Evolving Cellular Automata. In: IJCNN 2000: IEEE-INNS-ENNS International Joint Conference Neural Networks, Como., vol. (I), pp. 577–581 (2000)
12. Vasiliauskas, A.: Selfish Gene Algorithm (2008), `http://coding-experiments.blogspot.com/2008/04/selfish-gene-algorithm.html`
13. Garg, P.: A Comparison between Memetic algorithm and Genetic algorithm for the cryptanalysis of Simplified Data Encryption Standard algorithm. International Journal of Network Security & Its Applications (IJNSA) 1(1) (2009)
14. El-Mihoub, T.A., Hopgood, A.A., Nolle, L., Battersby, A.: Hybrid Genetic Algorithms: A Review. Engineering Letters 13(2), EL_13_2_11 (2006)
15. Premalatha, K., Natarajan, A.M.: Hybrid PSO and GA for Global Maximization. Int. J. Open Problems Compt. Math. 2(4) (2009); ISSN 1998-6262, Copyright © ICSRS Publication
16. Brumby, S.P., Theiler, J., Perkins, S.J., Harvey, N.R., Szymanski, J.J., Bloch, J.J., Mitchell, M.: Investigation of image feature extraction by a genetic algorithm. In: Proc. SPIE, vol. 3812, pp. 24–31 (1999)
17. Brumby, S.P., Davis, A.B., Harvey, N.R., Rohde, C.A., Hirsch, K.L.: Genetic refinement of cloud-masking algorithms for the multi-spectral thermal imager (MTI). In: Proc. IGARSS, vol. 3, pp. 1152–1154 (2001)
18. Gunn, S.R., Nixon, M.S.: Snake Head Boundary Extraction using Local and Global Energy Minimisation. In: Proc. IEEE Int. Conf. on Pattern Recognition, Vienna, Austria, pp. 581–585 (1996)
19. Yen, G.G., Nithianandan, N.: Facial Feature Extraction Using Genetic Algorithm. In: Proceedings of the IEEE 2002 Congress on Evolutionary Computation, Honolulu, USA, vol. 2, pp. 1895–1900 (2002)
20. Radtke, P.V.W., Wong, T., Sabourin, R.: A Multi-objective Memetic Algorithm for Intelligent Feature Extraction. In: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (eds.) EMO 2005. LNCS, vol. 3410, pp. 767–781. Springer, Heidelberg (2005)
21. Ghosh, P., Melanie, M.: Segmentation of Medical Images Using a Genetic Algorithm. In: GECCO 2006, Seattle, Washington, USA (2006); Copyright ACM 1-59593-186-4/06/0007
22. Ganesan, R., Radhakrishnan, S.: Segmentation of Computed Tomography Brain Images using Genetic Algorithm. International Journal of Soft Computing 4, 157–161 (2009)
23. Zhang, J., Yuan, X., Buckles, B.P.: A fast evolution strategies-based approach to image registration. In: Genetic and Evolutionary Computation Conference, New York (2002)

24. Cordon, O., Damas, S., Santamaria, J.: A practical review on the applicability of different evolutionary algorithms to 3D feature-based image registration. In: Genetic and Evolutionary Computation for Image Processing and Analysis, p. 241 (2009)
25. Munteanu, C., Rosa, A.: Color image enhancement using evolutionary principles and the retinex theory of color constancy. In: Proceedings IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing XI, pp. 393–402 (2001)
26. Wetcharaporn, W., Chaiyaratana, N., Huvanandana, S.: Enhancement of an Automatic Fingerprint Identification System Using a Genetic Algorithm and Genetic Programming. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) EvoWorkshops 2006. LNCS, vol. 3907, pp. 368–379. Springer, Heidelberg (2006)
27. Paulinas, M., Usinskas, A.: A Survey of Genetic Algorithms Applicatons for Image Enhancement And Segmentation. Information Technology and Control 36(3), 278–284 (2007)
28. Harvey, N.R., Theiler, J., Brumby, S.P., Perkins, S., Szymanski, J.J., Bloch, J.J., Porter, R.B., Galassi, M., And Young, A.C.: Comparison Of GENIE And Conventional Supervised Classifiers For Multispectral Image Feature Extraction. IEEE Transactions on Geoscience and Remote Sensing 40(2) (February 2002)
29. Mohammad, D.: Multi Local Feature Selection Using Genetic Algorithm For Face Identification. International Journal of Image Processing 1(2), 1–10 (2007)
30. Ammar, H.H., Tao, Y.: Fingerprint Registration Using Genetic Algorithms. In: Proceedings of 3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology, pp. 148–154 (2000)
31. Yuizono T., Wang Y., Satoh K., Nakayama S.: Study On Individual Recognition For Ear Images By Using Genetic Local Search. In: Proceeding Congress Evolutionary Computation, pp. 237-242 (2002)
32. Maludrottu, S., Sallam, H., Regazzoni, C.S.: Sparse Shapes Prototype Modeling Using Genetic Algorithms. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 978–971 (2010) ISSN: 1522-4880, E-ISBN: 978-1-4244-7993-1, Print ISBN: 978-1-4244-7992-4
33. Ninot, J., Smadja, L., And Heggarty, K.: Road Sign Recognition Using A Hybrid Evolutionary Algorithm And Primitive Fusion. In: Paparoditis, N., Pierrot-Deseilligny, M., Mallet, C., Tournaire, O. (eds.) IAPRS, Saint-Mandé, France, September 1-3, Part 3A, vol. XXXVIII (2010)
34. Jabeen, H., And Baig, A.R.: Review of Classification Using Genetic Programming. International Journal of Engineering Science and Technology 2(2), 94–103 (2010)
35. Trujillo, L., Legrand, P., Olague, G., Pérez, C.B.: Optimization of the hölder image descriptor using a genetic algorithm. In: GECCO 2010: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, t Portland, Oregon, USA, pp. 1147–1154 (2010)
36. Downey, C., Zhang, M., Browne, W.N.: New crossover operators in linear genetic programming for multiclass object classification. In: GECCO 2010: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, Portland, Oregon, USA, pp. 885–892 (2010)
37. Sri Rama Krishna, K., Reddy, A.G., Giri Prasad, M.N.: Chandrabushan Rao K. & Madhavi M.: Genetic Algorithm Processor for Image Noise Filtering Using Evolvable Hardware. International Journal of Image Processing 4(3) (2010)

38. Venkatesan, S., Madane, S.S.R.: Experimental Research on Identification of Face in a Multifaceted Condition with Enhanced Genetic and ANT Colony Optimization Algorithm. International Journal of Innovation, Management and Technology 1(5) (2010) ISSN: 2010-0248
39. Goranin, N., Cenys, A.: Evolutionary Algorithms Application Analysis in Biometric Systems. Journal of Engineering Science and Technology Review 3(1), 70–79 (2010)
40. Miller, J.F., Smith, S.L., Zhang, Y.: Detection of microcalcifications in mammograms using multi-chromosome cartesian genetic programming. In: GECCO 2010: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, Portland, Oregon, USA, pp. 1923–1930 (2010)
41. Ghosh, P., Mitchell, M., Gold, J.: LSGA: combining level-sets and genetic algorithms for segmentation. Evolutionary Intelligence 3(1) (2010)
42. Aljuaid, H., Muhammad, Z., Sarfraz, M.: A Tool to Develop Arabic Handwriting Recognition System Using Genetic Approach. Journal of Computer Science 6(5), 490–495 (2010) ISSN 1549-3636
43. Kharrat, A., Gasmi, K., Messaoud, M.B., Benamrane, N., And Abid, M.: A Hybrid Approach for Automatic Classification of Brain MRI Using Genetic Algorithm and Support Vector Machine. Leonardo Journal of Sciences (17), 71–82 (2010) ISSN 1583-0233
44. Ramos, V.: The Biological Concept of Neoteny in Evolutionary Colour Image Segmentation - Simple Experiments in Simple Non-Memetic Genetic Algorithms. In: Applications of Evolutionary Computation. LNCS, Springer, Heidelberg (2010)
45. Pedrino, E.C., Saito, J.H., Roda, V.O.: A Genetic Programming Approach to Reconfigure a Morphological Image Processing Architecture. Hindawi Publishing Corporation International Journal of Reconfigurable Computing 2011, Article ID 712494, 10 (2010) doi:10.1155/2011/712494
46. Cattani, P.T., Johnson, C.G.: Typed cartesian genetic programming for image classification. In: Proceedings of the 2009 UK Workshop on Computational Intelligence, University of Nottingham, pp. 106–111 (September 2009)
47. Santamaria, J., Cordon, O., Damas, S., Garcia-Torres, J.M., Quirin, A.: Performance evaluation of memetic approaches in 3D reconstruction of forensic objects. Soft. Computing 13(8-9), 883–904 (2009)
48. Charbuillet, C., Gas, B., Chetouani, M., Zarader, J.L.: Optimizing Feature Complementarity by Evolution Strategy: Application to Automatic SpeakerVerification Université Pierre et Marie Curie-Paris6, UMR 7222 CNRS, Institut des Syst‘emes Intelligents et Robotique (ISIR), Ivry sur Seine, F-94200 France (2009)
49. Singh, T., Kharma, N., Daoud, M., Ward, R.: Genetic Programming Based Image Segmentation with Applications to Biomedical Object Detection. In: GECCO 2009: Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation Montréal, Québec, Canada (2009) Copyright ACM 978-1-60558-325-9
50. Anam, S., Islam, M.S., Kashem, M.A., Islam, M.N., Islam, M.R., Islam, M.S.: Face Recognition Using Genetic Algorithm and Back Propagation Neural Network. In: Proceedings of the International MultiConference of Engineers and Computer Scientists 2009, IMECS, Hong Kong, March 18–20, vol. I (2009)
51. Kowaliw, T., Banzhaf, W., Kharma, N., Harding, S.: Evolving novel image features using genetic programming-based image transforms. In: Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2009), pp. 2502–2507 (2009)

52. Ebner, M.: Engineering of computer vision algorithms using evolutionary algorithms. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) Advanced Concepts for Intelligent Vision Systems, Bordeaux, France, pp. 367–378. Springer, Berlin (2009)
53. Senthilkumaran, N., Rajesh, R.: Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches. International Journal of Recent Trends in Engineering 1(2) (May 2009)
54. Chen, X., Liu, X., Jia, Y.: Combining evolution strategy and gradient descent method for discriminative learning of bayesian classifiers. In: Proceeding GECCO 2009 Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation (2009) ISBN: 978-1-60558-325-9
55. Hemanth, D.J., Vijila, C.K.S., Anitha, J.: A Survey On Artificial Intelligence Based Brain Pathology Identification Techniques In Magnetic Resonance Images. International Journal of Reviews in Computing (2009)
56. Li, Y.: Vehicle extraction using histogram and genetic algorithm based fuzzy image segmentation from high resolution UAV aerial imagery. In: IAPRS, vol. XXXVII, part B3b, pp. 529–534 (2008)
57. Seixas, F.L., Ochi, L.S., Conci, A., Saade, D.M.: Image registration using genetic algorithm. In: GECCO 2008: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation (2008)
58. Harding, S., Banzhaf, W.: Genetic programming on gpus for image processing. In: Proceedings of the First International Workshop on Parallel and Bioinspired Algorithms (WPABA 2008), Toronto, Canada, pp. 65–72. Complutense University of Madrid Press, Madrid (2008)
59. Kadar, I., Ben-Shaharv, O., Sipper, M.: Evolving boundary detectors for natural images via genetic programming. In: Proceedings of the 19th Internation Conference on Pattern Recognition (2008)
60. Lu, X., Zhou, J.: Applications of Evolutionary Programming in Markov Random Field to IR Image Segmentation. In: Proceedings of the IEEE/ASME International Conference on Advanced Intelligent Mechatronics, Xi'an, China (2008)
61. Song, A., Ciesielski, V.: Texture segmentation by genetic programming. Evolutionary Computation 16(4), 461–481 (2008)
62. Trujillo, L., Olague, G.: Automated Design of Image Operators that Detect Interest Points. Evolutionary Computation 16(4), 483–507 (2008)
63. Ciesielski, V., Song, A., Lam, B.: Visual Texture Classification and Segmentation by Genetic Programming. In: Ciesielski, V., Song, A., Lam, B., Cagnoni, Lutton, Olague (eds.) Genetic and Evolutionary Image Processing and Analysis. Hindawi Publishing Corporation (2007)
64. Braik, M., Sheta, A., Ayesh, A.: Image Enhancement Using Particle Swarm Optimization. In: Proceedings of the World Congress on Engineering, WCE 2007, London, U.K, July 2-4, vol. I (2007)
65. Wijesinghe, G., Ciesielski, V.: Using restricted loops in genetic programming for image classification. In: Proc. IEEE Congr. Evol. Comput., pp. 4569–4576. IEEE, Singapore (2007)
66. Espejo, P., Ventura, S., Herrera, F.: A Survey on the Application of Genetic Programming to Classification. IEEE Transactions on Systems, Man and Cybernetics 40(2), 121–144 (2010)
67. Sheng, W., Howells, G., Fairhurst, M., Deravi, F.: A memetic fingerprint matching algorithm. IEEE Transactions on Information Forensics and Security 2(3), 402–412 (2007)

68. Kucukural, A., Yeniterzi, R., Yeniterzi, A., Sezerman, O.U.: Evolutionary Selection of Minimum Number of Features for Classification of Gene Expression Data Using Genetic Algorithms. In: GECCO 2007, London, England, United Kingdom (2007); Copyright ACM 978-1-59593-697-4/07/0007
69. Imam, M.H.: An Extremely Simple Operation For Drastic Performance Enhancement Of Genetic Algorithms For Engineering Design Optimization. International Journal of Engineering Science and Technology 2(11), 6630–6645 (2010)
70. Pérez, O., Patricio, M.A., García, J., Molina, J.M.: Improving the segmentation stage of a pedestrian tracking video-based system by means of evolution strategies. In: 8th European Workshop on Evolutionary Computation in Image Analysis and Signal Processing. EvoIASP, Budapest, Hungary (April 2006)
71. Zhang, Y., Rockett, P.I.: A generic optimal feature extraction method using multiobjective genetic programming: Methodology and applications. Submitted to IEEE Transactions on Knowledge and Data Engineering (2006)
72. Quintana, M.I., Poli, R., Claridge, E.: Morphological algorithm design for binary images using genetic programming. Genetic Programming and Evolvable Machines 7(1), 81–102 (2006) ISSN 1389-2576
73. Ji, Z., Dasgupta, D., Yang, Z., Teng, H.: Analysis of Dental Images using Artificial Immune Systems. In: IEEE Congress of Evolutionary Computation (CEC), Vancouver, BC, Canada (2006)
74. Su, L., Liu, X., Wang, X., Jiang, N.: Dimensional Reduction In Hyperspectral Images By Danger Theory Based Artificial Immune System. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, vol. XXXVII, Part B7 (2008)
75. Jackson, J.T., Gunsch, G.H., Claypoole, R.L., Lamont, G.B.: Blind Steganography Detection Using a Computational Immune System Approach. A Proposal Work in Progress International Journal of Digital Evidence (Winter 2002)
76. Zheng, H., Li, L.: An Artificial Immune Approach for Vehicle Detection from High Resolution Space Imagery. IJCSNS International Journal of Computer Science and Network Security 7(2) (February 2007)
77. Wachowiak, M.P., Smolíková, R., Zheng, Y., Zurada, J.M., Elmaghraby, A.S.: An Approach to Multimodal Biomedical Image Registration Utilizing Particle Swarm Optimization. IEEE Transactions on Evolutionary Computation 8(3), 289 (2004)
78. Kundra, E.H., Panchal, V.K., Singh, K., Kaura, H., Arora, S.: Extraction of Satellite Image using Particle Swarm Optimization. International Journal of Engineering (IJE) 4(1) (2010)
79. Kwok, N.M., Ha, Q.P., Liu, D.K., Fang, G.: Intensity-Preserving Contrast Enhancement for Gray-Level Images using Multi-objective Particle Swarm Optimization. In: Proceeding of the IEEE International Conference on Automation Science and Engineering, Shanghai, China, October 7-10 (2006)
80. Wang, C.M., Kuo, C.T., Lin, C.Y., Chang, G.H.: Application of Artificial Immune System Approach in MRI Classification. EURASIP Journal on Advances in Signal Processing, Article ID 547684, 8 (2008)
81. Corno, F., Reorda, M.S., Squillero, G.: A New Evolutionary Algorithm Inspired by the Selfish Gene Theory. In: Proceedings of the ACM Symposium on Applied Computing, San Antonio, Texas, United States, pp. 333–338 (1998) ISBN:1-58113-086-4
82. Das, S., Abraham, A., Konar, A.: Spatial Information Based Image Segmentation Using a Modified Particle Swarm Optimization Algorithm. In: Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications, vol. 02, pp. 438–444 (2006) ISBN:0-7695-2528-8

83. Eiben, A.E., Smith, J.E.: What is an evolutionary algorithm? In: Introduction to Evolutionary Computing. Springer, Heidelberg (2003)
84. Kanungo, P., Nanda, P.K., Samal, U.C.: Image Segmentation Using Thresholding and Genetic Algorithm. In: Proceedings of the Conference on Soft Computing Technique for Engineering Applications, Rourkela, India, pp. 24–26 (2006)
85. Foon, D.W., Mandava, R., Ramachandram, D.: Deformable Boundary initialization for object Detection in Natural Images Using Multiple Scale Edges, Computer Science Postgraduate Colloquium, School of Computer Sciences, Universiti Sains Malaysia (USM), Penang (2004)
86. Ibrahim, S., Abdul Khalid, N.E., Manaf, M.: Particle Swarm Optimization – Brain Abnormalities Segmentation. In: International Conference on Robotics, Vision, Information and Signal Processing, ROVISP 2009, Langkawi, Malaysia (2009)
87. Ooi, T.H., Ngah, U.K., Abd. Khalid, N.E., Venkatachalam, P.A.: Mammographic Calcification Clusters Using The Region Growing Technique. In: New Millenium International Conference on Pattern Recognition, Image Processing and Robot Vision (PRIPOV 2000), pp. 157–163. Terengganu Advanced Technical Institute (TATI), Terengganu (2000)
88. Ji, Z., Dasgupta, D., Yang, Z., Teng, H.: Analysis of dental images using artificial immune systems. In: Proceedings of Congress on Evolutionary Computation (CEC), pp. 528–535. IEEE Press, Los Alamitos (2006)
89. Zhang, Y.: Multiobjective genetic programming optimal search for feature extraction. Ph.D. thesis, University of Sheffield (2006)
90. Afifi, A., Nakaguchi, T., Tsumura, N., Iyake, Y.: 2Shape and Texture Priors for Liver Segmentation in Abdominal Computed Tomography Scans Using the Particle Swarm Optimization Algorithm. Medical Imaging Technology 28(1) (2010)
91. Wang, C.M., Kuo, C.T., Lin, C.Y., Chang, G.H.: Application of Artificial Immune System Approach in MRI Classification. EURASIP Journal on Advances in Signal Processing, Article ID 547684, 8 (2008)
92. Hofmeyr, S.A., Forrest, S.: Immunity by design: an artificial immune system. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), pp. 1289–1296. Morgan Kaufmann, San Francisco (2004)
93. Poli, R.: Analysis of the Publications on the Applications of Particle Swarm Optimisation. Journal of Artificial Evolution and Applications, Article ID 685175, 10 (2008), doi:10.1155/2008/685175
94. Aickelin, U., Dasgupta, D.: Artificial immune systems tutorial. In: Burke, E., Kendall, G. (eds.) Search Methodologies—Introductory Tutorials in Optimization and Decision Support Techniques, pp. 375–399. Kluwer, Dordrecht (2005)
95. Eberhart, R.C., Shi, Y.: Comparison between genetic algorithms and Particle Swarm Optimization. In: Porto, V.W., Saravanan, N., Waagen, D., Eiben, A.E. (eds.) Evolutionary Programming VII, pp. 611–616. Springer, Heidelberg (1998)
96. Felzenszwalb, P., Huttenlocher, D.: Efficient Graph-Based Image Segmentation. Int'l J. Computer Vision 59(2), 167–181 (2004)
97. Lange, H., Ferris, D.G.: Computer-aided-diagnosis (CAD) for colposcopy. In: Proceedings of Medical Imaging: Image Processing, vol. 5747, pp. 71–84 (2005)
98. Clow, B., White, T.: An evolutionary race: A comparison of genetic algorithms and particle swarm optimization for training neural networks. In: Proceedings of the International Conference on Artificial Intelligence, IC-AI 2004, vol. 2, pp. 582–588. CSREA Press (2004)

99. Jadhav, D.G., Pattnaik, S.S., Devi, S., Lohokare, M.R., And Bakwad, K.M.: Approximate Memetic Algorithm For Consistent Convergence. In: National Conference on Computational Instrumentation, NCCI 2010, CSIO Chandigarh, India (2010)
100. Ciesielski, V., Mawhinney, D.: Prevention of early convergence in genetic programming by replacement of similar programs. In: Xin, Y. (ed.) Proceedings of the Congress on Evolutionary Computation (2002)
101. Popa, R.: Hybridated Selfish Gene Algorithm. In: IEEE International Conference on Artificial Intelligence Systems, ICAIS 2002 (2002)
102. Beyer, H.-G., Schwefel, H.P.: Evolution strategies: A comprehensive introduction. Nat. Comput. 1(1), 3–52 (2002)
103. Angeline, P.J.: Genetic programming and emergent intelligence. In: Kinnear Jr, K.E. (ed.) Advances in Genetic Programming, ch. 4, pp. 75–98. MIT Press, Cambridge (1994)
104. Digalakis, J., Margaritis, K.: Performance comparison of memetic algorithms. Journal of Applied Mathematics and Computation 158, 237–252 (2004)
105. Villalobos-Arias, M., Coello Coello, C.A., Hernández-Lerma, O.: Convergence analysis of a multiobjective artificial immune system algorithm. In: Nicosia, G., Cutello, V., Bentley, P.J., Timmis, J. (eds.) ICARIS 2004. LNCS, vol. 3239, pp. 226–235. Springer, Heidelberg (2004)
106. Timmis, J.: Artificial immune systems: Today and tomorow. Natural Computing 6(1), 1–18 (2007)
107. Yang, C., Li, Y., Lin, Z.: SGEGC: A Selfish Gene Theory Based Optimization Method by Exchanging Genetic Components. In: Cai, Z., Li, Z., Kang, Z., Liu, Y. (eds.) ISICA 2009. LNCS, vol. 5821, pp. 53–62. Springer, Heidelberg (2009)

The Use of Elimination Method and Nearest Neighbor for Oil Palm Fruit Ripeness Indicator

Mohamad Fatma Susilawati, Azizah Abdul Manaf, and Suriayati Chuprat

Advanced Informatic School, UTM International Campus,
Jalan Semarak, 54100 Kuala Lumpur
{fatma,azizah07,suria}@unisza.edu.my,
{azizah07,suria}@ic.utm.my

Abstract. Fruit ripeness identification is hard to measure especially when it involves color as main indicator. Suitable color model must be chosen to determine the right color for the ripeness identification. Hue, Saturation and Value (HSV) are proved to be a better choice because it can define the color intensity. Besides, it also helps to choose colors which are similar to the eyes. Manual grading process by human graders at oil palm mills lead to misconduct and mistakenly claimed unripe fruits as the ripe ones. This will cause trouble when the error report arrives at the production site for the oil production sterilization process. Furthermore, research done by the Federal Land Development Authority (FELDA) in Malaysia stated that an approximation of 60% palm oil are coming from ripe fruit meanwhile 40% are from underripe and 20% from unripe fruit minus water and dirt. This is proved the importance of identifying the right fruits for the purpose of oil palm production is extremely important. This paper studies the use of elimination method and nearest neighbor for oil palm fruit ripeness indicator. Result shows value gives the best indicator by providing the highest recognition rate towards ripe and unripe category.

Keywords: Fruit Ripeness Identification; HSV; Elimination Method; Nearest Neighbor ; Recognition Rate.

1 Introduction

Recent development of oil palm research undergoes the determining of oil content by providing a computer-based grading system. Currently, manual system conducted by human grader at oil palm mills lead to misconduct and disputes. Since color is main indicator of ripeness, it is important to research for a right technique to determine the fruit ripeness identification. As mentioned by [1], color provides valuable information in estimating the maturity and examining the freshness of fruits and vegetables. Color recognition model has been applied widely in industrial sectors, commercial fields as well as in social responsibilities. For instance, it is used as a powerful and reliable parameter in robotics machines, aid for the blind and the color blind people, diamond color sorting, quality control for the manufacture colored paper, [2] and in characterizing the thermal paints. [3].

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 674–686, 2011.

The normal human eyes have three types of sensors and the signal of these three sensors determine the color response of the observer. The response of this system produces the three-dimensional phenomenon of three dimensional spaces. When a human sees something, the light enters the eye and hit the light detector on the retina. This behaves similarly to a digital camera that records more reading whenever more lights hit the light detector on the back of the camera [4]. Two main characteristics that are decisive for visual inspection and classification of fruits are color and shape [5].

It is very important to correctly identify the fruit ripeness category due to its correlation to their oil content as mentioned by [15], otherwise the grading process will be time consuming and unreliable.

Table 1. Estimated Oil Content

Fruit Category	Estimated Oil Content
Ripe	60%
Underripe	40%
Unripe	20%

Tab. 1 shows the estimated oil palm content for ripe, underripe and unripe category of FFB by research group of Federal Land Development Authority (FELDA) [15] at oil palm mills. The estimated of oil content is made after the consideration of the presence of water and dirt for every FFB. From the table, we can see that the ripe category of fruit provide the highest percentage of estimated of oil content and this indicate the importance of the right identification of FFB before it was taken for the sterilization process for oil production.

2 Related Work

Several studies have been conducted to measure the ripeness of oil palm fruits especially to measure its oil content. Among the studies are explained below.

An open-ended coaxial probe is used by [6] as a moisture sensor to determine the moisture content *m.c* of oil palm fruits of various degrees of fruit ripeness at room temperature by using reflection techniques. Furthermore, a model is developed to describe the relation between reflection coefficient and moisture content *m.c* of the oil palm mesocarp for frequency range 1 GHz to 5 GHz.

[7] in their work introduced a color meter to measure ripeness of oil palm fruits. Therefore, the trials had shown that there was no significant difference in fruit ripeness between using the color meter and the human grader. However, color meter may give consistent results compared to human grader as the latter can be biased.

Meanwhile [8] present an automated technique of quantifying fruit color components based on digital images captured in Joint Photographic Experts Group (JPEG) format. This technique is based on a simple computer program written in Visual Basics and interfaced with ILWIS 3.2. Using the empirical relationship between oil content/DOBI and fruit color, an additional step of estimating % total oil content and/or DOBI values is also made possible. They employed a digital imaging approach to quantify fruit color.

A non-destructive technique for measuring the colour of oil palm fruit (Elaeis guineensis) bunches at different stages of maturity and correlating the colour data with the oil content in the fruit bunch is investigated by [9] in their research study. The study showed statistical evaluation showed that the digital imaging technique for determining FFB oil content can be successfully used on a homogeneous population of palms that display similar change of colour during the ripening process.

A computer assisted photogrammetric methodology is developed by [10] to correlates color of oil palm fruits with their ripeness by calculating the color Digital Numbers (DN). The result of developing a complete automation grading system is achieved. However, the images taken a day after the fruits were delivered will cause the fruits color change, less freshness, and this will affect the oil content in the fruits. This is in parallel with [11] which says that fruits must be graded within 24 hours after harvested.

Another automated grading system is also developed by [12] based on RGB color model. The color elements of Red, Green and Blue were analyzed using this grading system. The mean color intensity is used to differentiate between different color and its ripeness. However, results are only limited to Ripe, Under Ripe and Over Ripe categories of fruits which are insufficient to detect for the ripeness without take into account for Unripe fruits. This is important as major category of fruits for the ripeness indicator.

3 Methodology

An ongoing study is conducted to use similarity measures for oil palm fruit application. The approach is intended to classify the ripeness of oil palm fruit based on histogram. We are going to explore the potential features from the color histogram and incorporate these features into the distance measurement. For this purpose of study, Nearest Neighbor Distance is chosen for its suitability for Histogram Distance. Histogram is explored for the ability of increasing the global contrast of many images, especially when the usable data of the image is represented by close contrast values. Through this adjustment, the intensities can be better distributed on the histogram. This allows for areas of lower local contrast to gain a higher contrast without affecting the global contrast [13]. Histogram equalization accomplishes this by effectively spreading out the most frequent intensity values.

In order to develop a ripeness indicator, the following tasks have been carried out:

- Data collection
- Process Flow
- HSV Color Model
- Elimination Method
- Nearest Neighbor Distance

3.1 Data Collection

Images of oil palm fruits are captured by using a digital camera by an authorized grader of Felda mill at Bukit Sagu, Kuantan and Felda mill at Jerangau, Terengganu. For a training set of data, 30 images of FFB bunches are captured for a ripe category (based on prior knowledge) and another 30 bunches of FFB for unripe category. For testing set of data, 30 images of FFB bunches are captured for the unknown category that we assume ripe and another 30 bunches for unknown unripe.

As mentioned by [6,7,8,9,10,11,12] who studied in various applications, we are currently working on estimating the ranges of features extracted from the identified categories of oil palm fruits. Since there are not many studies done on this particular topic, especially in identifying the oil palm fruit ripeness by using histogram-based distance metric, we are going to explore the possibility of using histogram and make use of its features and tested by using Nearest Neighbor Distance. The ripeness bunches of fruits are identified based on prior knowledge. A data of approximately 30 bunches of fruits are collected for each category. The ranges are then computed for each fruit category. Sample of two category of oil palm fruits with their HSV color model histogram are shown below:

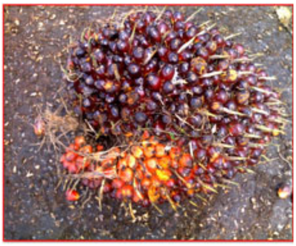


Fig. 1. Ripe FFB

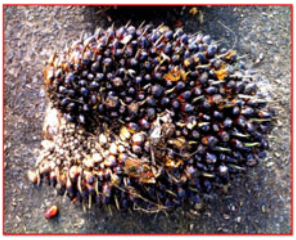


Fig. 2. Unripe FFB

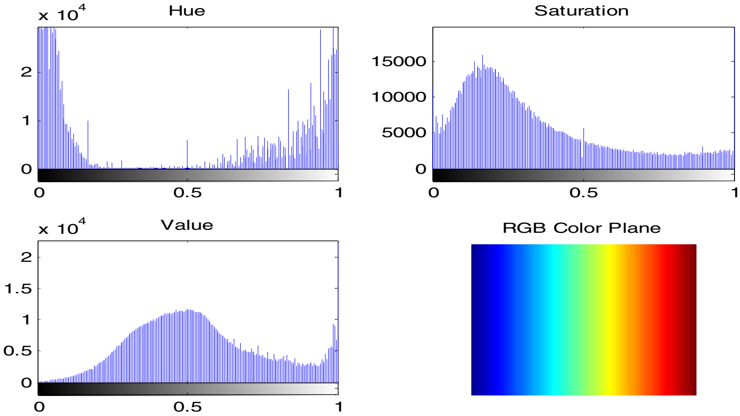


Fig. 3. HSV Histogram with RGB Color Panel for Ripe FFB

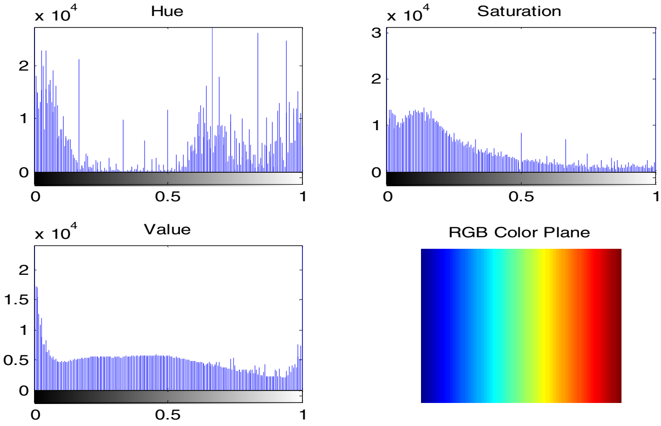


Fig. 4. HSV Histogram with RGB Color Panel for Unripe FFB

Figure 1 and 2 shows the sample image of ripe and unripe images of FFB which taken at the grading site at oil palm mill. Meanwhile Figure 3 shows the histogram for Hue, saturation and value also the RGB color panel for the ripe FFB and Figure 4 shows the HSV histogram and the RGB color panel for the unripe FFB.

3.2 Process Flow

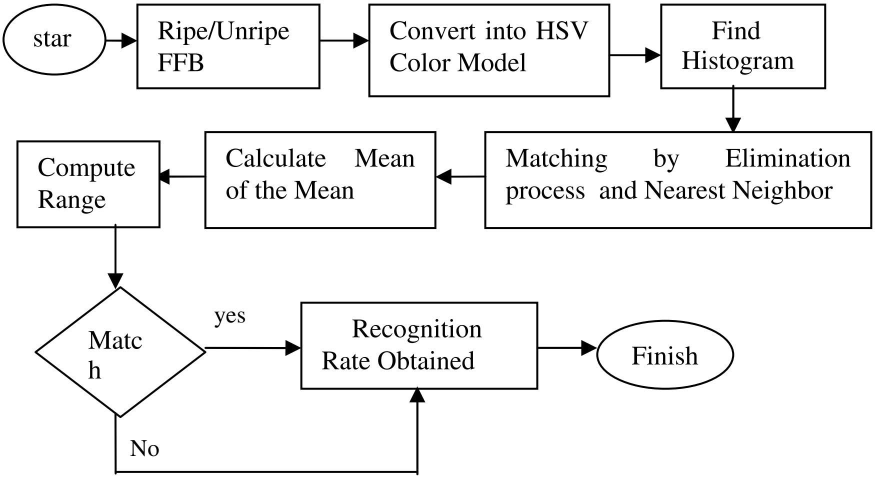

star
Ripe/Unripe FFB
Convert into HSV Color Model
Find Histogram
Matching by Elimination process and Nearest Neighbor
Calculate Mean of the Mean
Compute Range
Matc h
yes
Recognition Rate Obtained
Finish
No

Fig. 5. Process Flow

Refer to Fig.5., images of oil palm fruits in the form of JPEG format are captured at oil palm mill. Images are then converted to HSV color model from RGB color model. Then the histogram features are extracted for every element of Hue, Saturation and Value of HSV color model. Next, calculations are performed for the defined features extracted (mean of the mean value). Then, H, S, and V values are calculated for every bunch using Nearest Neighbor Distance. After that, a range of min and max values are also calculated for each H, S and V for every bunch of oil palm fruits for the ripe and unripe category. The same processes are then repeated for an unknown sample of ripe fruits (30 bunches) and unripe fruits (30 bunches). Lastly, comparisons are made to match the unknown fruits with known category of fruits and the correct matches are then calculated

3.3 HSV Color Model

HSV color space is chosen because it has information about color in one channel [3]. HSV stands for Hue, Saturation and Value. Hue is the color itself. Saturation is the "quantity" of colors, meanwhile Value is a kind of brightness. Typical users perceive colors normally in the form of HSV color space. Besides, HSV color space is also chosen when you want to match for the right colors or to choose colors which looks similar.

Figure 1 and 2 show images of oil palm fruits in the form of JPEG format are captured at oil palm mill. Images are then converted to HSV color model from RGB color model. Then the histogram features are extracted for every element of Hue, Saturation and Value of HSV color model. Next, calculations are performed for the defined features extracted (mean of the mean value). Then, H, S, and V values are calculated for every bunch using Nearest Neighbor Distance. After that, a range of min and max values are also calculated for each H, S and V for every bunch of oil palm fruits for the ripe and unripe category. The same processes are then repeated for an unknown sample of ripe fruits (30 bunches) and unripe fruits (30 bunches). Lastly, comparisons are made to match the unknown fruits with known category of fruits and the correct matches are then calculated.

HSV color model is defined as follows [16]:

$$H = \begin{cases} \frac{60(G-B)}{R} & if\, MAX = R \\ \frac{60(B-R)}{G} & if\, MAX = G \\ \frac{60(R-G)}{B} & if\, MAX = B \\ Not\ defined & \end{cases} \tag{1}$$

$$S = \begin{cases} \frac{\delta}{MAX} & if\, MAX \neq 0 \\ 0 & if\, MAX = 0 \end{cases}$$

$$V = MAX$$

where δ = (MAX - MIN), MAX = max(R, G, B), and MIN = min(R, G, B). Note that the R, G, B values in Equation (1) are scaled to [0, 1]. In order to confine H within the range of [0, 360],

$$H=H+ 360, \text{ if } H<0.$$

3.4 Elimination Process

Elimination process is applied during feature matching process. For example Ripe1 is compared to Ripe2, Ripe3, Ripe4 until Ripe 30. Then, after completed one cycle of process, Ripe1 is eliminated. Then Ripe2 is compared with Ripe3 until Ripe 30 and Ripe2 is eliminated. This process will be repeated until the last cycle which is Ripe29 is compared with Ripe30. The same process will repeat for unripe category of FFB for both training and tenting set of data. For every cycle of process, a set of mean of the mean for each HSV element is obtained.

Table 2. Elimination method matrix table

		R1		
No	Test	Hue	Saturation	Value
1	R1,R2	0.3184	0.3569	0.5563
2	R1,R3	0.393	0.3569	0.5327
3	R1,R4	0.3555	0.3569	0.5533
4	R1,R5	0.4023	0.3569	0.5512
5	R1,R6	0.2379	0.3569	0.5542
6	R1,R7	0.2818	0.3569	0.547
7	R1,R8	0.1949	0.3569	0.5106
8	R1,R9	0.2855	0.3569	0.4741
9	R1,R10	0.4231	0.3569	0.5614
10	R1,R11	0.3	0.3569	0.5304
11	R1,R12	0.2342	0.3569	0.5276
12	R1,R13	0.4154	0.3569	0.5276
13	R1,R14	0.2621	0.3569	0.5398
14	R1,R15	0.305	0.3569	0.5452
15	R1,R16	0.3395	0.3569	0.5332
16	R1,R17	0.4226	0.3569	0.5616
17	R1,R18	0.2772	0.3569	0.5471
18	R1,R19	0.3487	0.3569	0.5119
19	R1,R20	0.2347	0.3569	0.5502

Table 2. (*continued*)

20	R1,R21	0.2347	0.3569	0.5488
21	R1,R22	0.4231	0.3569	0.5616
22	R1,R23	0.1208	0.3569	0.5616
23	R1,R24	0.2319	0.3569	0.5616
24	R1,R25	0.1705	0.3569	0.5425
25	R1,R26	0.1572	0.3569	0.5616
26	R1,R27	0.1502	0.3569	0.5616
27	R1,R28	0.2319	0.3569	0.5277
28	R1,R29	0.1056	0.3569	0.5616
29	R1,R30	0.2999	0.3569	0.5599
MEAN OF THE MEAN		**0.28129655**	**0.3569**	**0.543582759**

3.5 Nearest Neighbor Distance

We use nearest neighbor (NN) distance to compare the similarity between HSV elements of the histograms. In this study, we use histogram as a feature vector. Mean value are extracted from the histogram for each H,S and V element. Nearest neighbor is used to compare between the histograms until a set of range value is obtained. The formula used is shown below:

$$(h_1,h_2) = \sum_{i=n}^{n} \min(h_1,h_2) \tag{2}$$

h_1 is the known category of FFB while h_2 is the individual FFB which falls under unknown category. In this case, we have 4 different set of FFB bunches (30 bunches each) and the dataset obtained are from prior knowledge by appointed mill grader. The processes are divided into two stages. Details of the processes are explained below:

3.5.1 Known Category

First, mean value is extracted from HSV histogram for ripe and unripe category of FFB. Total number of FFB used in this study is 30 for ripe and another 30 for unripe category. After that, we compare every FFB within the same category of fruits one by one until all 30 bunches using Nearest Neighbor Distance. Then, we calculate mean value for each matrix table for FFB 1 until FFB30. Since the comparison is made between FFB1 and FFB2, FFB1 and FFB3 until 30 bunches of fruits, the total number of FFB after the comparison is just 29. After that, for every mean value of 29 FFB, we compute the min and max value for the range. Then, mean of the mean value for every min and max for the total bunches are calculated for that particular category of each HSV elements.

3.5.2 Unknown Category

For unknown category, we compare unknown set of 30 fruits for every FFB bunch using Nearest Neighbor Distance for every HSV element until mean is computed. Then, for every bunch of FFB, we calculate min and max to compute range value. After that, we match every bunch of FFB one by one with known range of particular category of fruits whether ripe or unripe. Finally, recognition rate is obtained for every HSV element.

4 Experimental Design

Fig.6 shows general block diagram for oil palm fruit ripeness identification. The processes start from Image Acquisition, Feature Extraction, Feature Matching and Image Identification until Recognition Rate is obtained.

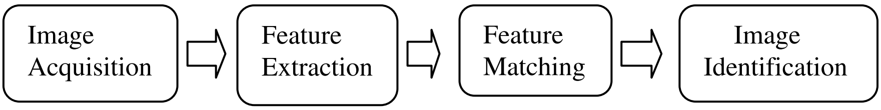


Fig. 6. General Block Diagram

The processes start from Data Gathering, Feature Extraction, Feature Matching, until the Recognition Rate obtained. Details of processes are shown below.

Step 1: Take 30 sample of ripe and another 30 sample of unripe oil palm fruits.

Step 2: Find mean value for each H, S, and V for every bunch of fruits

Step 3: Use Nearest Neighbor Distance and calculate min and max value for each H, S, and V for every bunch of fruits

Step 4: Calculate mean of min and mean of max value for each H, S, and V for every bunch of fruits

Step 5: Calculate mean of min and mean of max for each ripe and unripe category of oil palm fruits

Step 6: Take another sample of 30 bunches of unknown ripe fruits and another 30 bunches of unknown unripe fruits and repeat step 2 – step 5

Step 7: Match the result of unknown bunches of oil palm fruits with known ones. i.e. unknown ripe with ripe and unknown unripe with unripe

Step 8: Find the Recognition Rate or “match” rate as shown in Table 1 – 7.

5 Result and Analysis

Elimination method is used to compare between FFB for both category of fruits. Therefore, we are able to finalize the mean of the mean by using Nearest Neighbor Distance during elimination process. Based on the algorithm above, results are depicted in table 3 until 9 and divided into 2 categories which are Ripe and Unripe. Recognition rate is obtained and shown for Hue, Saturation and Value as below:

5.1 Ripe FFB

At this stage, unknown category of FFB (individual fruits which we assume ripe) are matched with the known category of ripe FFB (which we already tested earlier). Every element of H, S, and V for unknown FFB are matched with H,S, and V element of known FFB which in this case is ripe and the results are shown below.

Table 3. Hue ripe

Hue	
Total No of Fruits	29
Matched	11
Recognition Rate	38%

Table 3 shows the recognition rate for Hue element of ripe category of oil palm fruits. From the table, 11 out of 29 bunches of FFB are matched within the range value of known ripe category and the recognition rate is 38%

Table 4. Saturation ripe

Saturation	
Total No of Fruits	29
Matched	15
Recognition Rate	52%

Table 4 shows the Saturation element of the ripe category. Out of 29, 15 bunches of fruits are matched within the category and the recognition rate is slightly more than half, which is 52%.

Table 5. Value ripe

Value	
Total No of Fruits	29
Matched	26
Recognition Rate	90%

In table 5 Value proved to be the best ripeness indicator which provide 90% recognition rate. Out of 29 bunches of FFB, 26 are matched. This provide good indicator for ripeness identification.

5.2 Unripe FFB

At this stage, individual element of H,S,V of unknown (unripe) FFB is matched with the known category of unripe FFB. Results are as below.

Table 6. Hue unripe

Hue	
Total No of Fruits	29
Matched	7
Recognition Rate	24%

Table 6 shows Hue result for the unripe category. From 29 bunches of FFB, only 7 bunches are matched and the recognition rate is 24%.

Table 7. Saturation unripe

Saturation	
Total No of Fruits	29
Matched	14
Recognition Rate	48%

In Table 7, from the total number of 29, 14 bunches are matched and this provide 48% recognition rate for the Saturation value.

Table 8. Value unripe

Value	
Total No of Fruits	29
Matched	24
Recognition Rate	83%

Table 8 shows the highest recognition rate for Value (unripe category) which is 83%. Out of 29 bunches, 24 are matched. This proved to be a significant indicator for ripeness identification.

Table 9. Unknown ripe vs. unknown unripe

	Hue	Saturation	Value
Ripe	38%	52%	90%
Unripe	24%	48%	83%

Table 9 shows the comparison result between unknown bunches which we assume ripe and unknown bunches which we assume unripe (based on prior knowledge). Overall result shows the recognition rate for Nearest Neighbor Distance falls under Value element of HSV color model. Value is proved to be a good indicator for ripeness identification. 90% of oil palm fruits are correctly identified as ripe and 83% for unripe. In manual grading process, colors are one of the main indicators to correctly identify whether the fruits are ripe or unripe. This is in line with the definition set by Malaysian Palm Oil Board (MPOB) which "ripe bunch is a fresh bunch which has reddish orange color" [10] which is hard to define since it involves color intensity. This is the main reason why HSV color space is chosen compare to RGB color space which is describing only on 3 primary colors without detailing on color intensity. Furthermore as described by [9], "value is the color lightness", this research proves to be very promising and encouraging research work for fruit ripeness identification.

6 Conclusion and Future Work

Promising result in comparing ripe and unripe category of fruits were successful obtained. Color lightness proved the most distinctive difference between ripe and unripe. Further investigation will look deep into color lightness as a major indicator. Nearest Neighbor proves as a good distance measurement for histogram-based features. Currently, Furthest Neighbor and Mean Distance are also experimented to find the most promising result. All the Distances will be compared to find the best result. We also try to combine the extracted features using PCA, K-SOM and few other techniques to be incorporated with Similarity Measurement Distance and find the best matching technique for fruit ripeness identification.

Acknowledgement

Special thanks to UTM International Campus and University of Sultan Zainal Abidin for the help and support of this research.

References

1. Alfatni, M.S.M., Mohamed Sharif, A.R., Mohamad Shafri, H.Z.: Oil Palm Fruit Bunch Grading System Using Red, Green and Blue Digital Number. Journal of Applied Sciences 8(8), 1444–1452 (2008)
2. Stoksik, M., Nguyen, D.T., Czemkowaki, M.: A Neural Net Based Color Recognition System. University of Tasmania, Australia
3. Lalanne, T., Lempereur, C.: Color Recognition with a Camera, A Supervised Algorithm Recognition, Toulouse Cedex, France
4. Peter, S.: Fundamentals of Computer Graphics. A K Peters, Ltd., Standford (2005); is a leading independent scientific technical publisher based in Wellesley, Massachusetts
5. Effendi, Z., Ramli, R., Ghani, J.A.: A Back Propagation Neural Networks for Grading Jatropha curcas Fruits Maturity. American Journal of Applied Sciences 7(3), 390–394 (2010); ISSN 1546-9239 © 2010 Science Publications
6. Yeow, Y.K., Abbas, Z., Khalid, K.: Application of Microwave Moisture Sensor for Determination of Oil Palm Fruit Ripeness. Measurement Science Review 10(1) (2010)
7. Omar, I., Khalid, M.A., Harun, M.H., Wahid, M.B.: Colour Meter For Measuring Fruit Ripeness. MPOB Information Series (2003)

8. Balasundram, S.K., Mohd Hanif, A.H., Abd Rahim, A.: computerized digital imaging technique to estimate palm oil content and quality based on fruit color
9. Tan, Y.A., Low, K.W., Lee, C.K., Sang, K.: Imaging technique for quantification of oil palm fruit ripeness and oil content. Eur. J. Lipid Sci. Technol. 112, 838–843 (2010)
10. Jaffar, A., Jaafar, R., Jamil, N., Low, C.Y., Abdullah, B.: Photogrammetric Grading of Oil Palm Fresh Fruit Bunches. International Journal of Mechanical & Mechatronics Engineering IJMME 9(10)
11. MPOB, Fresh Fruit Bunch Grading Manual. Ministry of Primary Industries, Malaysia (1995)
12. Alfatni, M.S.M., Mohamed Shariff, A.R., Mohd Shafri, H.Z., Ben, O.M., Eshanta, O.M.: Oil Palm Fruit Bunch Grading System Using Red, Green and Blue Digital Number. Journal of Applied Sciences 8(8), 1444–1452 (2008); ISSN 1812-5654 @ 2008 Asian Network for Scientific Information
13. Histogram Equalization. Wikipedia The Free Encyclopedia (2010)
14. NetMBA. Internet Center for Management and Business Administration, Inc. The Histogram, `http://www.netmba.com/statistics/histogram/` (accessed)
15. Felda Agricultural Services Sdn Bhd. Analisa Oil MPD Kilang-Kilang FPISB (2010)

Crowd Analysis and Its Applications

Nilam Nur Amir Sjarif[1], Siti Mariyam Shamsuddin[2],
Siti Zaiton Mohd Hashim[3], and Siti Sophiayati Yuhaniz[4]

Soft Computing Research Group,
K-Economy Research Alliance
Universiti Teknologi Malaysia, Skudai, Johor
nilamnini@gmail.com, mariyam@utm.my,
sitizaiton@utm.my, sophia@utm.my

Abstract. Crowd is a unique group of individual or something involves community or society. The phenomena of the crowd are very familiar in a variety of research discipline such as sociology, civil and physic. Nowadays, it becomes the most active-oriented research and trendy topic in computer vision. Traditionally, three processing steps involve in crowd analysis, and these include pre-processing, object detection and event/behavior recognition. Meanwhile, the common process for analysis in video sequence of crowd information extraction consists of Pre-Processing, Object Tracking, and Event/Behavior Recognition. In terms of behavior detection, the crowd density estimation, crowd motion detection, crowd tracking and crowd behavior recognition are adopted. In this paper, we give the general framework and taxonomy of pattern in detecting abnormal behavior in a crowd scene. This study presents the state of art of crowd analysis, taxonomy of the common approach of the crowd analysis and it can be useful to researchers and would serve as a good introduction related to the field undertaken.

Keywords: Crowd analysis, pre-processing, object tracking, event behavior recognition.

1 Introduction

Handling the situation that is related with the abnormal in a crowd is not as simple easy [1]. The most important think that the researcher should consider in probing the problem of the crowd includes: i) How crowded the scene is, and ii) whether the situation is normal or abnormal. The issue such as complexity and abstraction of identifying and detecting abnormal behavior in crowd scene has attracted many researchers [2]. However, there are some difficulties in analyzing behavior in a crowd scene. In order to achieve the goal, the analyzing procedure must be done comprehensively through video surveillance; hence earlier works have been excellently reviewed by [3-5].

With the intelligent digital camera technology Closed-Circuit Television (CCTV), video surveillance has becoming more important. CCTV is used to observe parts of a process from control environment which is required in every intelligent crowded

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 687–697, 2011.

scene. One of the trendy topics in video surveillance is on crowd analysis to automatically detect the anomalies and alarms [5]. Crowd analysis consists of four phases: crowd density estimation, crowd motion detection, crowd tracking and crowd behavior understanding [2-4, 6].

The remainder of the paper is organized as follows: Section 2 presents the crowd analysis and related studies. Section 3 consists of framework of crowd analysis. Section 4 describes common approaches of crowd analysis. Section 5 consists of application of crowd analysis. Finally, in Section 6 we draw the conclusion together with some discussion.

2 Crowd Analysis and Related Studies

The terms of crowd or known as 'mob' or 'mob rule' can be define as a collective characteristic such as 'an angry crowd', a peaceful crowd', and 'a panic crowd' are well accepted. Crowd is made up of the independent individual's parts, whereby each of them have their own objectives and behavior pattern which differ from the expected individually from its participants. In a crowded scene, the individual making much more variable and complex and need to do some mathematical rules of behavior that might be useful to approximate the behavior [7]. There are various ways in analyzing the crowd for detection the abnormal such as using crowd density estimation, crowd motion detection, crowd tracking and crowd behavior recognition.

Crowd scene can be divided into two types: structured crowded scene and unstructured crowded scene [8] . The terms *structured crowded scene* can be described as crowd moves coherently in common direction, motion direction does not very time, each spatial locations of the scene supports only one dominant crowd behavior over the time. For example marathon race, queues of people event and traffic on the road. Meanwhile the term *unstructured crowded scene* represents the random crowd motion; different participants moving different direction at different times, each spatial location supports multi –modal, and crowd behavior. For example people walking on a zebra crossing in opposite directions, exhibitions, sporting event, railway stations, airport and motion biological cells. Fig 1 shows the sample of image structured and unstructured crowded scene based on human.

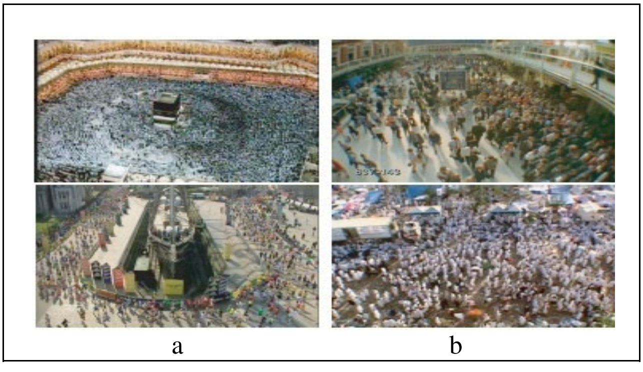


Fig. 1. Image of crowded scene (a) Structured, (b) Unstructured [8]

3 A Framework of Crowd Analysis

The important component attributes in a crowd consists of density, location, speed, color and etc. By using the computer visions, the information can be extracted either automatically or manually. Two types of sensors are used to capture the scene process include typology sensor and topology sensor. To get more accurate information in a crowd scene, the process of extraction information should be depending on the conditions of environment such as illumination changes (transition from day to night, shadow of background images and non static background like leaves blown by the wind could be detected as moving object), handling the occlusion, multiple input channel and amount number of cameras, the changes of motion and detecting different characteristic either human or object. Usually the crowd model is developed base on the extracted information that represent the status either implicitly or explicitly while the event discovery is accomplished using the computational model. Both of models are updated with the new information extraction [5]. Fig 2 illustrates the framework of crowd analysis and its processing.

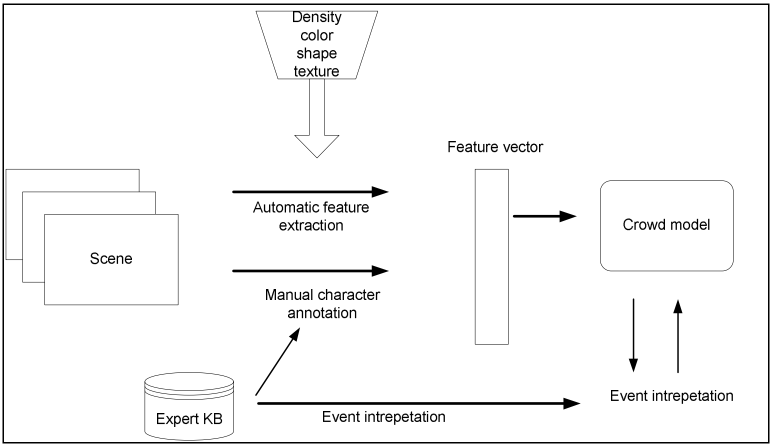


Fig. 2. A framework of crowd analysis

The potential of crowd analysis lend itself to a new application domain such as automatic detection of riots or chaotic acts in crowds and localization of the abnormal regions in scenes for high resolution. The common process for analysis in video sequence of crowd information extraction composed the following main three steps include [6, 9-11] i) Pre-Processing, ii) Object Tracking, iii) Event/Behavior Recognition. In addition, Microscopic, Macroscopic and Mesoscopic or Hybrid are the three main category modeling approaches which are familiar in the crowd [12]. Table 1 shows the description of categories of crowd scene.

Table 1. Terms and description in crowd analysis and modeling

Terms	Description
Crowd Density Estimation	To measure a crowd status. Find out the level of the crowd in a space or to detect abnormal changes of the crowd overtime
Crowd Motion Detection	To describe the characteristic of a crowd. Identify pattern of behavior in crowd
Crowd Tracking	To acquire the trajectories of movement. Determine whether abnormalities occur.
Crowd Behavior Recognition	To analyze the behavior of the crowd. Extract motion information and Model abnormal crowd
Structured Crowded Scene	Crowd moves coherently in common direction, motion direction does not very time, each spatial locations of the scene supports only one dominant crowd behavior over the time. *Example:* Marathon race, queues of people event, traffic on the road.
Unstructured Crowded Scene	Crowd motion is appeared random, different participants moving different direction at different times, each spatial location supports multi –modal, crowd behavior. *Example:* People walking on a zebra crossing in opposite directions, exhibitions, sporting event, railway stations, airport, motion biological cells.
Pre-Processing	*Responsibility* : Detect and classify *Category :* Rigid object or Non-Rigid Object *Analysis/ Features/ Approach :* Pixel Based Analysis, Texture Based Analysis, Region Based Analysis, Frame Based Analysis. *Example :* Feature extraction (foreground detection, optical flow), object detection, classification (color, edge, shape, head, body)
Object Tracking	*Responsibility :* Analyze target movement *Category:* Tracking individual objects and tracking the group of object. *Analysis/ Features/ Approach :* Region-based, active contour-based, feature-based, model-based tracking *Example :* Tracking speed and direction
Event/Behavior Recognition	*Responsibility:* Analyze pattern or behavior of the object. *Category :* Individual or crowd behavior recognition *Analysis/ Features/ Approach :* Object approach, Holistic approach *Example:* Occlusion, moving object (running, walking, jumping).

Table 1. (*continued*)

Microscopic	Defines the object movement and treats crowd behaviors as a result of a self organization process.
Macroscopic	Focus on goal-oriented crowds which determined a set of group-habits based on the goals and destination of the scene.
Mesoscopic / Hybrid	Inherit from Microscopic and Macroscopic

4 Common Approaches in Crowd Analysis

Fig 3 below shows the common crowd analysis approaches that is used to detect abnormal behavior in crowd scene. Preprocessing consist of; pixel level analysis, texture level analysis; object level analysis; frame level analysis. Object tracking consist of; region based approach; active contour based approach; feature based approach; model based approach. Event/ behavior recognition consist of object approach and holistic approach.

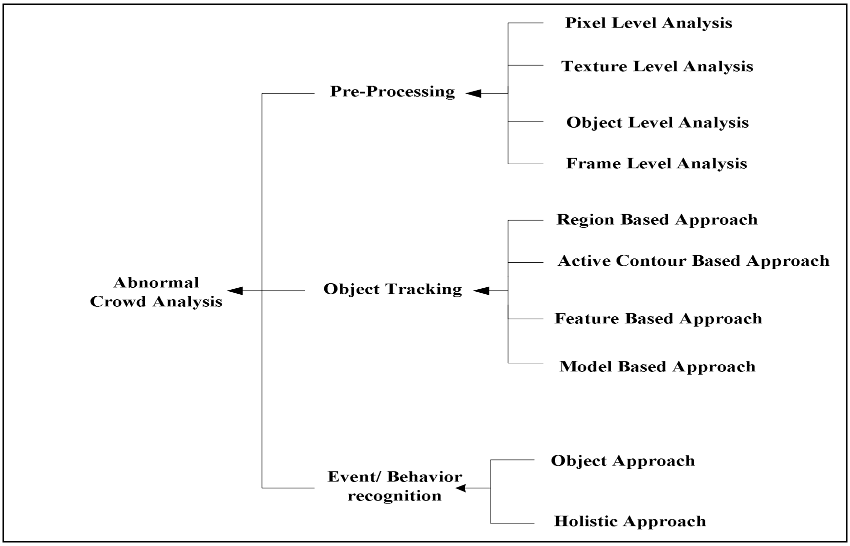


Fig. 3. Taxonomy approaches in crowd analysis

4.1 Preprocessing

One of the important steps in pre-processing is feature extraction. Feature extraction always deal with the crowd density which is very useful in source information. Due to the strength of the feature extraction in detection, most of the researchers are intended to analyze and learn the pattern of abnormal in crowd scene through this analysis. The common analysis uses during the preprocessing are pixel level analysis, texture level analysis, object level analysis and frame level analysis. *Pixel level analysis* is obtained through edge detection or background/foreground subtraction. Mostly focus on a low level features where extract the information based on density estimation rather than counting. For example Andrade and colleagues [9] used Expectation Maximization (EM) algorithm by analyzing pixel based to determine the variables and to update the equation of the probability distribution function. In Xiaogang and colleagues [13] works, simple "atomic" activities and interaction in low level features is proposed for unsupervised learning framework. Both activities and interaction are clustered into different class (eg : moving pixel- atomic; short video clips - interaction). The solution such as transparent, clean and probabilistic are formulated for solving the surveillance issues. *Texture level analysis* is similar like pixel level analysis that is used to estimate the number of people rather than identifying individual in a scene. The analysis of image patches is required for modeling and mostly focus on high level features. For example Xinyu and colleagues [14] is analyzing the texture based on the contour for human blob in order to learn different scale of group using Gauss Lapcian kernel function. Another author that learning the abnormal detection based in texture is Kilambi and colleagues [15] by learning the shape model to estimate the accurate people in the scene *Object level analysis*: is identifying individual object in a scene. More accurate result will be produced when compared to pixel and texture analysis. For example, in Khansari and colleagues [16] works, the direction and speed of the object motion is found, and inter frame texture analysis is adapted for the searching window. To perform the best matching region, frame is successfully generated with feature vector in a search window. *Frame level analysis* model behaviors of the full scene within the field of view of a camera. For example, in Oliver and colleagues [17], robust 2D blob features is presented in which the Eigenspace describes the appearance in a covariance data whilst principal component analysis is used to reduce the dimensionality space. However, eigenspace could not distribute the moving object efficiently in the background. Therefore, to get more accurately detection of abnormal to characterize the shape of each person, the portions of containing image moving present in the scene is well managed using frame by frame examination.

4.2 Object Tracking

The next steps after extracting the features from the image sequence is object tracking. Object tracking in a crowd attempt to minimizing the constraint such as occlusion, color intensity, illumination condition, appearance and etc. Past few years, multiple human objects tracking approach has been applied by researcher for recognize and detecting the behavior in a crowd, which is consist of identifying the position of each person in the same video sequence. There is various effective parameters ways. For example color, trajectory, body contour (head, hand, foot), and etc. Color distribution is commonly used in tracking to differentiate the object in a crowd [18] [16] [19]. Note that, it is easier way to track and understand the behavior of the people in crowd rather that individually as long as their moving in the same direction. By assigning the object distance , the occlusion in a crowd could easily track independently, as in [18]. However, if the pixel could not classify in the object, it is difficult to find the reliable central of the occlusion as long as the presence probability is updated for every pixel in the frame correctly.

Categories of object tracking approach include region based approach, active contour based approach, feature based approach and model based approach. *Region Based approach* is a robust computer vision in unconstrained crowd scene which is the information such as density, direction and velocity is extracted using optical flow technique. For example Weighted Maximum Cardinality Matching scheme with disparity estimation technique are presented by Kellly and colleagues [19] to evaluate both environment condition either indoor and outdoor (eg : varying cloud clover, shadows, reflections on windows and moving background). *Active Contour Based approach* is used to model the target partial occlusion and to extent some noise. Typically has been used a color histogram, however the weaknesses by using this technique is hardly change the color histogram when impair with similar object such as head in a crowd Khansari and colleagues [16]. *Feature based approach* is presented in feature image by describing the blob level feature. The examples are size, shape, elongatedness, luminance histogram and displacement histogram. Each feature image is transform from original blob level features into probabilistic appearance manifolds for each class. For example Peng and colleagues [20] introduces a pixel wise to detect anomalies in individual events sequence by constructing feature images. Each feature image is transform from original blob level features into probabilistic appearance manifolds for each class. *Model based approach* can solve blob merge and split constraint. This approach is used to segment and track multiple people occlusion. Bottom up image analysis is used to improve efficiency in computer vision. For example Yao-Te and colleagues [18]introduces model based object to estimate the positions corresponding to optical segmentation, which allow multiple object to be detected and track in crowded scene. To differentiate each object, color model of two region of body is presented called torso and bottom.

4.3 Event/Behavior Recognition

Another important process in a crowd analysis is event/behavior recognition. It can be characterized by regular motion patterns such as direction, speed, etc [6]. In the early work, crowd behavior analysis has been attempted in research topic of the computer vision especially in simulation [21-24] and graphic field. Monitoring and modeling the crowd is not so much to analyze normal crowd behavior, but to detect something different behavior from it. These are referred to as anomalous or abnormal.

Two types of approach are commonly used in this analysis includes object based approach and holistic based approach. *Object based approach* means a crowd is analyzes by treating a collection of individual to estimate the velocities, direction and abnormal motion. The complexity occur when the occlusion exist that maybe could be affected the process of analyzing such as detection of object, tracking trajectories and recognizing activities in a dense crowd. For example Weina Ge and colleagues [25]. Jacques and colleagues [26] use a position of each individual in parameter to obtain and characterize (voluntary or involuntary) the formation in a group and Voronoi diagram was used to understand people motion. Two approaches were proposed include feature correlation and binary function. Feature correlation was use to find the approximate position of the center of head while binary function is defined to represent distance between agents. *Hollistic based approach* means a crowd is analyzes by treating a single entity to estimate the velocities, direction and abnormal motion. The analysis covers medium to high density scene in global entity. For example Mehran and colleagues [12] integrates holistic approach with a particle advection method. They underlie the flow field with social force to extract interaction to determine the change interaction of time of behavior of the crowd for mapping to the image frame. However, using holistic approach application is still have weaknesses because in the dense crowd image of the object have a low resolution; and consists of dynamic and static occlusions. Thus, to get more accurate estimation parameter, the object based approach is still better.

5 Applications of Crowd Analysis

Crowd analysis becomes the most oriented research and trendy topic in computer vision and pattern recognition nowadays. The crowd phenomenon has been growth along urbanization frequently which gives the great interest in a large number of applications such as crowd management, public space design, virtual environments, visual surveillance, and intelligent environments (Fig 4) [5] and Table 2 describes the applications of crowd analysis.

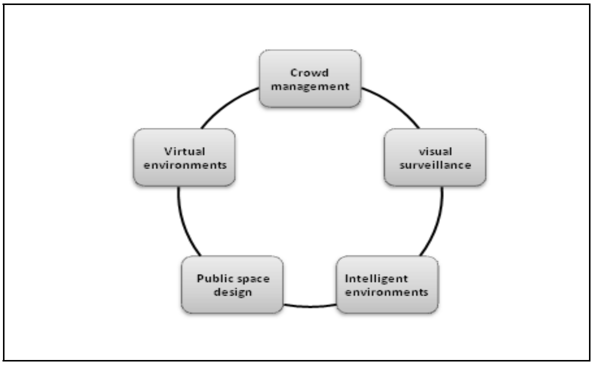


Fig. 4. Application of crowd analysis

Table 2. Applications of Crowd Analysis

Application	Description
Crowd management	Consist of developing crowd management strategies especially for increasingly more frequent and popular events like sport matches, concert events, public demonstrations and etc in order to avoid crowd disasters and ensure the public safety. Crowd management mostly studied by the sociologist, psychologist and civil engineers.
Virtual environments	Consist of mathematical models of crowds can be employed in virtual environment in order to enhance the simulations of crowd phenomena, to enrich the human life experience. Virtual environment are commonly studied by the computer graphic researchers.
Visual surveillance	It is used to detect anomalies and alarms automatically. Virtual surveillance is commonly studied by computer vision.
Intelligent environment	Involve a pre-requisite for assisting the crowd or an individual in the crowd. For example how to divert a crowd based on the pattern of crowd in a outdoor environment like parking lot.
Public space design	Consist of guidelines for the design of public space for example to optimize the space usage of an office.

6 Conclusion

In this paper, we present the state of the art of crowd analysis and the taxonomy of the common approach of the crowd analysis that could be useful to researchers and would serve as a good introduction related to the field undertaken. Crowd analysis is the most important concept for understanding the behavior especially in analyzing the abnormal behavior in a crowded scene. Three processing steps are involved in crowd analysis: preprocessing, object tracking and event/behavior recognition. Every step in crowd analysis has different analysis such as for pre-processing phase: the analysis for pixel level analysis, texture level analysis, object level analysis and frame level analysis is different. While for object tracking phase, which includes region based approach, active contour based approach, feature based approach and model based approach, the performance are measured differently. Finally, for event/behavior recognition phase, the object based approach and holistic based approach scene is normally used to probe the significant of the method. Most of the researchers used these procedures to analyze, detect and recognize the abnormal in crowded scene. However, what we have indentified here, only a little attention has been paid to learning the motion pattern in a crowded scene despite its importance in video surveillance in which reliable track are harder to obtain. For future work, we will explore and classify the variant learning based motion pattern in a crowded scene.

Acknowledgement

This work is supported by Ministry of Higher Education (MOHE) under fundamental Reseach Grant Scheme (Vote No 78509) and Zamalah Scholarship, Universiti Teknologi Malaysia (UTM). Authors would like to thank *Soft Computing Research Group* (SCRG), for their continuous support and motivation in making this research a success.

References

1. Yufeng, C., Guoyuan, L., Ka Keung, L., Yangsheng, X.: Abnormal Behavior Detection by Multi-SVM-Based Bayesian Network. In: International Conference of Information Acquisition, ICIA 2007 (2007)
2. Husni, M., Suryana, N.: Crowd event detection in computer vision. In: 2nd International Conference Signal Processing Systems, ICSPS (2010)
3. Saxena, S., Brémond, F., Thonnat, M., Ma, R.: Crowd Behavior Recognition for Video Surveillance. In: Blanc-Talon, J., Bourennane, S., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2008. LNCS, vol. 5259, pp. 970–981. Springer, Heidelberg (2008)
4. Jacques Junior, J.C.S., Mussef, S.R., Jung, C.R.: Crowd Analysis Using Computer Vision Techniques. IEEE Signal Processing Magazine 27(5), 66–77 (2010)
5. Zhan, B., Jacques Junior, J.C.S., Mussef, S.R., Jung, C.R.: Crowd analysis: a survey. Machine Vision and Applications 19(5), 345–357 (2008)
6. Garate, C., Bilinsky, P., Bremond, F.: Crowd event recognition using HOG tracker. In: Twelfth IEEE International Workshop Performance Evaluation of Tracking and Surveillance, PETS-Winter (2009)
7. Davies, A.C., Jia Hong, Y., Velastin, S.A.: Crowd monitoring using image processing. Electronics & Communication Engineering Journal 7(1), 37–47 (1995)
8. Rodriguez, M., Ali, S., Kanade, T.: Tracking in unstructured crowded scenes. In: IEEE 12th International Conference of Computer Vision (2009)
9. Andrade, E.L., Blunsden, S., Fisher, R.B.: Hidden Markov Models for Optical Flow Analysis in Crowds. In: 18th International Conference Pattern Recognition, ICPR 2006 (2006)
10. Andrade, E.L., Blunsden, S., Fisher, R.B.: Modelling Crowd Scenes for Event Detection. In: 18th International Conference Pattern Recognition, ICPR 2006 (2006)
11. Andrade, E.L., Fisher, R.B., Blunsden, S.: Detection of Emergency Events in Crowded Scenes. In: The Institution of Engineering and Technology Conference on Crime and Security (2006)
12. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: IEEE Conference Computer Vision and Pattern Recognition, CVPR 2009 (2009)
13. Xiaogang, W., Xiaoxu, M., Grimson, W.E.L.: Unsupervised Activity Perception in Crowded and Complicated Scenes Using Hierarchical Bayesian Models. IEEE Transactions Pattern Analysis and Machine Intelligence 31(3), 539–555 (2009)
14. Xinyu, W., Guoyuan, L., Ka Keung, L., Yangsheng, X.: Crowd Density Estimation Using Texture Analysis and Learning. In: IEEE International Conference Robotics and Biomimetics, ROBIO 2006 (2006)
15. Kilambi, P., Masoud, O., Papanikolopoulos, N.: Crowd Analysis at Mass Transit Sites. In: Intelligent Transportation Systems Conference, ITSC 2006. IEEE, Los Alamitos (2006)

16. Khansari, M., Rabiee, H.R., Asadi, M., Ghanbari, M.: Crowded scene object tracking in presence of Gaussian White noise using undecimated wavelet features. In: 9th International Symposium Signal Processing and Its Applications, ISSPA 2007 (2007)
17. Oliver, N.M., Rosario, B., Pentland, A.P.: A Bayesian computer vision system for modeling human interactions. IEEE Transactions Pattern Analysis and Machine Intelligence 22(8), 831–843 (2000)
18. Yao-Te, T., Huang-Chia, S., Chung-Lin, H.: Multiple Human Objects Tracking in Crowded Scenes. In: 18th International Conference Pattern Recognition, ICPR 2006 (2006)
19. Kelly, P., O'Connor, N.E., Smeaton, A.F.: Robust pedestrian detection and tracking in crowded scenes. Image and Vision Computing 27(10), 1445–1458 (2009)
20. Peng, C., Li-Feng, S., Zhi-Qiang, L., Shi-Qiang, Y.: A Sequential Monte Carlo Approach to Anomaly Detection in Tracking Visual Events. In: IEEE Conference Computer Vision and Pattern Recognition, CVPR 2007 (2007)
21. Thalmann, D., Musse, S.R., Musse, S.: Relating Real Crowds with Virtual Crowds, in Crowd Simulation, pp. 125–148. Springer, London (2008)
22. Reynolds, C.: Steering Behaviors for Autonomous Characters. In: Game Developers Conference (1999)
23. Vigueras, G., Lozano, M., Orduña, J.M., Grimaldo, F.: A comparative study of partitioning methods for crowd simulations. Applied Soft Computing 10(1), 225–235 (2010)
24. Bruno, L., Tosin, A., Tricerri, P., Venuti, F.: Non-local first-order modelling of crowd dynamics: A multidimensional framework with applications. Applied Mathematical Modelling 35(1), 426–445 (2010)
25. Ge, W., Collins, R.T., Ruback, B.: Automatically detecting the small group structure of a crowd. In: Workshop Applications of Computer Vision, WACV (2009)
26. Jacques, J., Braun, A., Soldera, J., Musse, S., Jung, C.: Understanding people motion in video sequences using Voronoi diagrams. Pattern Analysis & Applications 10(4), 321–332 (2007)

Iris Segmentation: A Review and Research Issues

Abduljalil Radman, Kasmiran Jumari, and Nasharuddin Zainal

Department of Electrical, Electronic & Systems Engineering
Faculty of Engineering & Built Environment
National University of Malaysia
Bangi, Malaysia
abdu_rad@yahoo.com, kbj@eng.ukm.my, nash@eng.ukm.my

Abstract. In recent years, due to the crime spreading and the potential of security threats, a lot of efforts have been spent to develop a reliable surveillance system. Biometric systems have grown in popularity as a trustworthy way to authenticate the identity of individuals. There are several biometrics to identify the persons such as Fingerprint, Face, Iris, Ear and Gait. The iris biometric is considered the best biometrics, due to its speed and accuracy in the identification, and distinctive features along the individual's lifetime. This paper focuses on issues of the iris segmentation research in the existing techniques. The techniques reported here can be categorized based on image acquisition environment into two scenarios: close-up imaging settings and non-cooperative imaging settings. The theorem behind each strategy is presented along with the outlines of implementation. A review of the issues in the iris segmentation methods coupled with comparison to the relevant issues is included. The weakness of each method is also obtained.

Keywords: Biometrics, iris segmentation, non-cooperative iris recognition, active contour.

1 Introduction

At present, one needs a pin to access an ATM machine, a password to log onto notebooks or database access, identification card to pass through restriction areas in the airport and so on. In this context, great interest has been given to implement reliable biometric identification systems, in order to replace the conventional security systems. The biometric measurements offer excellent secure and reliable accessibility; and less likely to be fraud as credit cards, hacking or forgetting as passwords, or stolen as identification cards. The iris biometric provides very high accuracy for personal identification; furthermore, the quantity and unique information which can be measured in a single iris are much greater than other biometrics' information. Therefore, much effort has been exhausted for developing a robust iris recognition system in the recent years.

In spite of difference between the proposals' particulars, the iris recognition systems share in four typical stages (Fig. 1). Start with *iris segmentation*: locating iris in the eye image. Specifying the iris' inner (pupillary) and outer (limbic) boundaries

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 698–708, 2011.

is the core goal for this stage; after that, eliminating undesired data caused by the occlusion of eyelids, eyelashes, and reflections. *Normalization*: transferring the segmented iris data into a fixed length coordinate system, in order to avoid the pupillary dilation and imaging distance. *Feature Extraction*: the unique features of the iris are extracted and encoded. *Finally*: comparing the iris' signature is generated in the prior stage with the registered iris' signatures in the database, to decide if they are belonging to the same or different irises. This paper is intended to offer an extensive criticism for the iris segmentation techniques.

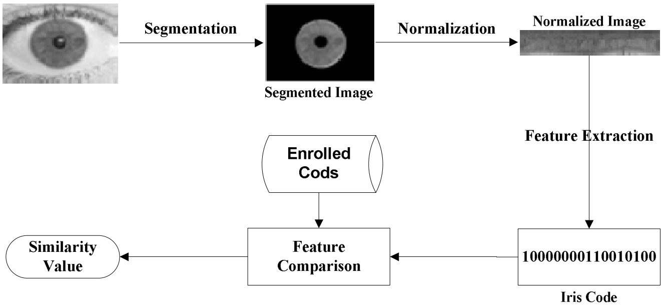


Fig. 1. Traditional phases of the iris recognition methods

1.1 The Path Forward for Iris Recognition Systems

The vantage of deployed the current iris recognition systems, returns to Daugman's system [1]. The largest part of these systems has proven under relatively rigid constraints; such as, capturing the images from closeness with a stop-and-stare interface; good lighting is utilized to attain sufficient contrast and discriminated iris features. Along with all aforementioned constraints, the subject's cooperation represents a crucial factor for the system success. In such systems, the near-infrared (NIR) illumination sources have been used. It provides good quality images but may hurt the eyes and causes permanent injury if the illumination is strong. In this paper "close-up system" refers to this kind of systems.

Often, the image should be taken under unconstrained conditions and without the subject's knowledge; for instance, in case of tracking a criminal. Therefore, recently the research path veered toward capturing the images under the visible wavelength (VW) light, on-the-move, and at-a-distance conditions (Non-cooperative system). This made the iris recognition systems more suitable for practical applications such as forensic and security. However, capturing images under this type of lighting enables much higher level of details; on the other hand, the quality of captured data is degraded. Furthermore, many noisy artifacts introduced namely the specular and light reflections. Compared with images acquired under the NIR wavelength, the sclera spectral reflectance is significantly higher. Acquiring images under on-the-move and at-a-distance circumstances (non-cooperative environment) leads to more undesired data (noise); such as, non-completed, closed eye, off-angle and poor focused iris images. Generally, the images that have been acquired in non-cooperative settings are noisier and considerably need more exertion to analyze. Table 1 provides a brief comparison of the two image acquisition scenarios in the existing iris recognition systems.

Table 1. Brief comparison of the two image acquisition environments

Parameter	Close-up	Non-cooperative
Illumination	NIR	VW
Subject's cooperation	Stop-and-stare	On-the-move and At-a-distance
Image quality	Good	Low
Feasibility	Low	High
Noise	Limited	High

Before presenting the previous work in the literature of the iris segmentation, it will be appropriate to discuss the frequently used terms in this context. Pupillary boundary is the inner border of the iris which located between the iris and pupil regions. Limbic boundary refers to the outer border of the iris region which located between the iris and sclera regions. Both define the iris spatial extent in the 2-D eye image. Pupillary zone refers to the circular aperture close to the iris center (Fig. 2).

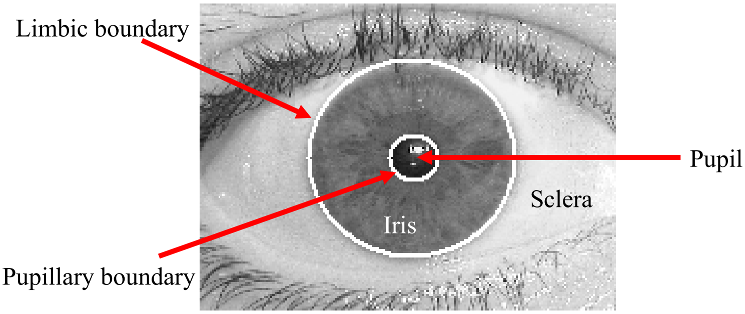


Fig. 2. Iris image structure

Lighting and specular reflections: this type of noises is caused by illumination sources. The former corresponds to the reflections from light sources near to the user, when the lighting is strong; whilst the second corresponds to the reflected objects from the environment that the subject surrounded by, when the illumination is not enough strong (Fig. 3).

The reminder of this paper is organized as follows. Section 2 and 3 present the early development of iris segmentation techniques related to one of the two imaging environments: close-up imaging environment or non-cooperative imaging environment. The grand challenges of iris segmentation research are presented in section 4. Section 5 reviews, the public and freely available iris image databases and their characteristics. Finally, section 6 discusses and concludes this paper.

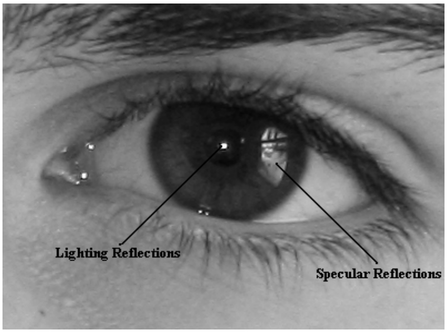


Fig. 3. Iris image containing lighting and specular reflections

2 Related Work in Close-Up Imaging Environment

2.1 Daugman's Approach

The most cited approach in the iris segmentation literature is that of Daugman [1]. It is the first algorithm to verify the people by their iris patterns. Daugman pioneered the integro-differential operator to detect the iris and pupil boundaries. This operator estimates the boundaries of the pupil and limbus as circles; each circle is defined by three parameters; the radius r and center coordinates (x_0,y_0). The integro-differential operator is defined as:

$$\max_{(r,x_0,y_0)}|G_\sigma(r) * \frac{\partial}{\partial r}\oint_{r,x_0,y_0}\frac{I(x,y)}{2\pi r}ds| \qquad (1)$$

where I(x,y) is an eye image. The operator searches for the maximum in the blurred partial derivative over the image domain (x,y) with respect to increasing radius r of the normalized contour integral of I(x,y); along an annular curve ds of radius r and center coordinates (x_0,y_0). The symbol * denotes convolution and $G\sigma$ is a smoothing function such as Gaussian of scale σ [1]. Daugman modeled the upper and lower eyelids as parabolic arcs [2-3]. When the coarse to fine iterative searches for inner and outer iris boundaries have reached single pixel precision; the same operator is used to localize eyelids' border, with adapting the search contour from circular to curvature. Daugman's algorithm produces no false matches in the close-up iris images; on the other hand, it frequently fails especially with the images that do not have sufficient contrast between the iris and sclera regions. The Daugman's operator presumes that the iris and pupil have circular boundaries; this was a source of error, where the pupillary and limbic borders are not always exactly circular. Furthermore, it endures a heavy computation.

2.2 Wildes' Method

One of the early history methods of the iris segmentation is that for Wildes [4]. Wildes used gradient maps with the Hough transform as a geometric fitting algorithm, to fix the two circular boundaries of the iris. The proposed method begins by obtaining the edge maps of the image; which consist of Gaussian kernel convolved with the magnitude of the image intensity gradient:

$$|\nabla G(x, y) * I(x, y)| \tag{2}$$

the limbus and pupil borders were assumed as circular shapes whilst the upper and lower eyelids were represented with parabolic curves. They were fixed by the same manner; the only difference in eyelids fitting, it votes for parabolic arcs instead of circles. Wildes' approach is steadier against the noise fluctuations. However, the Hough transform requires threshold values to perform the edge detection; as a result, significant edge-points can be removed and could cause failures to fit the contours. It requires large memory and causes slowness because its complex computation. Consequently, may not be appropriate for real time implementations.

Daugman's [1-3] and Wildes' [4] systems achieve accurate recognition in situations of high quality image. On the other hand, they do not take into account the non-iris regions, and such regions could lead to generate wrong biometric templates; consequently, poor recognition rates.

2.3 Others Methods

By combining the integro-differential operator [1-2] and Hough transform [4] algorithms, Tisse et al. [5] introduced a new method for iris segmentation. Ferreira et al. [6] altered the Daugman's operator [1-2] to fit the noisy iris images. The modification was based on the study of the databases' images; moreover, techniques for taking off the reflections and enhancing the pupil isolation were proposed. Based on thresholding technique, which produces a binary image from grayscale image; Lya Liam et al. [7] have calculated a threshold value to join the iris and pupil in one isotropic coarse region. Next, they formulated a ring mask and applied it on entire image to search for iris and pupil borders; which were assumed as circular shapes.

All iris segmentation techniques are described previously hypothesized that the pupil and iris boundaries are circular; therefore, their parametric models designed to identify the boundaries based on this basis. After detecting the iris' inner and outer boundaries, the noise generated by eyelashes and eyelids ought to be removed. This means more algorithms for different stages should be designed to do the segmentation. Hence, the active contour models turned out to be more attractive to accomplish such applications; where, the methods based on active contours have more flexibility in determining the shapes. This was the motivation of Daugman's work [8] for iris segmentation based on active contours. Based on this technique, Daugman opened the door for a new technique in iris segmentation; which, allows flexible shapes and coordinates. This algorithm contributes in the iris segmentation even with those occluded by eyelids and eyelashes. Ritter et al. [9] made use of active contour models for revealing the iris boundaries in the eye image. In other approaches [10-11], the active contour model based on the snake model is used for identifying pupil/limbic limits. Different approach for iris segmentation was suggested by Cui et al. [12]; the low frequency information of the wavelet transform is used for pupil detection, and the iris' boundary identified by the integro-differential operator [1].

3 Related Work in Non-cooperative Imaging Environment

3.1 Proenca and Alexandre's Technique

Similar to Wildes' method [4], Proenca and Alexandre proposed a method [13] for non-cooperative iris segmentation. A clustering process was used to increase the robustness against the noisy images. The clustered image is utilized to construct the edge-maps instead of use the original image; this is the difference between this method and Wildes' method. Next, the circular Hough transform is used to specify the pupil and limbic boundaries. Similar to Wildes' method, this approach suffers from the threshold values and intensive computation. Furthermore, circular boundaries for the iris were assumed just like Daugman's and Wildes' algorithms; eyelids and eyelashes detection did not take into account as well. Additionally, it should be pointed out, as far as this method was designed for non-cooperative iris segmentation the off-angle problem did not address; where, the off-angle situation frequently happened in non-cooperative iris recognition.

3.2 Jeong's et al. Approach

The authors expressed an algorithm [14] for isolating the iris' subject from noisy iris images. They used the Daugman's integro-differential operator (with a little modification to evade the eyelids' occlusion) as a circular edge detector to attain the limbic and pupillary boundaries. Often, the specular reflections mislead the edge detectors; hereby, the interpolation technique has been used to remove the discontinuous boundary caused by reflections. However, in case of failure; the eye detection algorithm using AdaBoost comes up to replace the former method. The authors classified the iris image which has no corneal specular reflections as a ''closed eye" image. After revealing the iris site precisely; the cross points between the iris' outer boundary and both the upper and lower eyelids, eyelid detection masks, and the parabolic Hough transform are utilized to detect the eyelids' boundaries. Next, the obstructions resulted by ghosting of the visible light are detected using the color segmentation strategy. Based on the detected iris, pupil, and eyelids regions; the authors used an eyelash-detecting mask to separate the eyelashes. In spite of the AdaBoost classifier algorithm carries out fine classification performance, it consumes lots of time for training; this designates a shortcoming for this method.

3.3 Labati and Scotti's Algorithm

Beginning with estimation of the pupil and iris centers, Labati and Scotti offered an algorithm [15] for iris segmentation in degraded images. Briefly, from our viewpoint their approach roughly combines between the integro-differential operator [1], and the new iris segmentation method of Daugman [8], which is based on the active contour model. The limitation of this method is the initial estimation of the iris boundaries' centers; where, poor initialization of centers/radii of the iris boundaries could lead to complete failure in the segmentation outputs. Even though this method may allow precise segmentation, it does not identify the eyelashes' region precisely; as a result, the iris border on the upper side drastically overestimates.

3.4 Li's et al. Method

The authors nominated an algorithm [16] to segment the iris image in non constrained imaging environment. At the beginning, they restrict the size and position of the eye; then, trace the limbus and pupil boundaries inside this region. Due to the importance of the iris segmentation stage, the authors suggest two methods. First, the Hough transform has been improved and merged with the K-means (based on the gray-level co-occurrence histogram) clustering algorithm to disclose the iris limits. Second, in the case of failure, the method which uses skin information to confine the eye area was suggested. Li's algorithm uses many threshold values; as a result, inaccurate choosing for any one of those values can cause an error in the subsequent localization. Furthermore, a complex algorithm for eyelid detection has been used; consequently, more time is required for fulfilling the iris segmentation.

3.5 Tan's et al. Technique

Based on the integro-differential operator [1], an approach for non-cooperative iris segmentation was proposed by Tan's et al. [17]. The first goal for this method was to cast off the noise artifacts produced by specular reflections from the image. Next, the position of the coarse iris region was approximated by a clustering scheme. Always, in the iris segmentation methods, eyelids and eyelashes are classified as a non-iris region; hereby, a curvature and prediction models were proposed to discover those regions. The iris' inner and outer boundaries were ascertained by the well-known integro-differential operator with minor adaptation against the various noise artifacts. In spite of the good progress that accomplished, the segmentation failed with noncircular iris.

3.6 Proenca' New Method

Proenca pioneered a technique [18] which can handle iris' subject from noisy iris images captured under less controlled circumstances. The author took advantage of the sclera region which is the most distinguishable part of the eye in the degraded images. Next, the adjacency of the iris and sclera is exploited to detect the noise-free iris regions; the Neural Network is utilized for this purpose.

Generally, most of the iris segmentation approaches in the literature depend on one of the three basic techniques: integro-differential operator, Hough transform, or Active contour. Table 2 provides the reader an overview of the different aspects of these three technologies.

Table 2. Brief comparison among, integro-differential operator, Hough transform, and active contour techniques

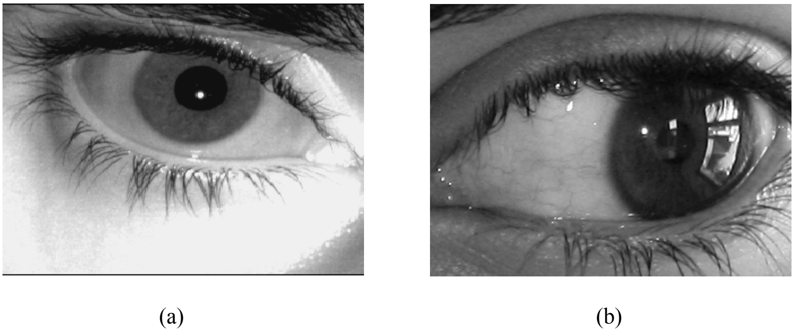
Parameter	Integro-differential	Hough transform	Active contour
Complexity	Low	High	Medium
Speed	Medium	Low	Low
Memory	Low	High	Medium
Accuracy	High	Medium	Low
Noise sensitivity	Medium	Low	High

4 Iris Segmentation Delimitations

Every research has challenges; some of the technical barriers to develop a good iris segmentation algorithm can be presented as follows. Firstly, illumination reflections which generated by the light sources; where the strong lighting reflects maximum values on the reflection area, or reflects some objects that the user surrounded by if the lighting is not strong enough; secondly, eyelids and eyelashes which sometimes occlude part of the iris patterns. Thirdly, the geometric issues are also challenging; the non-circular iris/pupil can affect the iris localization process ultimately. Good classifying for these issues can be found in [19].

5 Databases

The biometrics' research and development is demanded to be extremely accuracy; so, it is not logical to investigate the recognition algorithms with data captured on-the-fly. Standard biometrics' databases turn out to be very crucial in the development process; to allow fair comparison between recognition techniques. The databases in the iris biometric scope can be divided into two categories: *good quality iris image* databases, for instance, CASIA [20], IITD[21], ICE[22], MMU[23], and UPOL [24] databases; *noisy iris image* databases such as UBIRIS databases [25]. The foremost type contains images taken under controlled environment in terms of illumination sources and subjects' cooperation. These databases present quite homogenous and clearly images; moreover, their noise factors confined on iris obstructions by eyelashes and eyelids, poor focused images, and wittingly rotated images. In the second type, the images acquired under uncontrolled settings; and all the noise factors, including those mentioned in section 4 are induced, in order to produce an environment closer to reality (Fig. 4). Hereby, the databases in the first type suitable for evaluating the iris segmentation techniques in cooperative environments, whilst, the second type is nominated for non-cooperative iris segmentation researches.



(a) (b)

Fig. 4. Examples of iris images (a) acquired in NIR wavelength under constrained conditions, containing occlusion and wittingly rotated (MMU database [23]) (b) acquired in visible wavelength under uncontrolled imaging setting, containing off-angle rotation, illumination effects and occlusion (UBIRIS database [25])

6 Discussion and Conclusion

This paper summarizes the state-of-art iris segmentation techniques in the literature. Finding an iris in an eye image is the first step in iris recognition. Detecting the boundaries among pupil, iris, and sclera; segregating the occlusion caused by upper and lower eyelids if it is existed; and removing the non-iris regions may be caused by illumination reflections or eyelashes occlusion; all these processes cumulatively called iris segmentation. Inaccurate detection and representation of any boundary can lead to different representation of the iris pattern in its extracted features, and such difference could cause failures in the matching stage. The majority of relevant published iris segmentation strategies assumed that the iris has a circular shape, and significantly developed based on this assumption. Newly, one realized that inner and outer boundaries of the iris may not be perfect circular; moreover, iris and pupil are not concentric; but the pupil is fully inside the iris and both have closed curves. More research work should focus on improving disciplined algorithm for representing these boundaries whatever their contours. One of the most powerful ways to accomplish these objectives is active contours; which enable to confine irregular boundaries. On the other hand, methods based on an active contour are envisioned to consume long time to process; hence, much effort should be spent to develop a new algorithm to conform to the jagged contours.

The iris segmentation approaches are presented in the literature, have been proven on one of two environments. *The first*: close-up imaging distance, stop-and-stare interface, NIR illumination, in addition to subject's cooperation. The NIR illumination provides good quality images in terms of limited reflections on the sclera, and distinguished features; but extremely strong illumination may injure the eyes. Furthermore, not in all cases the subjects' cooperation is needed; where, sometimes the image should be taken without user's awareness; for instance, tracking criminals and terrorists. *The second*: non-cooperative imaging settings, without subjects' knowledge, from a distance, on the move, and under visible wavelength lighting. In contrast to the former type, it is more feasible and suitable for security purposes. On the contrary, the image quality is reduced; new artifacts noise also introduced, such as ghosts caused by light reflections. In addition, capturing an image under the visible wavelength light produces many spectral reflectances on the sclera. Moreover, acquiring images under on-the-move and at-a-distance circumstances may lead to non-completed, closed eye, off-angle and poor focused iris images. However, the iris images that have been acquired in non-cooperative settings are noisier and considerably need more effort to study.

Although, many attempts have been exhausted, and some progress has been made, the problem of automatic iris segmentation still far from being fully solved owing to its complexity. The factors, eyelids and eyelashes occlusion, eye position, closed eye, reflection and light conditions; undoubtedly affect the performance of iris segmentation algorithms. Consequently, there is lots of research yet to be done to make iris segmentation more feasible in less-controlled settings. In addition, further work should be done for isolating the noise regions which can lead to generate false biometric description; thus, poor recognition rates.

Acknowledgments. This research activity had been conducted in Computer & Network Security Laboratory, University Kebangsaan Malaysia (UKM). This project is supported by University Grant fund: UKM-AP-ICT-17-2009.

References

1. Daugman, J.G.: High Confidence Visual Recognition of Persons by a Test of Statistical Independence. IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 1148–1161 (1993)
2. Daugman, J.G.: High Confidence Recognition of Persons by Iris Patterns. In: The 35th International Carnahan Conference on Security Technology, pp. 254–263. IEEE, Los Alamitos (2001)
3. Daugman, J.: How Iris Recognition Works. IEEE Transactions on Circuits and Systems for Video Technology 14, 21–30 (2004)
4. Wildes, R.P.: Iris Recognition: an Emerging Biometric Technology. Proceedings of the IEEE 85, 1348–1363 (1997)
5. Tisse, C.-L.: Lionel Torres, Robert, M.: Person Identification Technique using Human Iris Recognition. In: the 15th International Conference on Vision Interface (2002)
6. Ferreira, A., Lourenço, A., Pinto, B., Tendeiro, J.: Tuning Iris Recognition for Noisy Images. In: Fred, A., Filipe, J., Gamboa, H. (eds.) BIOSTEC 2009. Communications in Computer and Information Science, vol. 52, pp. 211–224. Springer, Heidelberg (2010)
7. Lye Wil, L., Chekima, A., Liau Chung, F., Dargham, J.A.: Iris Recognition using Self-organizing Neural Network. In: Student Conference on Research and Developing Systemsm, SCOReD 2002, pp. 169–172. IEEE, Los Alamitos (2002)
8. Daugman, J.G.: New Methods in Iris Recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37, 1167–1175 (2007)
9. Ritter, N., Owens, R., Cooper, J., Van Saarloos, P.P.: Location of the Pupil-Iris Border in Slit-lamp Images of the Cornea. In: International Conference on Image Analysis and Processing, pp. 740–745 (1999)
10. Arvacheh, E.M.: Study on Segmentation and Normalization for Iris Recognition. Systems Design Engineering, MSc, p. 81. University of Waterloo, Waterloo (2006)
11. Liu, X.: Optimizations in Iris Recognition. Computer Science, Phd, p. 130. University of Notre Dame, Notre Dame (2006)
12. Cui, J.L., Wang, Y.H., Tan, T.N., Ma, L., Sun, Z.N.: a Fast and Robust Iris Localization Method Based on Texture Segmentation. In: Biometric Technology for Human Identification, pp. 401–408 (2004)
13. Proenca, H., Alexandre, L.A.: Iris Segmentation Methodology for Non-Cooperative Recognition. IEE Proceedings - Vision, Image and Signal Processing 153, 199–205 (2006)
14. Jeong, D.S., Hwang, J.W., Kang, B.J., Park, K.R., Won, C.S., Park, D.K., Kim, J.: a New Iris Segmentation Method for Non-ideal Iris Images. Image and Vision Computing 28, 254–260 (2010)
15. Labati, R.D., Scotti, F.: Noisy Iris Segmentation with Boundary Regularization and Reflections Removal. Image and Vision Computing 28, 270–277 (2010)
16. Li, P.H., Liu, X.M., Xiao, L.J., Song, Q.: Robust and Accurate Iris Segmentation in Very Noisy Iris Images. Image and Vision Computing 28, 246–253 (2010)
17. Tan, T.N., He, Z.F., Sun, Z.Z.: Efficient and Robust Segmentation of Noisy Iris Images for Non-Cooperative Iris Recognition. Image and Vision Computing 28, 223–230 (2010)

18. Proenca, H.: Iris Recognition: On the Segmentation of Degraded Images Acquired in the Visible Wavelength. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1502–1516 (2010)
19. Proenca, H.: Towards Non-Cooperative Biometric Iris Recognition. Department of Computer Science, Phd, p. 175. University of Beira Interior, Covilhã (2006)
20. 'Institute of Automation', Chinese Academy of Sciences' CASIA Iris Image Database (2004), http://www.sinobiometrics.com
21. 'Indian Institute of Technology Delhi', IIT Delhi Iris Database Version 1.0 (2007), http://web.iitd.ac.in/~biometrics/Database_Iris.htm
22. 'National Institute of Standards and Technology', Iris Challenge Evaluation Database (2006), http://iris.nist.gov/ICE/
23. 'Multimedia University', MMU Iris Image Database (2004), http://pesona.mmu.edu.my/~ccteo/
24. Dobes, M., Machala, L.: UPOL Iris Image Database (2004), http://phoenix.inf.upol.cz/iris/
25. Proença, H., Alexandre, L.: UBIRIS: A Noisy Iris Image Database (2005), http://iris.di.ubi.pt

Splitting Strategies for Video Images Processing in Medical Data Grid

Mien May Chong and Rohaya Binti Latip

Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
mienmay@hotmail.com, rohaya@fsktm.upm.edu.my

Abstract. There has a lot of improvements had been done on the diagnostic tools. Due to these improvements, most of the medical images are now migrating from 1D (e.g. cardiograms and encephalograms) and 2D (e.g. x-rays) images to extended 3D (e.g. tomography) and eventually 4D (3 spatial dimension + time) image. This migration process has let the volume of medical image become larger and become difficult to store. To overcome this kind of problems, most of the hospitals start to integrate the grid technology into their medical storage system. In this paper, besides presented the current related research projects in the medical field, we also briefly discussed on the details of medical data grid, and the strategies and techniques used in the video and image processing area. From the discussion, based on the parameters of the initial delay and the deadline miss, we found that the Fibonacci-based splitting strategy is the most appropriate strategy for used in the video images processing.

Keywords: Grid Computing, Medical Data Management, Digital Medical Images, Splitting Strategies, Image Processing, Video Processing.

1 Introduction

Grid computing is a collection of network connected computer resources that used by user to solve the large-scale of scientific, technical or business data or problems. Early of 1990s, “Grid Computing” terms had been interpreted as “Power Grid”. However, in year 1998, the term of “Grid” is defined by the Foster and Kesselman as “a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities” [1].

In the Grid environment, a virtual platform has been offered for the large-scale, resource-intensive, and distributed applications. Management and coordination of the various and scattered resources can be done due to its connectivity environment [2]. Grid enabled access to improve storage and computing capacity to provide a mechanism for sharing and transferring large-scale of data; while in other side, combine with distributed resources to conduct expensive computational procedures.

Nowadays, a lot of improvements have been done on the medical images, such as in the level of extracted data quality or its details. Moreover, the recent diagnostic tools also become more powerful, which allowing a migration from 1D

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 709–722, 2011.

(e.g. cardiograms and encephalograms) and 2D (e.g. x-rays) data to extended 3D (e.g. tomography) and eventually 4D (3 spatial dimension + time) data. Therefore, the volume of the medical data become huge and become not easily to store in database. Besides, users also need to use a lot of time to retrieve back the clearness original data from the database. Due to those reasons, recently, most of the hospitals are started to use the medical data grid for doing the analysis on the medical data.

In the other hand, grid computing use a common infrastructure based on open standards, therefore, it provides a platform for interoperability and interfacing between different Grid-based applications from the specific domain. Grid technology can potentially offer architecture to the medical applications, which it easy the user and access transparently to the distributed heterogeneous resources (e.g. data storage, networks, and computational resources) to across different organizations and administrative domains [2].

2 Backgrounds and Related Works

In this section, we briefly discussed on some latest related projects that use the grid technology in the medical area.

2.1 Biomedical Informatics Research Network (BIRN)

In medical area, there are many large-scale biomedical projects are launched by using the grid technology. One of the projects is under the Biomedical Informatics Research Network (BIRN). It is a national initiative that developed to assist the data sharing and online collaboration biomedical research. Besides, it also focuses on the neuroimaging studies [4]. Refer Figure 1 for the mapping of BIRN.

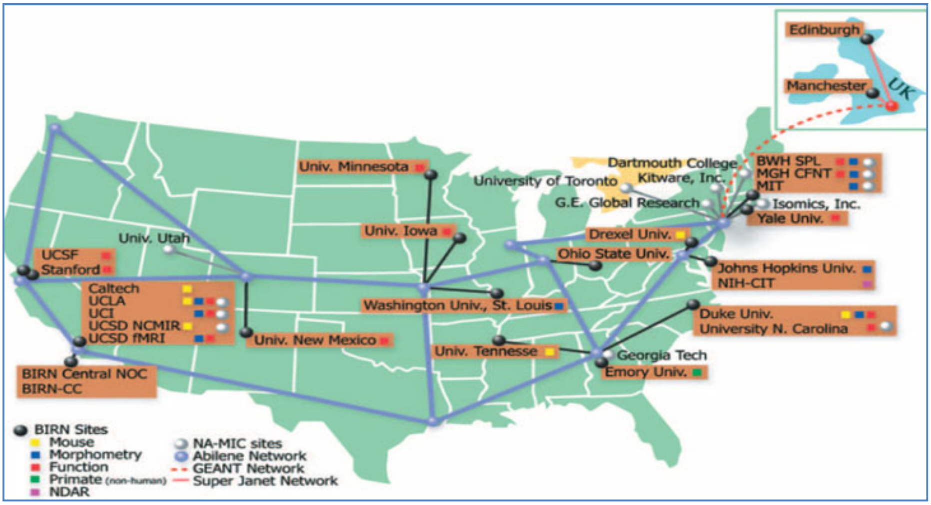


Fig. 1. Map of BIRN participating sites and BIRN Nodes [3]

A state-of-the-art open source toolkits and grid technologies built on the top of the high-speed internet-2/Abilene backbone has been used by the BIRN to develop and deploy a federated and distributed infrastructure for storage, retrieval, analysis, and documentation of biomedical data. BIRN is a virtual organization in sharing the resources. It have one central coordinating center, four test beds, and other associated collaborative projects such as National Alliance for Medical Imaging Computing (NAMIC) . The coordinating center is used to take care on the hardware, check on the security for grid middleware, manage the data, and computation, and also responsible on the data migration environment [3].

2.2 MammoGrid

For Europe Union, they have been funded a multi-institutional project called MammoGrid [4][5]. Inside the MammoGrid, there is a connection between the radiologist workstations across the grid. They use an "information infrastructure" and a DICOM-compliant object model residing in multiple distributed data stored in Italy and the UK [5][6]. It is aims to build a distributed database of mammograms by using the Grid middleware and tools. It also aims to prove that Grid infrastructures to facilitate the collaboration between the researchers and clinicians across the Europe Union [2]. Refer Figure 2 for the MammoGrid architecture.

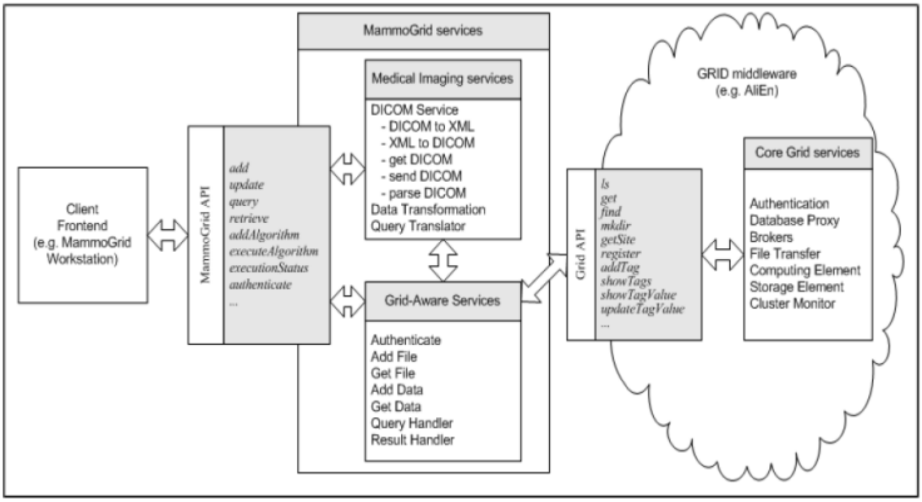


Fig. 2. High-Level of MammoGrid Architecture [2]

For MammoGrid, there is a clear need for an infrastructure, which is the large-scale of volume data can be associated with the regional breast cancer screening programs as well as can available across multiple medical centers in an acceptable time [2]. Standardization software is used for processing the image, thus, it can eliminate any biases introduced during the image acquisition. From this research, a comparison of images from different patients and centers can be done. Moreover, this kind of images must be made available to a data-mining engine, therefore, the queries based on patient details and text annotations can be resolved [2]. Besides, these kinds of images must also be accessible to analysis algorithms that offer the quantitative information.

2.3 MedIGrid

Another multi-institutional project is the MedIGrid [4]. MedIGrid is a project that funded by the German Federal Ministry of Education. As a member of German e-Science initiative, D-Grid, the MedIGRID is having permission on accessing the D-Grid resources by using the MedIGrid portal [7]. MedIGrid is used to investigate the application of Grid technologies for manipulating a large medical database. Refer Figure 3 for MedIGrid architecture. In MedIGrid, nuclear doctors can transparently use the high performance computers and storage systems for image processing and analysis.

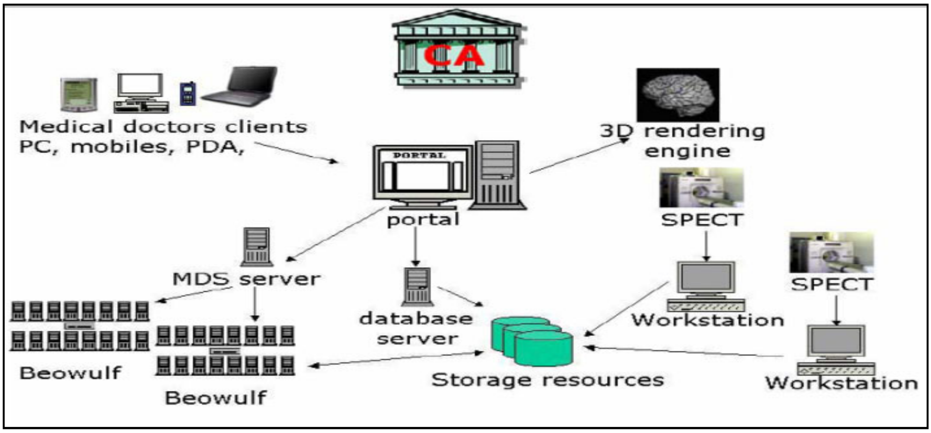


Fig. 3. MedIGrid architecture [7]

MedIGrid infrastructure is based on the grid middleware layer. In its infrastructure, Globus 4 had been used as its grid middleware. Specific virtualization services are implemented in the MedIGrid. Therefore, users are unable to know which resource data will be stored or which node is available to run the jobs. There are two types of virtualization, such as compute resource virtualization and data resource virtualization [8]. Compute resource virtualization is presented as the user shall not need to pick on the hardware and storage resources; while data resource virtualization is giving the user to access the data without showing the location of the data.

2.4 KnowledgeGRID Malaysia

In Malaysia, there is also a National Grid Computing Initiative, which known as KnowledgeGRID. KnowledgeGRID Malaysia was launched on 20th August 2007. The main objectives of the KnowledgeGRID is to provide Super Computing power to the Nation beyond the research communities, to provide highest level Cyberspace

Security for critical information, to achieve cost efficiency on Capital Investment for Knowledge Info Structure and also create new web services industries through pay-per-use [15].

For the KnowledgeGRID, a draft of national technology roadmap has been done. The objectives to develop this roadmap are to define potential key focus areas of the research in Grid Computing, to identify areas where Grid Computing can be applied, to further encourage multidisciplinary research and development, to establish a proper framework and policies and also to accelerate, educate and promote the potential use and benefits of Grid Computing to industries and public in general [15].

There are 4 main domains in National Grid Computing Roadmap, which are National Grid Facility, Info-structure and Security, Grid Middleware and Tools Enablers, Grid Applications and Policies and Governance [15]. Refer Figure 4 for the key domains for the National Grid Computing Roadmap.

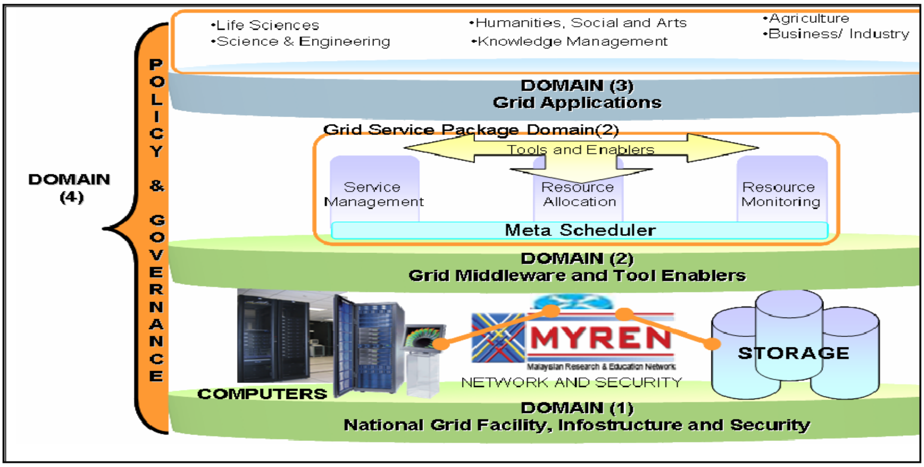


Fig. 4. Key domains for the National Grid Computing Roadmap [15]

In domain 1, it is the combination of the sub-domains of National Grid Facility, Info structure and Security. The core foundation will be formed in this domain for successful grid applications. The 2nd Domain is known as Grid Middleware and Tool Enablers domain. This domain is providing a standard platform for services to operate, coordinate and manage grid services for resource integration. This platform also used for monitoring the middleware level interface in Grid environment [15].

The 3rd domain is Grid Applications. This domain is concerned on the availability of Grid enabled applications. One of the sub-domains for this Grid Application domain is the Life Science area. This area includes the projects of bio-medical and bioinformatics. The classifications of the applications are as shown ass below:-

1) Short Term application (2007 -2008)
   - The applications have being used or in the prototype stage and ready to be grid enabled.

2) Medium Term application (2009 – 2010)
   - Application that important for Malaysia but the technologies and enablers need more time to develop and mature.

3) Future applications (beyond 2010)
   - Potential grid applications that may not have any mainstream player or the technologies do not exist yet.

The last domain for the National Grid Computing Roadmap is the Policy and Governance domain. It will addressed the issues on the policy and governance and make sure all the aspects in the roadmap is executed properly and must be in line with the nation's goals and needs. Refer Figure 5 for the National Grid Computing Roadmap [15].

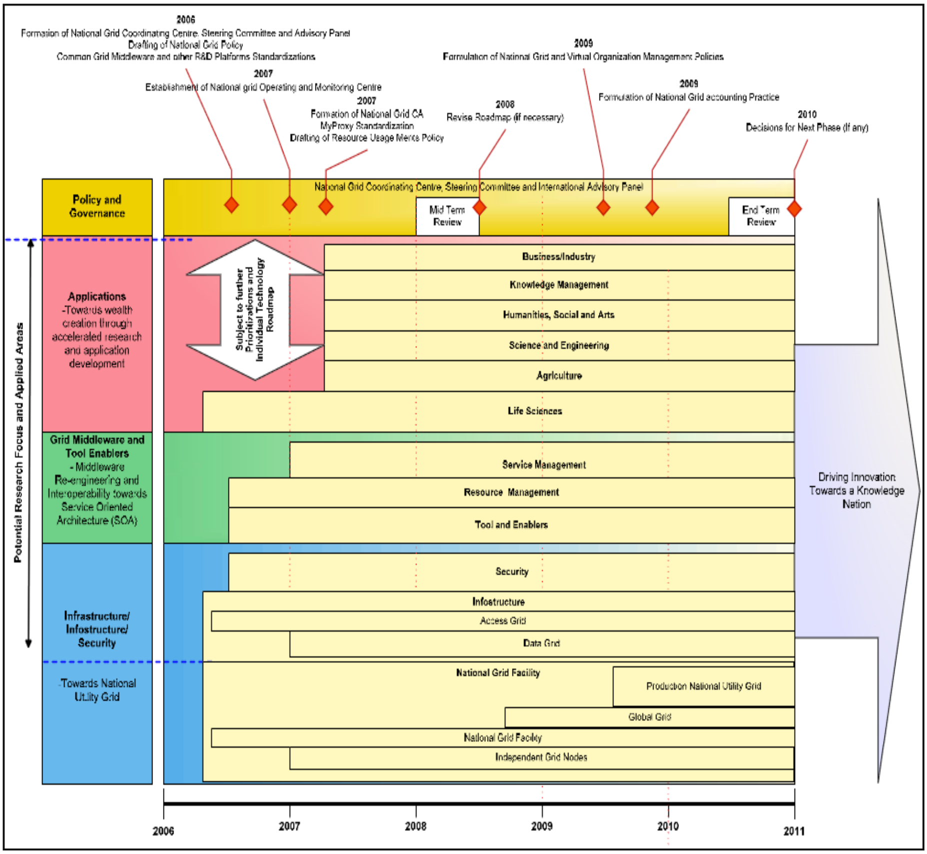

2006
Formation of National Grid Coordinating Centre, Steering Committee and Advisory Panel
Drafting of National Grid Policy
Common Grid Middleware and other R&D Platforms Standardizations
2007
Establishment of National grid Operating and Monitoring Centre
2007
Formation of National Grid CA
MyProxy Standardization
Drafting of Resource Usage Merits Policy
2008
Revise Roadmap (if necessary)
2009
Formulation of National Grid and Virtual Organization Management Policies
2009
Formulation of National Grid accounting Practice
2010
Decisions for Next Phase (if any)
Policy and Governance
National Grid Coordinating Centre, Steering Committee and International Advisory Panel
Mid Term Review
End Term Review
Potential Research Focus and Applied Areas
Applications -Towards wealth creation through accelerated research and application development
Subject to further Prioritizations and Individual Technology Roadmap
Business/Industry
Knowledge Management
Humanities, Social and Arts
Science and Engineering
Agriculture
Life Sciences
Grid Middleware and Tool Enablers - Middleware Re-engineering and Interoperability towards Service Oriented Architecture (SOA)
Service Management
Resource Management
Tool and Enablers
Infrastructure/ Infostructure/ Security
Security
Infostructure
Access Grid
Data Grid
-Towards National Utility Grid
National Grid Facility
Production National Utility Grid
Global Grid
National Grid Facility
Independent Grid Nodes
Driving Innovation Towards a Knowledge Nation
2006
2007
2008
2009
2010
2011

Fig. 5. National Grid Computing Roadmap [15]

3 Overview on Medical Data Grid

In this section, we briefly discussed on the medical data grid architecture and how the analysis of medical data implemented on the grid computing.

3.1 Medical Image Storage (Grid-Based PACS)

Medical image is one of the medical data that are produced in the healthcare centers every day. The large amount of medical images will be managed by the Picture Archiving and Communication System (PACS) in hospitals [9]. However, large sizes of medical images are difficultly managed by a standalone PACS System. Therefore, nowadays, most of the hospitals' PACS Systems have been integrated with the grid technology for reducing its management burdens. Refer Figure 6 for Grid-based PACS architecture.

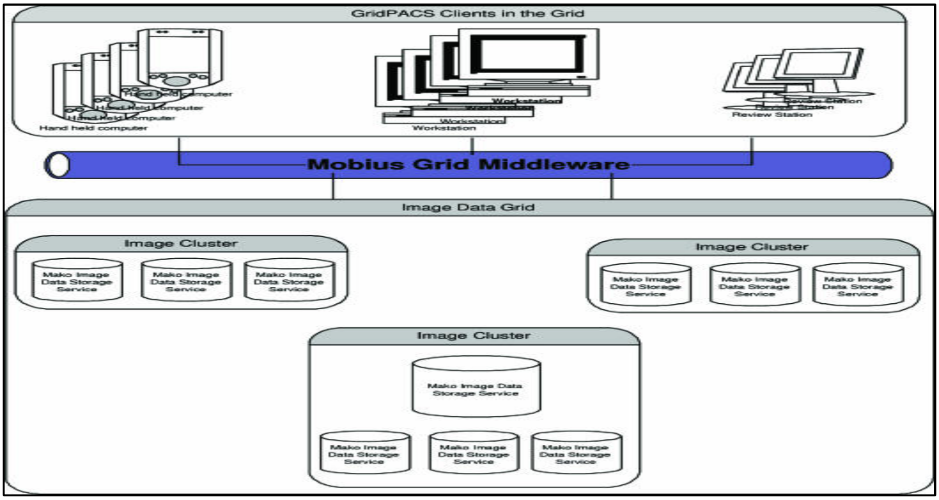

GridPACS Clients in the Grid
Hand held computer
Hand Held Computer
Workstation
Workstation
Review Station
Review Station
Mobius Grid Middleware
Image Data Grid
Image Cluster
Mako Image Data Storage Service
Mako Image Data Storage Service
Mako Image Data Storage Service
Image Cluster
Mako Image Data Storage Service
Mako Image Data Storage Service
Mako Image Data Storage Service
Image Cluster
Mako Image Data Storage Service
Mako Image Data Storage Service
Mako Image Data Storage Service
Mako Image Data Storage Service

Fig. 6. Grid-based PACS Architecture [10]

Typically, the images that stored inside the PACS are under the DICOM standard format. In DICOM standard, there are specified communication procedures among the application entities; but for the storage management, there is no standardization for them. To support distributed storage, retrieval, and querying of image data and descriptive metadata, a Grid-based PACS is developed [10]. Grid-based PACS allows the clusters of storage and computes machines to store and serving image data. It manages the server resources storage and also packages the efficient storage strategies for the image dataset. Besides, it use for maintaining the metadata describing image analysis workflows and support efficient execution of image analysis.

3.2 Medical Data Management

Distributed Medical Data Manager (DM^2) is developed inside the MedIGrid project. DM^2 represented the interface between the grid middleware and medical server. Refer Figure 7 for Distributed Medical Data Manager. Due to the reason of images are compatible to the DICOM, the images that produced inside the hospital will be transferred to the DICOM Server.

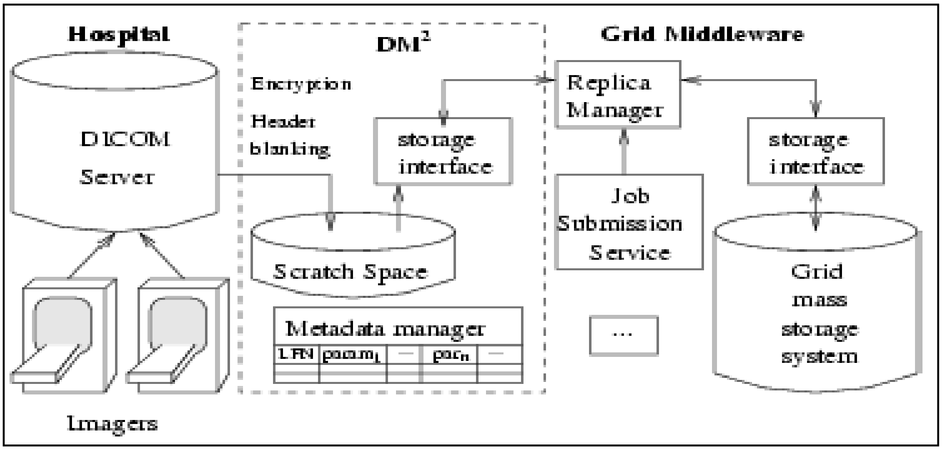


Fig. 7. Distributed Medical Data Manager [9]

DM^2 can communicate with single central or several distributed DICOM Servers that available in the hospital. In the grid side, DM^2 provides a grid storage interface. Each recorded image on the DICOM server will be registered in the DM^2 grid data management system.

For the grid side, authorized users are able to access to any file that have been registered. A grid node can access the image through the DM^2 interface. The grid incoming request is translated into a DICOM request and the desired file is returned. However, DM^2 is offers the additional services, such as automatic data anonymization and encryption to secure the critical data from being access by non-authorized users [9].

4 Medical Image and Video Processing

Typically, a registered patient's medical image will be diagnosed by his/her medical doctor. To confirm his/her diagnosis, the doctor is searching on the database to find out the similar known medical cases. By querying the image metadata, the target candidates are firstly selected out from the database. During comparison step, several similarity criterions are used. The similarity computation algorithms return in score and the images ranking are depending on the score. The highest score image is downloaded by the physician. This application requires one similarity job to be started for each candidate image to be compared to the database. To speed up the search, the computations are distributed over the available grid node [9].

4.1 Echocardiography System in PPUKM

In the Cardiac Care Unit of National University of Malaysia Medical Centre (PPUKM), they use the Acuson Squoia C512 Ultrasound Machine [16] to retrieve the echocardiography data. Then the data will be transferred to the computer for temporary storage. To avoid the missing of these echocardiography images data and also for reducing the usage of the computers storage, commonly, the technician will using the compact disk or other device storage to permanently store the echocardiography data. Refer to Figure 8 for the process of retrieval, transferring and permanent store the Echocardiography Data in Cardiac Care Unit PPUKM. Most of the echocardiography data are permanently stored in the video type's format.

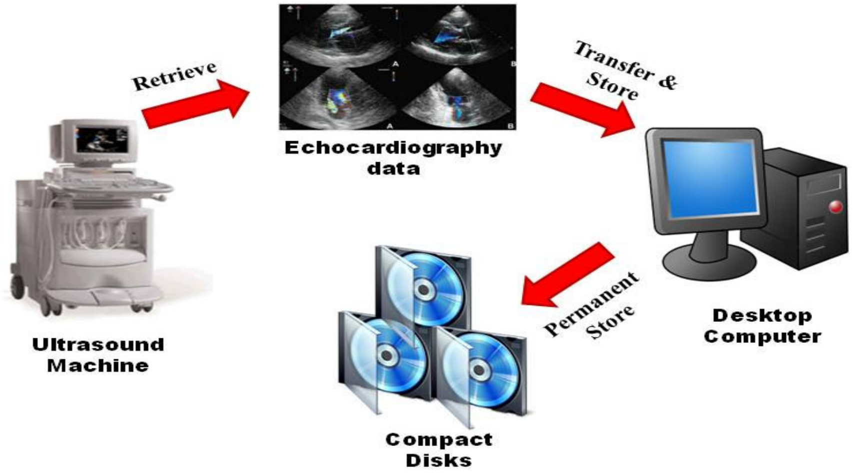


Fig. 8. Process of Retrieval, Transferring and Permanent Store the Echocardiography Data in PPUKM

4.2 Splitting Strategies

However, for long period, due to having more and more patients, there will be more and more compact disks used in permanently store the echocardiography data. Thus, there will be much more places need to use for storing these echocardiography data disk. Besides, management on those disks also a problem due to increasing of the disks' volume. To avoid the missing of those disks, an idea on using the grid computing storage for storing those data is introduced.

In the grid environment, the original video file is firstly uploaded in Grid node. Then, it will be split into many chunks for data dissemination or parallel tailoring. Typically, to respond the different quality of service request, different chunk splitting of same media file will be done. At last, in storage phase, each chunk will be sent to the Grid storage system and registered on the Replica Catalog [11].

However, the volume of those video type medical data is much bigger than the image data. Therefore, for long period, it is also increasing the usage of the grid storage. Thus, the splitting strategies and compression techniques need to be used for reducing the processing time and the data volume inside the grid storage. In this section, we briefly discussed on the existing splitting strategies and video compression techniques that common used in the video processing area.

In video processing, splitting process is one of the important parts. The term of "Splitting" is the mechanism to separate an object or idea into two or more parts. Number of chunks and the chosen splitting strategies are influence the system performances (e.g. delays during streaming, service continuity, or overhead due to parallel execution, in terms of system efficiency) [11]. There are three different splitting strategies will be discussed by us, such as the Uniform Splitting Strategies, First-reduced Splitting Strategies, and Fibonacci-based Splitting Strategies.

4.2.1 Uniform Splitting

For uniform splitting, each chunk will be split into the same size, which is,

$$V^i = \frac{V}{n}$$

Where:
V is assumed as the size of the whole video file,
V^i is the size of i^{th} chunk, with $i = 1,\ldots,n$,
n is the number of chunks

Although the uniform splitting strategy respects all the deadlines, the initial delay is higher than the others splitting strategy and giving a low value of system efficiency [12].

4.2.2 First-Reduced Splitting

In order to reduce the initial delay, the first-reduced splitting is used. For this strategy, the size of first chunk will be smaller than the others chunk. Following is the formula for the first-reduce splitting strategy,

$$V^1 = \frac{V}{1 + k(n-1)}$$

$$V^i = \frac{kV}{1 + k(n-1)}$$

Where:
V is assumed as the size of the whole video file,
V^i is the size of i^{th} chunk, with $i = 1,\ldots,n$,
V^1 is the size of 1^{st} chunk
n is the number of chunks
k is parameter

First-reduced splitting strategy can slightly improve the performance. But, for the second chunk, a structural deadline miss is occurs. This problem happens due to the great difference between the value for V^1 and V^2 [11].

4.2.3 Fibonacci-Based Splitting

To overcome the problem that occurs in the first-reduced splitting strategy, a Fibonacci-based splitting strategy is introduced. Fibonacci-based splitting is based on the Fibonacci number, where the size of i^{th} chunk is assumed to be proportional to the i^{th} Fibonacci number. Below is the formula for this strategy,

$$V^i = F(i) * \frac{V}{\sum_{i=1}^{n} F(i)}$$

Where:
V is assumed as the size of the whole video file,
V^i is the size of i^{th} chunk, with $i = 1,\ldots,n$,
n is the number of chunks
$F(i)$ indicates as i^{th} term of Fibonacci series.

The advantages of using this strategy are the size of the first chunk is reducing, while the size of the others is also slowly increasing. Thus, the structural deadline miss can be avoided and improve the system efficiency. Besides, the initial delay and the probability of deadline miss are reducing due to the smaller size of first chunks. But, there still a problem on this strategy, where the size of the first chunk has to be inferiorly limited. When n increases, the value of V^1 is rapidly decreases. Thus, this has reached the values that not efficiently used in multimedia streaming [11].

4.3 Video Compression Techniques

The term of "Video Compression" is means as the process to reduce the quality of the video data images. It also defined as a combination of spatial image compression and temporal motion compensation. Motion compensation is the process to apply the motion vectors into an image for synthesis the transformation to the next image. Usually, to reduce the volume of data, the compression process will be used. Several of the video compression techniques have been introduced in the previous works, such as the motion estimation techniques, and fast intra-mode selection techniques. However, the common used compression technique is the motion estimation technique. Thus, in this section, we briefly discussed on the motion estimation technique.

4.3.1 Motion Estimation

In video compression, combination between the motion estimation and motion compensation is an important key part for it. Motion estimation is the mechanism use to determine the motion vectors that describe the transformation from an image to another. Typically, it happens on the adjacent frames in a video sequence. There are 2 methods usually used to find the motion vectors, such as pixel based methods ("direct methods") and feature based methods ("indirect methods").

For direct methods, there are few algorithm are used, such as block-matching algorithm, phase correlation and frequency domain methods, MAP/MRF type "Bayesian" estimators, Pixel recursive algorithm and optical flow algorithm; while for indirect methods, they usually used the Harris corners, and match corresponding features between frames algorithm. In direct methods, several common evaluation metrics are used, such as mean squared error (MSE), Sum of Absolute Differences (SAD) and Mean Absolute Difference (MAD).

4.3.2 Mean Absolute Difference (MAD)

Typically, MAD will investigate the motion activity of each block. Then, each frame's block will be compared with the next frame's block to compute a motion vector for each block. By using the parallel image processing model with the MAD evaluation metrics, time execution speed of video compression is reduced by multiplicity of number of computers. Although reduced the time execution speed, there still a problem on the computational load [12][13].

4.3.3 Sum of Absolute Difference (SAD)

SAD is widely used in finding the correlation between the image blocks. It extracts the area by comparing on the absolute difference between each original block's pixel and the corresponding block's pixel. It is the simplest metric, which it just takes into count every pixel in a block. Due to its simplicity, this metric can reduce the computational complexity [14]. Since SAD analyzes each pixel separately, it also easy to parallelize. However, by using the SAD metric, there still a slightly increment on the total bit rate [14].

5 Conclusion

In this paper, we have discussed the current related medical data grid research projects and the techniques used for splitting and compressing the video images data.

From the discussions on those splitting techniques, we can see that most of the splitting techniques are facing the initial delay problem and the structural deadline miss problem. Based on the parameters of the initial delay and the deadline miss, we found that the Fibonacci-based splitting strategy is the most appropriate strategy for used in the video images processing.

Acknowledgements

We would like to take this opportunity to thank the National University of Malaysia Medical Centre (PPUKM) for giving us a chance to cooperate with them in improving the performances of their medical data storage. We also would like to thank the InfoComm. Development Center (IDEC), which is providing the grid infrastructure for us to do the research.

References

1. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
2. Rogulin, D., Estrella, F., Hauer, T., McClatchey, R., Solomonides, T.: Grid Information Infrastructure for Medical Image Analysis. In: Distributed, Parallel, and Cluster Computing (2004)
3. Keator, D.B., Grethe, J.S., Marcus, D., Ozyurt, B., Gadde, S., Murphy, S., Pieper, S., Greve, D., Notestine, R., Bockholt, H.J., Papadopoulos, P.: A National Human Neuroimaging Collaboratory Enabled by the Biomedical Informatics Research Network (BIRN). IEEE Transactions on Information Technology in Biomedicine 12(2), 162–172 (2008)
4. Kumar, V.S., Rutt, B., Kurc, T., Catalyurek, U.V., Pan, T.C., Chow, S., Lamont, S., Martone, M., Saltz, J.H.: Large-scale Biomedical Image Analysis in Grid Environments. IEEE Transactions on Information Technology in Biomedicine 12(2), 154–161 (2008)
5. Amendolia, S.R., Estrella, F., Hauer, T., Manset, D., McClatchey, R., Odeh, M., Reading, T., Rogulin, D., Schottlander, D., Solomonides, T.: Grid Database for Shared Image Analysis in the MammoGrid Project. In: Proceedings 8th International Database Engineering and Applications Symposium, pp. 302–311 (2004)
6. Ni, Y.-J., Youn, C.-H., Song, H., Kim, B.-J., Han, Y.: A PACS-Grid for Advanced Medical Services based on PQRM. In: 3rd International Conference on Intelligent Sensors, Sensor Networks and Information, pp. 625–630 (2007)
7. Boccia, V., Guarracino, M.R.: A grid Enabled PSE for mMedical Imaging: Experiences on MedIGrid. In: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (2005)
8. Weisbecker, A., Falkner, J., Rienhoff, O.: MediGRID – Grid Computing For Medicine and Life Sciences. In: Lin, S.C., Yen, E. (eds.) Grid Computing, pp. 57–65. Springer Science+Business Media, LLC (2009)
9. Breton, V., Blanchet, C., Lagre, Y., Maigne, L., Montagnat, J.: Grid Technology for Biomedical Applications. In: Daydé, M., Dongarra, J., Hernández, V., Palma, J.M.L.M. (eds.) VECPAR 2004. LNCS, vol. 3402, pp. 204–218. Springer, Heidelberg (2005)
10. Hastings, S., Oster, S., Langella, S., Kurc, T.M., Pan, T., Catalyurek, U.V., Saltz, J.H.: A Grid-Based Image Archival and Analysis System. J. Am. Med. Inform. Assoc. 12(3), 286–295 (2005)
11. Bruneo, D., Iellamo, G., Minutoli, G., Puliafito, A.: GridVideo: A Practical Example of Nonscientific Application on the Grid. IEEE Transactions on Knowledge and Data Engineering 21(5), 666–680 (2009)
12. Jeyakumar, S., Sundaravadivelu, S.: An Enhanced Two-Pass Motion Estimation Algorithm with Reduced Search for Fast Video Compression. In: First International Conference on Advanced Computing, pp, pp. 292–301 (2009)
13. Jeyakumar, S., Sundaravadivelu, S.: Implementation of Parallel Motion Estimation Model for Video Sequence Compression. In: International Conference on Computing, Communication and Networking, pp, pp. 1–5 (2008)

14. Cai, C., Zeng, H., Mitra, S.K.: Fast motion estimation for H.264. Signal Processing: Image Communication 24(8), 630–636 (2009)
15. Technology Roadmap for Grid Computing, `http://www.mosti.gov.my/mosti/images/pdf/grid-computing.pdf`
16. Saputro, A.H., Mustafa, M.M., Hussain, A., Maskon, O., Noh, I.F.M.: Motion Estimation along The Myocardial Boundary using Boundary Extraction and Optical Flow. In: Proceedings of the World Congress on Engineering, London, U.K, vol. 1 (2010)

Comparison of Linear Discriminant Analysis and Support Vector Machine in Classification of Subdural and Extradural Hemorrhages

Hau-Lee Tong[1], Mohammad Faizal Ahmad Fauzi[2], Su-Cheng Haw [1], and Hu Ng[1]

[1] Faculty of Information Technology,
Multimedia University, Jalan Multimedia,
63100 Cyberjaya, Selangor, Malaysia
[2] Faculty of Engineering,
Multimedia University,
Jalan Multimedia,
63100 Cyberjaya, Selangor, Malaysia
{hltong, faizall, schwa, nghu}@mmu.edu.my

Abstract. This paper describes new features for the classification of different types of extra-axial intracranial hemorrhages namely subdural hemorrhage(SDH) and extradural hemorrhage(EDH) on brain computed tomography(CT) scans. The main objective is to create an automatic retrieval system to reduce the time spent searching manually for the hemorrhagic images. Besides, the challenge is to locate suitable features to differentiate the SDH and EDH. One of the methods to distinguish EDH and SDH is through their shapes. Thus, a shape-based feature extraction is proposed in order to differentiate the SDH and EDH. For the classification part, we present a comparative study of linear discriminant analysis(LDA) and support vector machine(SVM) with linear kernel for the classification of SDH, EDH and normal regions. Both pattern classification techniques map pattern vectors to a high dimensional feature space to construct the optimal margin separating hyperplane. To conclude, SVM outperforms LDA from the obtained classification results.

Keywords: Computed tomography, hemorrhage classification, linear discriminant analysis, support vector machine.

1 Introduction

Extra-axial hemorrhage can be defined as bleeding within the intracranial space that happens within the skull but outside of the brain tissue. CT has been the prime modality for the detection of extra-axial hemorrhage because it is quick to perform, widely available and readily revealing extra-axial hemorrhage. The category of extra-axial hemorrhage mainly includes SDH and EDH. The major differences between

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 723–734, 2011.

SDH and EDH are on their shape and location. EDH occurs closer to the surface of the skin and usually appears to be bi-convex while SDH occurs closer to the brain and usually appear to be crescent in shape. Generally, classification of hemorrhage can be divided into supervised and unsupervised classification. From the existing works, various supervised and unsupervised classification techniques have been adopted. The challenge is rely on discovery of the appropriate and good descriptive features and how well the classification technique work together with the extracted features to obtain the promising results.

For the unsupervised classification, simple rule-based approach [1] has been proposed to identify the hemorrhage by using the priori-knowledge of region intensity, adjacent neigbhour and region size. Another approach common known as midline approach has been adopted by Mayank et al [2]. In their work, intracranial is divided into left and right hemispheres. Then, the histogram analysis is performed, followed by the dissimilarity comparison between left and right hemispheres. If dissimilarity above certain threshold, hemorrhage is considered exist.

Similar midline approach was also adopted by Liu Y. et al [3] and Hara et al [4] respectively, but with different extracted features for the similarity or dissimilarity test. However, the drawback of the midline approach is in the case when hemorrhage happens to be on left and right hemisphere with similar symmetrical position, midline approach will be inefficient to detect the hemorrhage.

On the other hand, for supervised classification, Liu R. et al in [5] proposed a SVM with a linear kernel for hemorrhagic slices detection by extracting the global texture features such as entropy, energy and so on. Besides, Gong et al [6] proposed trained decision tree by using local region features such as area, eccentricity, extent and so on to differentiate normal regions with different types of hemorrhagic regions. However, trained decision tree produces relatively low accuracy obtained for hemorrhages as compared to the accuracy of normal regions.

The main objective of this work is to develop a computer-aided detection system that could automatically detect different types of extra-axial intracranial hemorrhage. The classified CT scan can be used for retrieving process. With this retrieval system, medical doctors and students can retrieve the images for further study and analysis. This saves substantial time for going through the whole database to query for specific type of images.

In this paper, we evaluate and compare the different supervised pattern classifications namely LDA and SVM with linear kernel's results to test the efficiency and feasibility of the new shape features. Both LDA and SVM generate the hyperplanes that are optimal with respect to their respective objectives. Besides, both LDA and SVM with linear kernel are linear classifier. As such, these techniques are selected in order to have more proper comparison.

2 Overview of the Proposed Scheme

The proposed scheme is illustrated in Fig. 1. First of all, the original image will undergo three stages of preprocessing. The main objective of preprocessing is to equip the original image for the subsequent processes. The first preprocessing is to automatically adjust the contrast of the original image to a desirable contrast. The second preprocessing is to segment intracranial contents. The segmentation is achieved through global thresholding and morphological operations. The final preprocessing is to refine the contrast of hemorrhage to a desirable range to ease the detection of the hemorrhage.

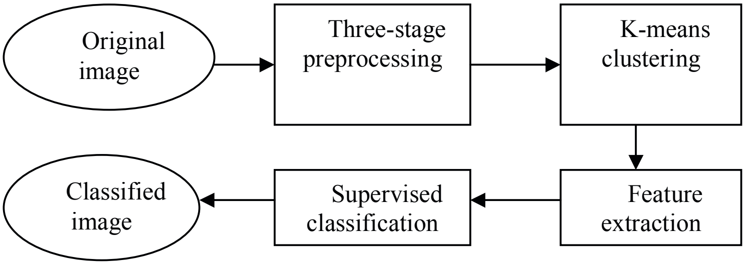


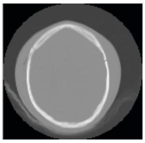
Fig. 1. Schematic diagram of the proposed scheme

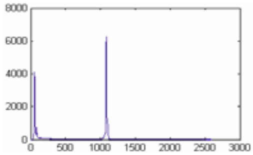
After preprocessing, the preprocessed image will go through k-means clustering to cluster all potential hemorrhage into a single cluster based on their intensity. Then, the shape features are extracted from each potential hemorrhage. Lastly, features are used in supervised classification system to annotate normal region, EDH and SDH.

3 Three-Stage Preprocessing

3.1 First Level Preprocessing: Original Image Contrast Enhancement

The first level preprocessing is to adjust the contrast of the original image to a desirable range. This is to improve the visibility of the region of interests(ROIs) as original image is with very low visibility for ROIs as shown in Fig. 2. Therefore, an automatic-contrast adjustment system is proposed to achieve this.





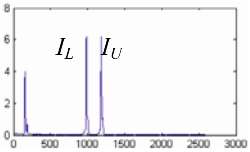


Fig. 2. Original image **Fig. 3.** Constructed histogram **Fig. 4.** Absolute first difference

During automatic-contrast adjustment, the following steps will be executed:

i) Construct the histogram for the original image. As shown in Fig. 3, the constructed histogram consists two major peaks but the only the rightmost peak is contributed by the ROIs.

ii) Smoothen the curve by using convolution operation to ease the acquisition process of the upper and lower bounds from the rightmost peak.

iii) Transform the smoothened curve into absolute first difference as shown in Fig. 4 and find the upper and lower bounds from the left and right peaks for the absolute first difference.

iv) Channel the appropriate I_L and I_U into equation (1) for linear contrast stretching.

$$F(i,j) = I_{\max} \frac{(I(i,j) - I_L)}{(I_U - I_L)} \tag{1}$$

Where $I_{max}I(i,j)$ and $F(i,j)$ denote the maximum intensity in the image, original pixel intensity and contrast enhanced pixel intensity respectively. After the contrast adjustment, the ROIs visibility improved image is shown in Fig. 5.

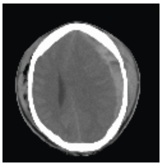
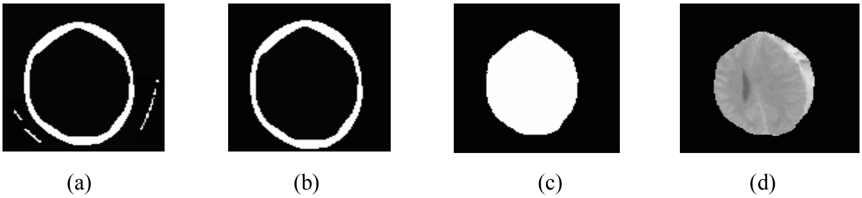


Fig. 5. Contrast adjusted image

3.2 Second Level Preprocessing: Segmentation of Intracranial Contents

Intracranial contents includes brain tissue and cerebral spinal fluid(CSF). Segmentation of intracranial contents is to strip off the skull, scalp and background from intracranial contents. This is achieved by global thresholding that normally remains the skull and background as shown in Fig. 6(a). The acquisition of the skull is to generate the intracranial contents mask.

Identification of skull is done through connected components analysis and locating the largest connected components as skull always appears to be largest region after thresholding. After acquisition of the skull, a flood fill operation is applied to fill up the holes within the skull. Then intracranial mask is generated by setting the intensity of the skull to zero as shown in Fig. 6(c). At last, intracranial area is acquired as shown in Fig. 6(d).



(a) (b) (c) (d)

Fig. 6.(a) Thresholded image (b) Skull (c) Intracranial mask (d) Intracranial contents

3.3 Third Level Preprocessing: Contrast Adjustment of Candidate SDH and EDH

The aim in this section is to adjust the contrast of the candidate SDH and EDH to a better range. This adjustment is to ease the subsequent clustering process. The adjustment is based on priori-fact that the intensities of the SDH and EDH are usually located within the range after the peak position of the global image. Therefore, adjustment of the contrast is focus on after peak position.

Firstly, histogram of the intracranial contents is constructed. Then lower limit is automatically located from the peak position of the histogram. From obtained lower limit, upper limit is derived from

$$I_U = I_L + I_\alpha$$

Where, I_α is predefined at 500 obtained from experimental observation.

Once I_L and I_U have been identified, these values are used for the linear contrast stretching. To reduce noise, median filter is applied. Median filter is applied as it is less sensitive to extreme values and is able to remove noise without reducing the sharpness. The image obtained after contrast adjustment and noise reduction is shown in Fig. 7.

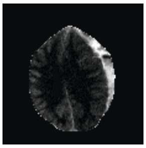


Fig.7. Hemorrhage contrast adjusted image

4 Potential Candidate SDH and EDH Clustering

The aim for this section is to cluster potential hemorrhagic regions into a single cluster. In order to achieve this, firstly image is partitioned into two clusters. From these two clusters, the low intensity cluster without potential hemorrhagic regions is ignored. Only the high intensity cluster which consists of potential hemorrhagic regions is considered. In other words, the high intensity cluster can consists of high intensity normal regions and hemorrhagic regions. Therefore, classification is adopted in later section to differentiate the normal regions, SDH and EDH.

Four clustering techniques which are Otsu thresholding, fuzzy c-means(FCM), k-means and expectation-maximization(EM) are attempted in order to select the most appropriate technique for the potential candidate for EDH and SDH clustering. The comparison results are shown in Fig. 8.

From the results obtained, Otsu thresholding, FCM and EM encountered the over-segmentation as hemorrhagic region merged together with surrounding pixels. This directly causes the hemorrhagic region to be distorted from their original shape. On the other hand, k-means conserves the original shape most of the time and produces less noise. Thus, k-means clustering is adopted to obtain more proper shape of SDH and EDH.

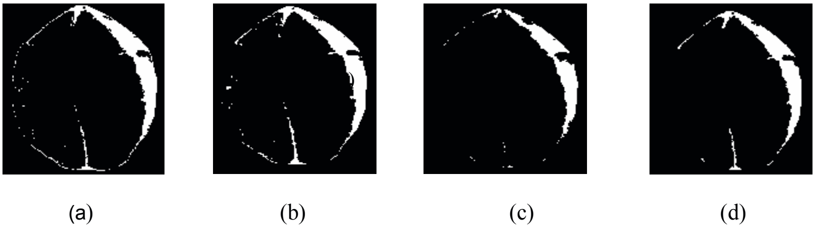


Fig. 8. Clustering results by (a) Otsu thresholding (b) FCM clustering (c) K-means clustering and (d)EM clustering

5 Extraction of Shape Features

Radiologists usually based on intensity, size, shape and position to judge the existence of hemorrhagic regions. Therefore, quantified features are extracted to describe the shape for the automatic classification. The features are selected based on two criteria which are to differentiate (1) normal regions from hemorrhagic regions and (2) SDH from EDH. The seven shape features considered are region area, border contact area, linearity, concavity, ellipticity, circularity and triangularity.

Region area and border contact area are used to differentiate normal regions from hemorrhagic regions. Region area is to differentiate high intensity noise which relatively small compared with the hemorrhagic regions. Border contact area is a measure of number of pixels contact with the skull. Border contact area is adopted based on the priori-knowledge that normal high intensity region such as falx, tentorium and noise having less contact with the skull.

The remaining five features are mainly used to differentiate SDH from EDH. EDH and SDH always appear to be bi-convex and elongated crescent in shape respectively. Generally, EDH is more elliptic, circular and triangular as compared to SDH. However, SDH is more concave and linear than EDH. Based on these priori-facts, ellipticity, circularity, triangularity, linearity and concavity are proposed to quantify the shape of EDH and SDH. Firstly, circularity[7] is computed based on moments as given in equation (2).

$$\text{Circularity, } \partial(ROI) = \frac{(\mu_{0,0}(ROI))^2}{2\pi(\mu_{2,0}(ROI) + \mu_{0,2}(ROI))} \tag{2}$$

where the (i,j)-moment, $\mu_{i,j}(ROI)$ as defined in equation (3):

$$\mu_{i,j}(ROI) = \iint_{ROI} x^i y^j dxdy \tag{3}$$

For triangularity and ellipticity, the affine moment invariant used to characterise the triangle and ellipse is given in equation (4).

$$I = \frac{\mu_{2,0}(ROI)\mu_{0,2}(ROI) - (\mu_{1,1}(ROI))^2}{(\mu_{0,0}(ROI))^4} \tag{4}$$

From the affine moment invariant, triangularity[8] and ellipticity[8] of a ROI are derived as shown in equation (5) and equation (6) respectively.

$$\text{Triangularity}, \angle(ROI) = \begin{cases} 108I & if\ I \leq \dfrac{1}{108} \\ \dfrac{1}{108I} & otherwise \end{cases}, \tag{5}$$

$$\text{Ellipticity}, \ell(ROI) = \begin{cases} 16\pi^2 I & if\ I \leq \dfrac{1}{16\pi^2} \\ \dfrac{1}{16\pi^2 I} & otherwise \end{cases}, \tag{6}$$

Linearity[9] that is used to represent the elongated shape of SDH is defined as in equation (7).

$$\text{Linearity}, L(ROI) = 1 - \frac{\text{minor axis}}{\text{major axis}} \tag{7}$$

where, the major and minor axes are the longest and shortest diameter of the ROI respectively.

We proposed a new concavity measure to measure the degree of concaveness for SDH. This new concavity takes into consideration the contours and overlapping area in order to acquire the concave area. The derivation of the concavity is based on the following steps:

i) Locate inner contour(without contact with skull) and outer contour(with contact with skull) as shown in Fig 9(b) and 9(c) respectively.

ii) Locate the two endpoints of the inner contour and outer contour by using the 3 X 3 neighbourhood (x_1^{outer} , y_1^{outer}), (x_2^{outer} , y_2^{outer}) (x_1^{inner} , y_1^{inner}), (x_2^{inner} , y_2^{inner}).

iii) Interpolate the linear line to connect the endpoints, (x_1^{inner} , y_1^{inner}) with (x_2^{inner}, y_2^{inner}) and (x_1^{outer} , y_1^{outer}) with (x_2^{outer} , y_2^{outer}) in order to generate the closed inner and outer contour based on the equation (8) as shown in Fig. 9(d) and 9(e).

$$y = y_1 + \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)} \tag{8}$$

iv) Fill up the closed inner and outer contours by using morphological operation as shown in Fig. 9(f) and 9(g).

v) Acquire the concave area by overlapping the filled inner contour with filled outer contour as shown in Fig. 9 (h): Concave area $=A_{filled\_contour} \cap A_{filled\_inner}$.

vi) Divide the concave area by area of filled inner contour to normalize the concave area to the interval [0, 1]. With this, concavity of a ROI is defined as:

$$\text{Concavity}, \lambda(ROI) = \frac{A_{filled\_inner} \cap A_{filled\_outer}}{A_{filled\_inner}}$$

The overlapping area should be equal to the filled inner contour area in the case if filled inner contour is purely concave shape which gives the maximum of 1.

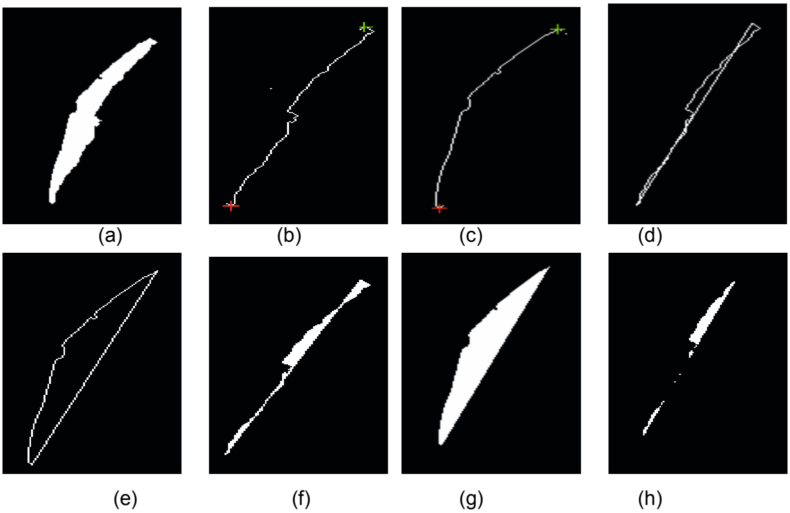


Fig. 9. (a) SDH region. (b) Inner contour with located endpoints (c) Outer contour with located endpoints (d) Inner closed contour (e) Outer closed contour (f) Filled up inner contour (g) Filled up outer contour (h) Overlapping area

6 LDA vs. SVM Classification

Generally, in pattern recognition applications, the classifier contributes to significant final results of the previous processes. Therefore, it is important to select the appropriate techniques for the betterment of the final decision. In order to have the proper comparison for our data, linear SVM and LDA are adopted in this research. Both SVM[10] and LDA[11] are supervised pattern classification techniques. Therefore, the data is divided into training and testing sets. Both techniques tend to map a testing feature vector, Y={y_1 , y_2,…., y_7} in $\Re$ of a region to a most closest class set {Ci}, where i=class number. In simplest form, both techniques seek the optimal classification hyperplane with respect to their respective objective.

However, there are some differences between LDA and linear SVM. Suppose we have a set of sample patterns: $\{x_1 , x_2 ,…., x_N\}$, each of which is annotated to their respective class. LDA computes a transformation that maximizes the between-class scatter, S_{bc} and minimizes the within class scatter, S_{wc} :

$$\text{Maximize: } \frac{\det(S_{bc})}{\det(S_{wc})}$$

The within-class scatter and between-class scatter are defined in equation (9) and (10) respectively.

$$S_{wc} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (x_j - \mu_i)(x_j - \mu_i)^T \tag{9}$$

$$S_{bc} = \sum_{i=1}^{C} (\mu_i - \mu)\ (\mu_i - \mu)^T \tag{10}$$

Where, μ denotes mean of the entire data;

μ_i denotes mean vector of class ***i***=1,2…,C;

N_i denotes number of samples in class ***i***.

For LDA, all classes are assumed to own identical covariance matrices. The discriminant functions for LDA are defined by equation (11).

$$d_i(x) = \ln \pi_i - \frac{1}{2} \mu_i^T \sum\nolimits_{pooled}^{-1} \mu_i + x^T \sum\nolimits_{pooled}^{-1} \mu_i \tag{11}$$

Where, ***x*** is the new data point to be classified, π_i is the estimated prior probability of class ***I*** and $\sum_{pooled}^{-1}$ is the estimated inverse pooled covariance matrix. From equation (11), the new data point, ***x*** is classified by the classification rule as specified by equation (12).

$$cf(x) = \arg \max_i d_i(x) \tag{12}$$

The resulting LDA decision boundaries between classes are linear.

For SVM, a hyperplane can be defined by $(w \bullet x_i) + b$=0, where w is the vector perpendicular to the hyperlane and $\boldsymbol{b}$ is the bias. SVM seeks optimal hyperplane by solving the following optimization problem:

$$\text{Minimise: } \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i \tag{13}$$

Subject to $y_i (w \bullet x_i) + b \geq 1 - \xi_i, \xi_i \geq 0$. Where ξ_i denotes the slack variable. Parameter $\boldsymbol{C}$ is penalty factor which maintain the balance of classification accuracy. SVM linear classifier classifies the given test instance$x$ based on the decision function as in equation (14).

$$f(x) = \text{sign}(\sum_{x_i \in sv} \alpha_i y_i (x_i \bullet x) + b) \tag{14}$$

Wheresv is the set of support vectors andα_i is a Lagrange multiplier. The value of$f(x)$ denotes the distance of the test instance from the separating hyperlane and the sign indicates the class label.

7 Experimental Protocol and Discussion

The data used in this work consists of 121 CT scans were retrospectively obtained from two collaborating partners. These CT scans are generated from different CT scanners i.e., Siemens Somaton Plus 4 CT scanner and Toshiba Aquilion scanner. The CT is routinely performed by using 512X512 matrix and the slice thickness ranges from 6 to 10mm. In total, there are 130 hemorrhagic regions and 457 normal regions. These normal regions have similar intensity with the hemorrhagic regions. The experiments were performed on a computer with Intel Pentium-4 2.9GHz and 1024MB main memory where the average processing time is less than 15 seconds for the entire classification. The visual classification results are shown in Fig. 10. As depicted in Fig. 10, the hemorrhagic regions are well segmented by using the proposed approach.

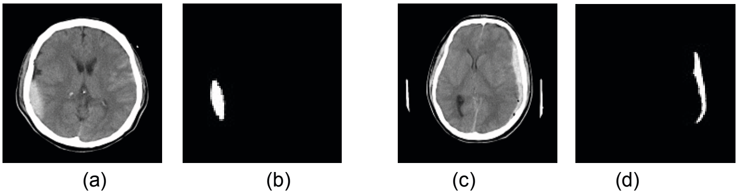


Fig. 10. (a) EDH image (b) Detected EDH (c) SDH image (d) Detected SDH

For the testing part, ten-fold cross validation was considered on our dataset by partitioning 587 regions into 10 disjointed subsets. Each disjointed subset acts as a new incoming data that needed classification. The cross-validation process is iterated for 10 times so that every feature will be validated by LDA and SVM classifiers. Both LDA and linear SVM are experimented rather than non-linear methods such as SVM using a radial-basis function kernel and quadratic LDA. This is because from our extended experiments, non-linear methods do not yield satisfactory results. This indirectly implies that the extracted features are more toward linear type. The results are obtained from both methods as depicted in Table 1 and Fig. 11.

Table 1. The accuracy of the classification

	LDA	SVM
SDH	0.8500	0.9000
EDH	0.9000	0.9000
Normal region	0.8709	0.9409
Overall	0.8705	0.9333

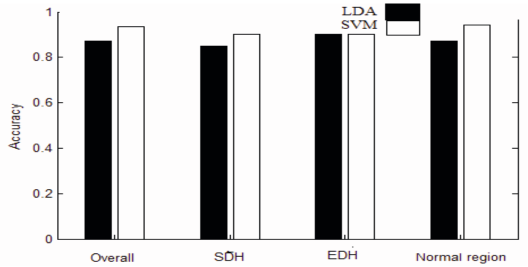


Fig.11. Comparison results for overall, SDH, EDH and normal regions

From Table 1, the experimental results show that the acceptable rates were attained in which the overall accuracies have achieved more than 87%. This is contributed by the features employed which can effectively and feasibly to differentiate each class to a certain extent. In other words, the features employed can describe well the characteristics of normal regions, SDH and EDH. From Fig. 11, linear classifier based on SVM outperforms linear classifier based on LDA in accuracy rate for SDH and normal region. Besides, the experimental results also show that SVM outperforms LDA for overall accuracy by around 6.28%. This outperformance is statistically significant for the improvement of classification accuracy. As such, we can conclude that separating hyperplanes constructed by SVM are more efficient and suitable for the classification of the three classes.

8 Conclusion

As a conclusion, the proposed system yielded promising results for the classification of normal regions, SDH and EDH. Besides, the proposed features are feasible and efficient to differentiate the normal regions, SDH and EDH. SVM with linear kernel outdoes LDA in which on the overall SVM and LDA achieved 93.33% and 87.05%

accuracies respectively. Future works to be considered will be directed towards classification of more abnormalities in the brain such as infarct, atrophy and so on.

Acknowledgments. The authors would like to express their gratitude to Dr. Fatimah Othman from Putrajaya Hospital, Dr. Ezamin Abdul Rahim and Dr. Noraini Abdul Rahim from Serdang Hospital for the image acquisitions, reviews and consultations.

References

1. Dubravko, C., Sven, L.: Rule-Based Labeling of CT Head Image. In: 6th Conference on Artificial Intelligence in Medicine, pp. 453–456 (1997)
2. Mayank, C., Saurabh, S., Jayanthi, S., Kishore, L.T.: A Method for Automatic Detection and Classification of Stroke from Brain CT Images. Engineering in Medicine and Biology Society (2009)
3. Liu, Y., Lazar, N.A., Rothfus, W.E., Dellaert, F., Moore, A., Schneider, J., Kanade, T.: Semantic-based Biomedical Image Indexing and Retrieval. In: Shapiro, Kriege, Veltkamp (eds.) Trends and Advances in Content-Based Image and Video Retrieval (2004)
4. Chan, T.: Computer aided detection of small acute intracranial hemorrhage on computer tomography of brain. Computerized Medical Imaging and Graphics 31(4-5), 285–298 (2007)
5. Liu, R., Chew, L.T., Tze, Y.L., Cheng, K.L., Boon, C.P., Lim, C.C.T., Qi, T., Tang, S., Zhang, Z.: Hemorrhage slices detection in brain CT images. In: 19th International Conference on Pattern Recognition, pp. 1–4 (2008)
6. Gong, T., Liu, R., Tan, C.-L., Farzad, N., Lee, C.K., Pang, B.C., Tian, Q., Tang, S., Zhang, Z.: Classification of CT brain images of head trauma. In: Rajapakse, J.C., Schmidt, B., Volkert, L.G. (eds.) PRIB 2007. LNCS (LNBI), vol. 4774, pp. 401–408. Springer, Heidelberg (2007)
7. Zunic, J., Hirota, K., Rosin, P.L.: A Hu moment invariant as a shape circularity measure. The Journal of the Pattern Recognition Society 43(1), 47–57 (2010)
8. Rosin, P.L.: Measuring shape: ellipticity, rectangularity, and triangularity. Machine Vision and Applications 14(3), 172–184 (2003)
9. Stojmenovic, M., Nayak, A., Zunic, J.: Measuring Linearity of a Finite Set of Points. In: 2006 IEEE Conference Cybernetics and Intelligent Systems (CIS), pp. 1–6 (2006)
10. Muthu, R.K., Shuvo, B., Chinmay, C., Chandan, C., Ajoy, K.R.: Statistical analysis of mammographic features and its classification using support vector machine. Expert Systems with Applications 37(1), 470–478 (2010)
11. Dave, P.B., Simon, P.K., Philip, C., Richard, B.R., Ciarán, F.: A Parametric Feature Extraction and Classification Strategy for Brain–Computer Interfacing. IEEE Transactions on Neural Systems and Rehabilitation Engineering 13(1), 12–17 (2005)

A New Approach for Adjusting Braille Image Skewness in Optical Braille Recognition

Abdul Malik S. Al-Salman*, Ali El-Zaart, and Abdu Gomai

Department of Computer Science, College of Computer and Information Sciences
King Saud University
Riyadh, Kingdom of Saudi Arabia
{salman,elzaart}@ksu.edu.sa, abdugomai@gmail.com

Abstract. Braille recognition is the ability to recognize Braille cells in Braille images. The result will be used in several applications such as Braille translating, Braille copying...etc. However, the performance of these applications will be affected by some factors such as skew of Braille documents through the scanning or embossing. In this work, we designed a new approach for adjusting Braille image skewness. This approach is summarized in three steps: (1) A Braille image is segmented by estimated thresholds values which are resulting from Beta distribution, (2) simple Braille cells are detected by using morphological operations; and (3) the rotation angle of Braille image is estimated by calculating the median of slope angles of all straight lines that pass through any three dots of detected Braille cells. The proposed method is tested on several Braille images. Results demonstrated that the method is able to adjust the skewed Braille images accurately.

Keywords: Braille image, Braille cell, Braille recognition, skewness, slope angle, recto dot, verso dot.

1 Introduction

Braille is a reading and writing system which enables blind and partially sighted persons to read and write through touch. Braille System was invented by Louis Braille in 1824. It generally consists of cells of six raised dots which are arranged in three rows and two columns. These six positions of dots used in arrangement to give just 64 different Braille codes "characters".

There are standards for Braille production that determine the height and diameter of a dot, the spacing between dots and between cells. Braille Image skewing correction plays an important role in automated Braille recognition systems. For some years there has been an increasing trend to use computers for entering, editing and printing Braille documents using special purpose software and printers. Computerized systems can now produce Braille documents from ASCII text. But these Braille documents are possibly to be weak or exposed to damage. Therefore, they must be

* This research is funded by NPST @ KSU-KACST, Saudi Arabia.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 735–746, 2011.

reproduced so that they can be preserved and accessed by more people. Since manual transcription is verbose and costly, there is an important need for an automatic system to reproduce Braille documents. Many methods have been conducted on Braille printing and translation. One of the existing methods is to use an Optical Braille Recognition system (OBR) to scan the Braille document and translate or convert it to normal ASCII text then printing it using a regular printer or a Braille embosser.

There are many attempts to perform optical Braille page recognition using different methods. In 1988, Dubus et al. [1] introduced an algorithm called *Lectobraille* which translates relief Braille into Black-ink. After that, image processing techniques are used to translate Braille pages to printed text. In 1993, Mennens et al. [2] developed an optical recognition system that can recognize a scanned Braille document using a commercially available scanner. However, the system cannot manipulate deformation in the dot grid arrangement. In 1994, Ritchings et al. [3] used a flatbed scanner to scan both single and double sided Braille documents at 100 dpi and at 16 grey-levels. The authors used a character segmentation and recognition to locate the lines of Braille cells. The results reported for double-sided Braille documents were around 96.5% correct. This approach used a fixed grid to handle the variation in positions of characters but the fixed grid causes some problems in Braille recognition. In 1995, Blenkhorn [4] presented a method for converting Braille dots into print by using a finite state system to hold the current context with right context and the checking was achieved by using matching algorithms. This system has been designed to be configurable for a wide range of languages and character sets, and uses a predominantly table driven method to achieve this. But this work didn’t use any technique of image processing for converting Braille images. Also, in 1995, Hentzschel and Blenkhorn [5] introduced a system for optical Braille recognition based on twin shadows technique, which subtracts two images of the same Braille page, where each image was taken under different illumination conditions. This system can locate and extract the Braille dots in Braille images as pairs of white and black spots, where each pair of white and black spots represents a single Braille dot, but if some pairs are lost, false ones are formed. In other hand, this system did not detect the distances between Braille dots automatically. In 1997, Oyama et al. [6] proposed a dot detection module incorporated in the OBR system to detect both recto and verso Braille dots on both single and double sided pages. The problem in this work was due to the difference in light reflectance between recto and verso dots. In 1999, Ng et al. [7] presented an automatic Braille recognition system to translate Braille documents into English or Chinese text using edge enhancement, noise filtering and boundary detection techniques. The recognition rates were good; but, there is no explanation of grid deformed input, nor its effectiveness. In 2004, [8] developed windows based commercial OBR software that allows reading single and double sided Braille documents with a standard scanner. Unfortunately, it supports few languages and does not preserve the layout. In 2001, Murray and Dais [9, 10] developed a portable device for optically scanning embossed Braille and conversion of the scanned text to binary Braille representation. Because of the user was in manage of the orientation of scanning and only a small part was scanned at each time, grid deformation wasn’t a main concern, and an easier algorithm was used to give efficient and immediate translation of Braille codes. In 2004, Wong et al. [11] proposed an OBR system that is capable of recognizing a single sided Braille page to produce a text file. The recognition module designed to work with threshold images resulting from the half-character detection module. In 2004, Antonacopoulos and Bridson [12] proposed a Braille recognition system to identify Braille dots on both single and double sided documents of average quality where many improvements were added to increase the cost-effectiveness and usability of the system. This approach is similar to the approach proposed by Ritchings in [3] but solves the fixed grid problem by using a flexible grid. In 2005, Falcón et al. [13] presented the development of *BrailLector* system that translates Braille scanned images into normal text. In 2006, Namba and Zhang [14] proposed the Braille image recognition system by CNN (Cellular Neural Network) for associative memory. Their system consists of three stages: preprocessing, feature extraction, and recognition. The authors have obtained a good recognition rate (87.9%). In 2007, Al-Salman et al. [15] developed a system to recognize a single and double sided images of embossed Arabic Braille and convert them to regular text. This work helped to build a fully functional Optical Arabic Braille recognition system. The recognition rate was around 99% correct. In 2008, Zhenfei et al. [16] proposed a Braille documents parameters estimation method based on Radon transform to automatically determine the skewness, indentations, and spacing in both vertical and horizontal directions. The authors of this paper denied the existence of recent attempts for Braille documents recognition such as [15], which adjusts skewed image with 4 degrees from either the left or right sided. As well as the literature review of [16] didn't include any recent works for Braille recognition systems [13-17]. The dataset used to perform a test in their system is too small (only four single sided images).

In this work, we present an efficient approach for correcting the skewing in Braille images to build an effective optical Braille recognition system. This approach is based on detecting a set of simple cells containing at least three dots on the same straight line, calculating the slope angles for each detected Braille cell and finding the median of those slope angles to get the rotated angle of Braille page.

The paper is organized as follows. Section 2 presents the proposed approach step by step. Experimental results are shown in Section 3. Conclusions are given in Section 4.

2 The Proposed Approach

Our developed method adjusts the skewed Braille images based on detecting simple cells containing at least three dots at the same line and estimating the slope angles between the first and last dot because the rotation angle of Braille image is the rotation angle of Braille cells. From the literature review and our observations, we found that the Braille dots aren't regular for the margins of Braille page, so, we can't be relied upon them to know the skewing angle of Braille page. This was the reason for proposing this method. The method consists of three major steps: Braille image segmentation, Simple Braille cells detection and Skewness angle estimation (See Fig. 1). In this section, we will explain these steps in detail.

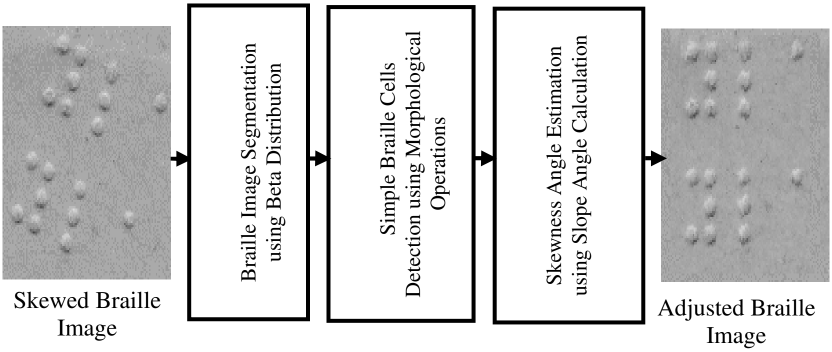

Skewed Braille Image
Braille Image Segmentation using Beta Distribution
Simple Braille Cells Detection using Morphological Operations
Skewness Angle Estimation using Slope Angle Calculation
Adjusted Braille Image

Fig. 1. The proposed method diagram

(a) (b)

Fig. 2. (a) Recto dot and (b) Verso dot

2.1 Braille Image Segmentation

The histogram of Braille image consists of three modes. The three modes represent the following three classes of pixels: (i) Mode 1: represents the dark region of a recto and verso dot. (ii) Mode 2: represents the background. (iii) Mode 3: represents the light region of a recto and verso dot (See Fig. 2a and Fig. 2b). The problem of segmentation is the estimation of thresholds T_1 and T_2 for separating the three classes. We assume that a histogram of a Braille image is a combination of three Beta distributions. The Beta distribution is a continuous probability distribution with the *probability density function* (pdf) defined on the interval [0, 1] [17]:

$$f(x, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\, x^{\alpha-1}\, (1 - x)^{\beta-1},$$

where α and β are the shape parameters of the distribution and must be greater than zero, x is a random variable, and it must be between 0 and 1. The Beta distribution can take different shapes depending on the values of its two parameters α and β. The histogram $h(x)$ of a Braille image can be written as follows:

$h(x) = p_1 f(x, \alpha_1, \beta_1) + p_2 f(x, \alpha_2, \beta_2) + p_3 f(x, \alpha_3, \beta_3)$ The estimated threshold T_i^{new} of a Braille image can be calculated using the following formula:

$$T_i^{new} = 1 - e^{\frac{-A - B\log(T_i^0)}{C}}$$

,where $B = \alpha_i - \alpha_{i+1}$, $C = \beta_i - \beta_{i+1}$, $K_r = \Gamma(\alpha_r - \beta_r)/\Gamma(\alpha_{r+1} - \beta_{r+1})$ and $r = i, i+1$

The statistical parameters of the histogram (p_i, α_i, β_i), i=1, 2, 3 are estimated using the stability of thresholding algorithm [17].

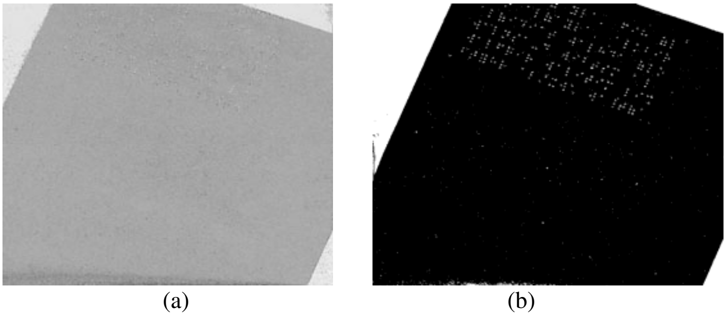


Fig. 3. (a) Original Braille image, and (b) Segmented Braille

1. Estimate initial threshold values T_1^0 *and* T_2^0 *,*

T_1^0 *will be the average gray level value of image starting from 0 to maximum value using the following formula:*

$$T_1^0 = \sum_{i=0}^{MaxIndex} (i * h(i)) \;/\; \sum_{i=0}^{MaxIndex} h(i)$$

T_2^0 *will be the average gray level value of image starting from maximum value to 255 using the following formula:*

$$T_2^0 = \sum_{i=MaxIndex}^{255} (i * h(i)) \;/\; \sum_{i=MaxIndex}^{255} h(i)$$

2. T_1^0 *and* T_2^0 *are then used to estimate the values for* α *,* β *and prior probability (P) for each mode of the histogram as follows:*

- *The values of* α *,* β *for mode i=1,2,3 are estimated using moment order.*
- *The prior probability* (P) *for a certain mode i is estimated according to the following formula:*

$$P_i = \sum_{j \in Mode_i} h(x_j) \;/ \sum_{j=0}^{255} h(x_j), \quad i = 1,2,3$$

3. The estimated values of α *,* β *and prior probability (P) for each mode i=1,2,3 are now used to recalculate the new thresholds* T_i^{new} *using the following formula:*

$$T_i^{new} = 1 - e^{\frac{-A - B\log(T_i^0)}{C}}$$

Fig. 4. Braille image segmentation algorithm

where $A = \log((p_i K_i)/(p_{i+1}K_{i+1}))$, $B = \alpha_i - \alpha_{i+1}$, $C = \beta_i - \beta_{i+1}$, *and* $K_r = \Gamma(\alpha_r - \beta_r)/\Gamma(\alpha_{r+1} - \beta_{r+1})$ and $r = i, i+1$.

4. The procedure is repeated until the error is zero. The error is calculated as the cumulative difference between the old threshold values and the new threshold values as the following equation:

$$error = \left|(T_1^0 - T_1^{new}) + (T_2^0 - T_2^{new})\right|$$

Otherwise, we set $T_i^0 = T_i^{new}$, $(i = 1,2,3)$ *and repeat steps 2 until 4. By the end of this algorithm the optimal values of* T_1^{new} *and* T_2^{new} *are found. After calculating threshold values, we then segment the Braille image.*

Fig. 4. (*continued*)

The output of this step is a segmented Braille image (See Fig. 3b) and the algorithm is summarized in Fig. 4.

2.2 Simple Braille Cells Detection

The segmented Braille image resulting from the step one is used to detect a set of simple Braille cells. This step is achieved by these operations:

1. Dilation operation is applied on segmented Braille image to increase the sizes of distorted dots, filling in holes and broken areas, and connecting areas that are separated by spaces smaller than the size of the structuring element (See Fig. 5).

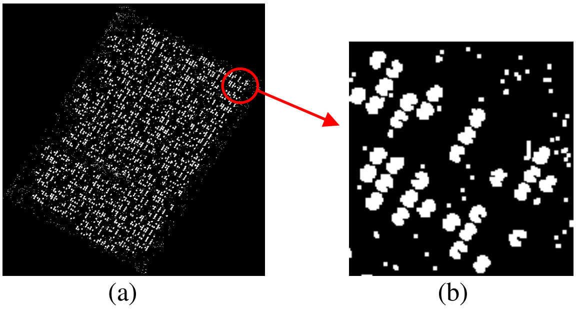


Fig. 5. (a) Segmented Braille image after dilation operation image, and (b) Enlarging some cells from the output image

2. All dots and connected components that have area less than dpi/2 white pixels and greater than 3*dpi white pixels are removed to remove the noise and undesirable components (See Fig. 6).

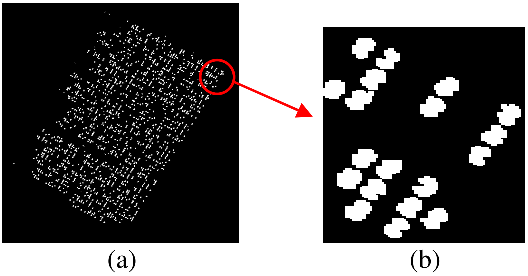

(a) (b)

Fig. 6. (a) Dilated segmented Braille image after removing the noise and undesirable dots, and (b) Enlarging some cells from the output image

3. Dilation operation is applied two times on dilated segmented Braille image after removing the noise and undesirable components to merge the dots on cells as connected components. The objective of this step is to access any cell as connected component instead of accessing each dot on all cells and search for neighboring dots that are irregular due to the skewness of Braille page on the scanner. Fig. 7 shows the Braille image after applied this operation.

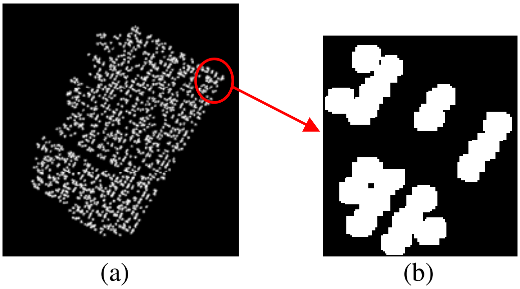

(a) (b)

Fig. 7. (a) Dilated segmented Braille image after applying the dilation operation two times, and (b) Enlarging some cells from the output image

4. Each connected component is localized by a bounding box. This box will enable us to access any cell on the dilated segmented Braille image that resulted from the second operation (See Fig. 8).

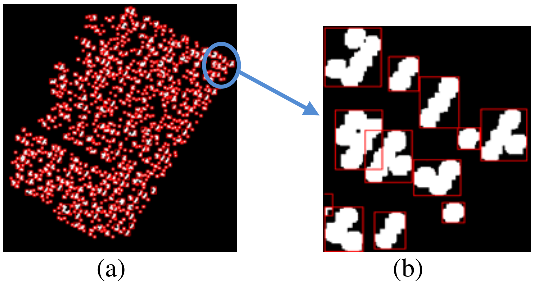

(a) (b)

Fig. 8. (a) Dilated segmented Braille image after localizing each connected component by a bounding box, and (b) Enlarging some cells from the output image

5. In this task, we determine the *x-axis* and *y*-axis coordinate of the center of first and last dots for all Braille cells containing three dots on the same straight line; then, we draw a line between them (See Fig. 9).

   The outputs of this step are a set of simple Braille cells containing three dots on the same straight line.

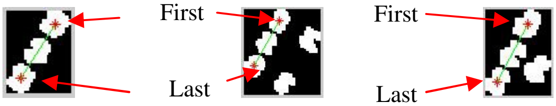


Fig. 9. Samples of detected Braille cells after localizing the center of first and last dots

2.3 Skewness Angle Estimation

The set of simple Braille cells $c_1,c_2,c_3,\ldots,c_n$ which contain three dots $d_1(x_1,y_1)$, $d_2(x_2,y_2)$ and $d_3(x_3,y_3)$ on the same straight line resulting from the previous step are used here to estimate the skewness angle of Braille image. Estimating skewness angle for each Braille cell is achieved based on calculating the slope of straight line between $d_1(x_1,y_1)$ and $d_3(x_3,y_3)$ where $d_2(x_2,y_2)$ is located in the middle for each detected Braille cell $c_1,c_2,c_3,\ldots,c_n$ using Equation 1. And then the angle of each slope $m_1, m_2, m_3,\ldots, m_n$ can be calculated by Equation 2. After that, the median of the calculated angles $(\theta_1,\theta_2,\theta_3,\ldots,\theta_n)$ is the estimated skewness angle of Braille image.

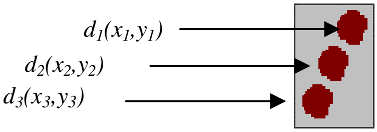


Fig. 10. One detected Braille cell (c_1) from a set of Braille cells(c_i) contains three coordinates (x_1,y_1), (x_2,y_2), and (x_3,y_3) of three dots d_1,d_2,d_3

$$\tan(\theta) = m = -\frac{(y_2 - y_1)}{(x_2 - x_1)} \tag{1}$$

$$\theta = \tan^{-1}(m) \tag{2}$$

Before calculating the slope of straight line for each Braille cell containing three dots, we rotate the segmented Braille image by 90^0, to know if the Braille image was rotated by 1^0 or -1^0. This is done because some rotation angles of Braille image have the same estimation angle (i.e. the Braille image that is rotated by zero and one degree have the zero estimation angle) due to the effect of segmentation step on the size of Braille dots that led to non-regular centers of Braille cells. The angle 90^0 is selected to correct the adjusted Braille image because most dots of Braille cells in this angle have the same *x-axis*. Therefore, if the adjusted Braille image is not rotated, the estimation skewness angle will be the same angle 90^0. After estimating the skewness angle of

Braille image, if this angle is 91^0 then the skewness angle of Braille image is 1^0, otherwise, if the skewness angle of Braille image is 89^0 then the skewness angle of Braille image is -1^0. Otherwise, if the skewness angle of Braille image is the estimated angle. Fig. 11 shows the flowchart of this step.

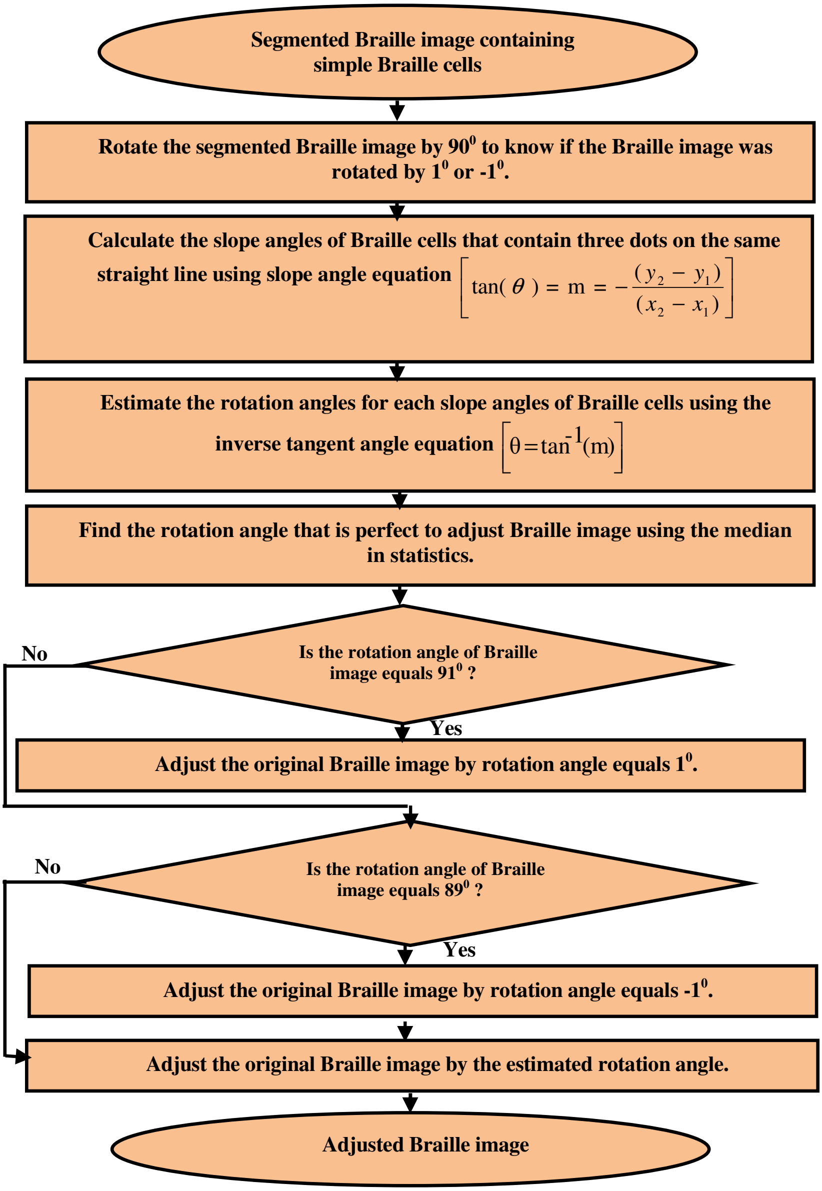
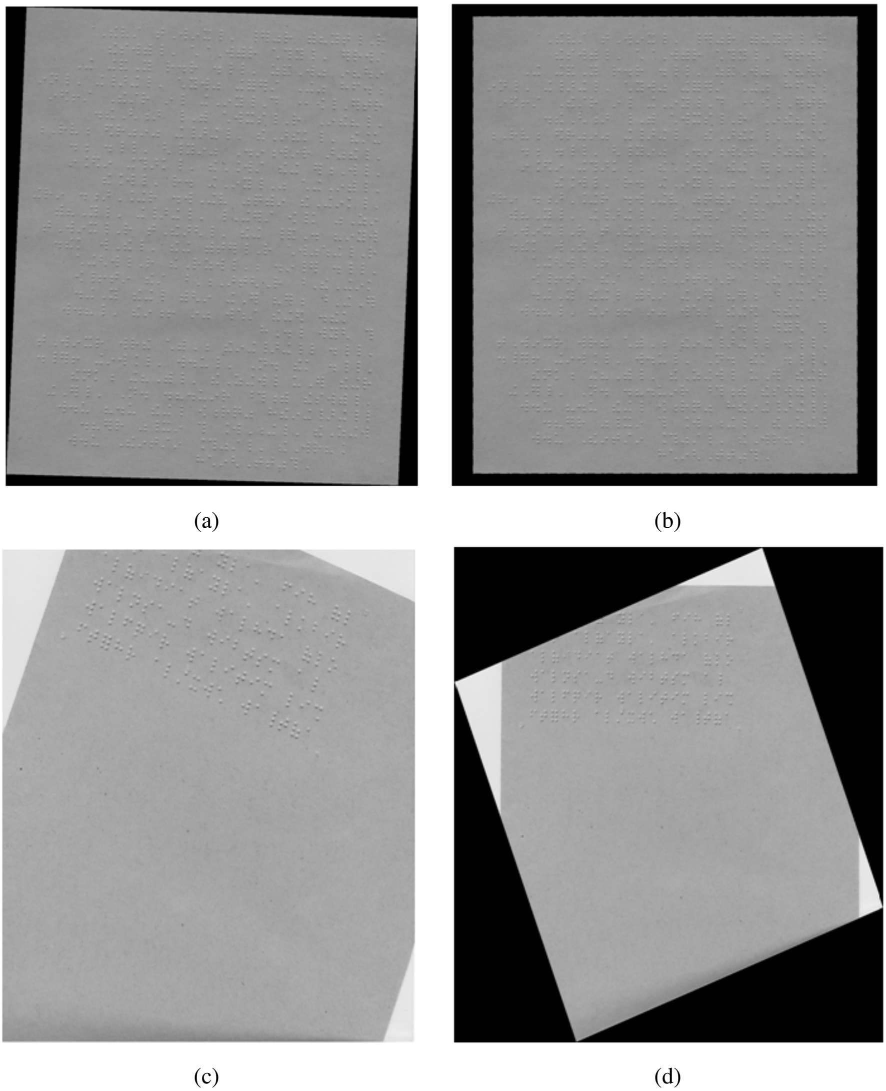

Segmented Braille image containing simple Braille cells
Rotate the segmented Braille image by 90^0 to know if the Braille image was rotated by 1^0 or -1^0.
Calculate the slope angles of Braille cells that contain three dots on the same straight line using slope angle equation $\left[\tan(\theta) = m = -\frac{(y_2 - y_1)}{(x_2 - x_1)}\right]$
Estimate the rotation angles for each slope angles of Braille cells using the inverse tangent angle equation $\left[\theta = \tan^{-1}(m)\right]$
Find the rotation angle that is perfect to adjust Braille image using the median in statistics.
Is the rotation angle of Braille image equals 91^0 ?
No
Yes
Adjust the original Braille image by rotation angle equals 1^0.
Is the rotation angle of Braille image equals 89^0 ?
No
Yes
Adjust the original Braille image by rotation angle equals -1^0.
Adjust the original Braille image by the estimated rotation angle.
Adjusted Braille image

Fig. 11. Skewness angle estimation flowchart

3 Experimental Results

Our proposed approach is implemented using Matlab version 7.5 and evaluation process is done using several single and double sided Braille documents which have been scanned by a flatbed scanner and rotated by different angles. The results show that the method is able to correct the skewed Braille images almost 100% accurately in a faster way. Fig. 12 shows the results of proposed approach on skewed Braille images.



(a) (b)

(c) (d)

Fig. 12. (a) and (c) original Braille images which are rotated by 2^0 and 22^0 angles; (b) and (d) resulted Braille images which are adjusted by -2^0 and -22^0 estimated angles

4 Conclusion

In this work, we developed a new approach for adjusting skewness of Braille image in optical Braille recognition system based on image processing techniques. This approach is able to detect the skewed Braille image and adjust it efficiently. The estimated rotation angle which is resulting from our approach is in the range from -90^0 to 90^0 and by symmetric property we can get the other angles for any rotation of Braille images but a context analysis is necessary to identify if the Braille images are rotated upside down. The results show that the method is able to correct the skewed Braille images almost 100% accurately in a faster way. This is has been increased the performance of Braille recognition system to recognize all Braille cells in Braille images.

References

1. Dubus, J., Benjelloun, M., Devlaminck, V., Wauquier, F., Altmayer, P.: Image processing techniques to perform an autonomous system to translate relief Braille into black-ink, called: Lectobraille. In: Proceedings in the IEEE Engineering in Medicine and Biology Society, pp. 1584–1585 (1988)
2. Mennens, J., Tichelen, L.V., Francois, G., Engelen, J.: Optical recognition of braille writing using standard equipment. IEEE Trans. on Rehabilitation Eng. 2(4), 207–212 (1994)
3. Ritchings, R.T., Antonacopoulos, A., Drakopoulos, D.: Analysis of Scanned Braille Documents. In: Dengel, A., Spitz, A.L. (eds.) Document Analysis Systems, pp. 413–421. World Scientific Publishing Company, Singapore (1995)
4. Blenkhorn, P.: A System for Converting Braille into Print. IEEE Trans. on Rehabilitation Engineering 3(2) (1995)
5. Hentzschel, T.W., Blenkhorn, P.: An Optical Reading Systems for Embossed Braille Characters using a Twin Shadows Approach. Journal of Microcomputer Apps., 341–345 (1995)
6. Oyama, Y., Tajima, T., Koga, H.: Character Recognition of Mixed Convex- Concave Braille Points and Legibility of Deteriorated Braille Points. System and Computer 28(2) (1997)
7. Ng, C., Ng, V., Lau, Y.: Regular feature extraction for recognition of Braille. In: Proceedings in Third International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 1999, pp. 302–306 (1999)
8. Optical Braille Recognition System User Manual. Version 3.7 (January 2004), http://www.indexbrailleaccessibility.com/downloads/obr/obrman37.pdf
9. Murray, I., Dias, T.: A portable device for optically recognizing Braille. part i: hardware development. In: The Seventh Australian and New Zealand Intelligent Information Systems Conference, pp. 129–134 (2001)
10. Murray, I., Dias, T.: A portable device for optically recognizing Braille. part ii: software development. In: The Seventh Australian and New Zealand Intelligent Information Systems Conference, pp. 141–146 (2001)

11. Wong, L., Abdulla, W., Hussmann, S.: A Software Algorithm Prototype for Optical Recognition of Embossed Braille. In: 17th Conference of the International Conference in Pattern Recognition, Cambridge, UK, pp. 23–26 (2004)
12. Antonacopoulos, A., Bridson, D.: A robust Braille recognition system. In: Marinai, S., Dengel, A.R. (eds.) DAS 2004. LNCS, vol. 3163, pp. 533–545. Springer, Heidelberg (2004)
13. Falcon, N., Travieso, C.M., Alonso, J.B., Ferrer, M.A.: Image Processing Techniques for Braille Writing Recognition. In: Moreno Díaz, R., Pichler, F., Quesada Arencibia, A. (eds.) EUROCAST 2005. LNCS, vol. 3643, pp. 379–385. Springer, Heidelberg (2005)
14. Namba, M., Zhang, Z.: Cellular Neural Network for Associative Memory and Its Application to Braille Image Recognition. In: Proc. of IJCNN 2006, pp. 4716–4721 (2006)
15. Al-Salman, A.S., Al-Ohali, Y., Al-Kanhal, M., Al-Rajih, A.: An Arabic Optical Braille Recognition System. In: ICTA 2007, Hammamet, Tunisia, April 12-14 (2007)
16. Tai, Z., Cheng, S., Verma, P.: Braille Document Parameters Estimation for Optical Character Recognition. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Remagnino, P., Porikli, F., Peters, J., Klosowski, J., Arns, L., Chun, Y.K., Rhyne, T.-M., Monroe, L. (eds.) ISVC 2008, Part II. LNCS, vol. 5359, pp. 905–914. Springer, Heidelberg (2008)
17. El-Zaart, A., Ziou, D.: Statistical Modeling of SAR Images. International Journal of Remote Sensing 28(10), 2277–2294 (2007)

Multilevel Thresholding Method Based on Aggressive Particle Swarm Optimization

Habibullah Akbar, Nanna Suryana, and Shahrin Sahib

Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka

Abstract. The multilevel thresholding problem based on Otsu criterion is examined in this paper. The problem is the computational time of Otsu method increases exponentially according to number of thresholds dimension. A new variant of PSO namely AgPSO is proposed. The contribution to the original PSO is highlighted by introducing an aggressive behavior to the movement rule. The proposed algorithm performance was evaluated based on three natural images. The experimental results showed the proposed PSO variant was more efficient according to other PSO variants.

Keywords: Multilevel Thresholding, Image Segmentation, Particle Swarm Optimization, Aggressive Behavior, Otsu Criterion.

1 Introduction

The advancement in image analysis and interpretation offers the vast area of research fields, not only limited to object recognition, industrial automation and remote sensing, but also telemedicine and exploratory observation. To all these applications, image segmentation is the critical step to delivers a successful application.

There have been large numbers of available image segmentation method in the literature. In fact, thresholding is one of the most simplest and popular techniques for image segmentation [1]. Among them, Otsu method, addressed as optimization problem, received more attention and widely studied experimentally [2]. Although the method perform well for most bilevel or binary thresholding application, however, the optimizing procedure becomes more and more complicated for multilevel thresholding which is required for advance application [3]. On the other hand, several bio-inspired algorithms, with the advantage of its reinforcement learning ability, have been proposed within last two decades ago. Among them, GA and PSO are two of most widely used algorithms [8]. According to [4], PSO is more efficient and precise for thresholds number greater than two. However, conventional PSO may easily get trapped into a local optimum when tackling complex problems [5].

This paper presents a new variant of PSO namely AgPSO to speed up the process of finding both local and global optimum solution. The aggressive behavior as a new velocity component is added to the PSO movement rule to help the swarm population flocking the target more aggressively. This way, each individual of swarm has the greater tendency to search locally faster and expected to search globally broader.

J.M. Zain et al. (Eds.): ICSECS 2011, Part I, CCIS 179, pp. 747–757, 2011.

Three benchmark images were used to evaluate the performance of the proposed algorithm due to multilvel thresholding problem. The experimental results showed that the proposed PSO variant was more efficient in comparison to other PSO variant. The remaining content of the paper is organized as follows. In Section 2, conventional PSO algorithm and the proposed AgPSO algorithm detailing the aggressive behaviour is briefly described. Section 3 gives the description of multilevel thresholding problem based on Otsu criterion. Section 4 presents the experimental result to evaluate the performance of the proposed algorithm. Finally, the conclusion is given in Section 5.

2 Aggressive Particle Swarm Optimization

2.1 Overview of Particle Swarm Optimization

The collective behavior of social animal agents within a population, such as bird flocking or fish schooling, inspire researches about the idea of swarm intelligence. The simplification of how bird or fish agents adjust their physical movement seeking food, mates and avoiding predators becomes a fruitful optimization tool. The method establishes the idea of basic principle in swarm intelligence, particularly aim to make the population carry out adaptive and simple computation models. On the other hand, the successful of vast application such a bio-inspired computing techniques do not reflect the real mechanism in the nature. Indeed, we just borrow some ideas based on our understanding due to natural mechanism to solve a real world problem. In fact, the real nature mechanism is too complex (where simple model can only approach simple system) and too perfect (where no mistake can take into account) to be modeled.

In Kennedy and Eberhart original version, the PSO basically has three components, the momentum, cognitive and social components [6]. In *D-dimensional* search-space, the actual particle position can be denoted as $x_i = (x_{i1}, x_{i2}, ..., x_{iD})$. Each particle travels towards the optimum solution, through the search-space, directed by best previous positions. Note particle and agent have the same meaning in this article; it described an active individual that is working together towards optimum solution.

The momentum component is based on previous velocity. The cognitive component is based on each particle experience, represented by individual previous best position p_i. The social component is based on entire swarm experience, represented by previous neighborhood best position p_{best}. Hence, the velocity and best visited position of particle can be represented as $v_i = (v_{i1}, v_{i2}, ..., v_{iD})$ and $p_i = (p_{i1}, p_{i2}, ..., p_{iD})$ respectively. Then, the next position of particles can be updated heuristically based on the following movement rule.

$$x_{i+1} = x_i + v_{i+1} \tag{1}$$

The next particle velocity v_{i+1} is a linear combination of momentum, cognitive and social components as shown in equation below.

$$v_{i+1} = v_i + a_1 \cdot rand_1() \cdot (p_i - x_i) + a_2 \cdot rand_2() \cdot (p_{best} - x_i) \tag{2}$$

The variable a_1 and a_2 are cognitive and social acceleration learning rate respectively, commonly both values are set as 2 to allow particles to "overfly" the target about half the time [9]. The variable $rand_1$, $rand_2$ are random numbers between 0 and 1 with uniform distribution. The variable velocity v_i can also be restricted into a boundaries between [$-v_{max}$, v_{max}] whereas the v_{max} value can be set as 2 [9]. According to [13], the PSO pseudo code for maximization can be represented as follows.

```
Initialize population
  Do
    For i = 1 to Population Size
    if f($x_i$) > f($p_i$) then $p_i$ = $x_i$
    if f($p_i$) > f($p_{best}$) then $p_{best}$ = $p_i$
     For d = 1 to Dimension begin
     $v_{id}$ = $v_{id}$ + $a_1$·rand()·($p_{id}$ - $x_{id}$) + $a_2$·rand()·($p_{best}$ - $x_{id}$)
     $v_i$ = max($v_{min}$, min($v_{max}$, $v_{id}$))
     $x_{id}$ = $x_{id}$ + $v_{id}$
     next d
    next i
  Until termination criterion is met
end.
```

Omran mentioned that the p_{best} component and the fast rate of information exchange between particles provokes particle into a single point [15]. Therefore, the conventional PSO is easily trapped into local optima when tackling complex problems [5]. Several variants of PSO have been proposed to improve original PSO which modified the movement rule. Among them, two of PSO variants are discussed in this paper.

The first variant was introduced by Shi and Eberhart. They add a new parameter to PSO namely variable inertia weight w, also called as momentum learning rate [9]. For comparison, we name this variant as wPSO. Hence, the movement rule becomes.

$$v_{i+1} = w \cdot v_i + a_1 \cdot rand_1() \cdot (p_i - x_i) + a_2 \cdot rand_2() \cdot (p_{best} - x_i) \tag{3}$$

The variable w can be represented as weighting function below.

$$w = w_{\max} - \frac{(w_{\max} - w_{\min})xIter}{Iter_{\max}} \tag{4}$$

where
w_{max} = initial weight
w_{min} = final weight
$Iter$ = actual iteration
$Iter_{max}$ = number of maximum iteration

The variable w linearly decreases from w_{max} equal to 0.9 down to w_{min} equal to 0.4 for each subsequent movement [7]. At first, higher value of w helps particles to explore global search and then lower value of w brought particles focus on local search.

Another variant of PSO was proposed by Clerc, with the introduction of constriction factor to ensure solution convergence [11]. The modified movement rule of Clerk modification can be represented as follows [10].

$$v_{i+1} = K \cdot [v_i + a_1 \cdot rand_1() \cdot (p_i - x_i) + a_2 \cdot rand_2() \cdot (p_{best} - x_i)] \tag{5}$$

where
$$K = \frac{2}{\left|2 - \alpha - \sqrt{\alpha^2 - 4\alpha}\right|}, \alpha = a_1 + a_2, \alpha > 4 \tag{6}$$

Eberhart and Shi (1995) found that particles can travel faster towards the optimum solution if the constriction factor is set as 4.1 and the v_{max} is set at the maximum search range x_{max} [10]. For comparison, we name this variant as *K*PSO.

2.2 Aggressive Behaviour in Particle Swarm

By nature, the aggressive behavior is one of the most fundamental characteristic of distinctive creatures. As explained by Fink, the aggressive behavior is highly functional form of social behavior in almost vertebrate animal [14]. In fact, offensive agent is more aggressive and proactive to locate resources while non-aggressive agent has less value for the swarm. Therefore, the aggressiveness behavior is expected to provoke agent to search optimum solution faster than normal agent. The aggressiveness itself can be embedded into PSO in many ways.

One way is to incorporate self-adaptive as new velocity component into the movement rule. In our model, the aggressive behavior is characterized by two concepts, non-aggressive agent must learn how to be more aggressive and aggressive agent can share their experience to non-aggressive agent. Non-aggressive particles should have higher acceleration learning rate while aggressive particles have lower acceleration learning rate as shown in Fig. 1.

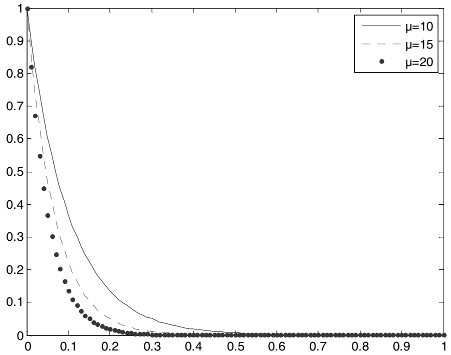


Fig. 1. Aggressive learning rate a_3 as function of aggressiveness Δx_i

The aggressiveness variable Δx_i is simply based on how far the particle move from its previous position. Non-aggressive particles have to learn being more aggressive through the experience of p_{best}. However, we found the aggressive behavior sometimes is destructive due to its nature of randomization. Hence, the aggressive is only applicable for low quality particles. Several experiments have been tested to meet the requirement of concepts above. Thus, we found one shape of the new movement rule can be defined as follows.

$$v_{i+1} = w \cdot v_i + a_1 \cdot rand_1() \cdot (p_i - x_i) + a_2 \cdot rand_2() \cdot (p_{best} - x_i) + v_\alpha \tag{7}$$

$$v_\alpha = a_3 \cdot rand_3() \cdot (\Delta x_i \cdot p_i) \tag{8}$$

$$a_3 = e^{-\mu \cdot \Delta x_i} \tag{9}$$

where
v_α = aggressive velocity
a_3 = aggressive learning rate
μ = decay constant
Δx_i = aggressiveness

Notice the aggressive acceleration learning rate a_3 is similar to the mutation rate mentioned in [16].

3 Multilevel Thresholding Selection of Image Segmentation

In this paper, the Otsu criterion [3] is used as the objective function because it is not only most popular but also effective [4], [6].

3.1 Bilevel Thresholding

Consider the input image can have any gray-level value between [0, 1, 2, ..., L – 1]. The occurrence probability of gray-level i can be denoted and normalized as follows.

$$p_i = \frac{n_i}{N}, \qquad 0 \le p_i \le 1 \tag{10}$$

where n_i is the number of pixels at gray-level i and N is the total number of pixels in the input image which satisfies the following condition.

$$N = n_0 + n_1 + n_2 + ... + n_{L-1.} \tag{11}$$

Suppose that the input image has an object (foreground) and its background. Therefore, the image pixels can be separated into two classes C_f and C_b by a threshold at a gray-level T. The foreground class C_f and the background class C_b consist of gray-levels range between [0,T] and [T+1, L-1] respectively.

The objective function of Otsu criterion can be represented as follows [3].

$$\sigma^2(T) = \frac{[\mu_T w(T) - \mu(T)]^2}{w(T)[1 - w(T)]} \tag{12}$$

The optimal threshold T^* value that maximizes the separability of the foreground and the background can be obtained through the following equation.

$$\sigma^2(T^*) = \arg\max_{0 \le T \le L-1} \sigma^2(T) \tag{13}$$

3.2 Multilevel Thresholding

The bi-level thresholding is simple and straightforward to compute. Unfortunately, this technique is heavily relies on the assumption that only two classes exist in the image. Since the real world image can have any number of objects, then it is necessary to use multilevel thresholding to detect more foregrounds.

Let m number of thresholds: $0 \leq T_1 \leq T_2 \leq ... \leq T_m \leq L$-1 which separate input image into m classes. Hence, the objective function becomes a function of m variables. The optimal thresholds T_1^*, T_2^* , ... ,T_m can be obtained by maximizing σ^2 through following equation.

$$\sigma^2(T_1^*, T_2^*, ..., T_m^*) = \max_{0 \leq T_1 \leq T_2 \leq ... \leq T_m \leq L-1} \sigma^2(T_1, T_2, ..., T_m) \tag{14}$$

However, the original Otsu method increases the processing time exponentially due to increment of the thresholds level. Indeed, bio-inspired computing can be employed to find the thresholds values in reasonable amount of time.

3.3 AgPSO Step-by-Step Procedure

Initially, the particles positions are generated randomly in between [0, L] where L is 255 according to 8-bit quantization of the standard input image [19]. Then, the proposed algorithm is described in five steps as follows.

Step 1. Set parameters input: N_{iter}, a_1, a_2, μ, v_{min}, v_{max} and generate swarm population with their positions and velocities randomly.

Step 2. Compute the fitness value of D particles according to Otsu criterion in equation (14).

Step 3. Update best position value of each particle and best position value of swarm.

Step 4. Update the position and velocity for each particle according to equation (**7**), (8), (9), and (1) to include the aggressive behavior.

Step 5. Repeat steps 2 to 4 until pre-defined stopping condition is achieved

Notice the major contribution of AgPSO is highlighted in *Step 4*.

4 Experiments and Performance Evaluation

The proposed algorithm was tested on three natural images; (i) Mud, (ii) Rainbow, and (iii) Twig, where can be found in [17]. The images are shown in Fig. 2.

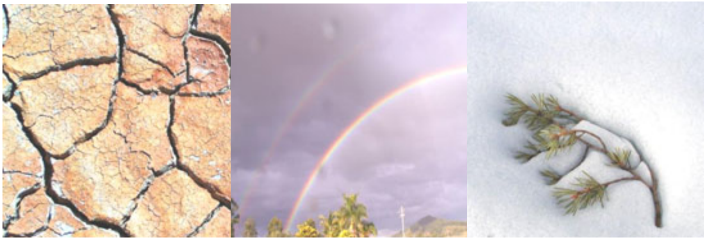
In this paper, the AgPSO is compared to PSO, *w*PSO, and *K*PSO. Several parameters have same value for fair comparison;

(i) $Iter_{max} = 100$.

(ii) $a_1 = a_2 = 2$.

(iii) $v_{max} = 2$ and $v_{min} = -v_{max}$.

(iv) Population size = 25.



(a) (b) (c)

Fig. 2. Original image and their histogram (a) Mud (b) Rainbow (c) Twig

The maximum iteration was chosen at 100 while population size was chosen at 25 according to [18]. Specific parameters values due to the PSO variants are given in Table 1.

Table 1. Parameter Setting

Algorithm	Additional Parameter
AgPSO	$\mu = 10$
*K*PSO	$\alpha = 4.1$, $v_{max} = x_{max}$
*w*PSO	$w_{max} = 0.9$, $w_{min} = 0.4$
PSO	-

Each were evaluated based on three different threshold level, $T_m = 2$, 3, and 4. The experiment was tested several times. The best result of the proposed algorithm performance and its comparison to other PSO variants are given in the following subsections.

4.1 Multilevel Thresholding Result

The average objective value corresponds to Otsu criterion is shown in Table 2. The AgPSO was better at $T_m = 3$ for Rainbow image but worse at $T_m = 4$. For other case, AgPSO has similar performance with *w*PSO and PSO. However, *K*PSO has the lowest performance for all problems. The correspondence threshold values are given in Table 3.

Table 2. Comparison of objective function value for aggressive PSO and others

Images	T_m	AgPSO	*K*PSO	*w*PSO	PSO
Mud	2	2414.1316	2375.6309	2414.1316	2414.1316
	3	2551.3532	2510.6970	2551.3532	2551.3532
	4	2609.3176	2548.0972	2614.5363	2614.5363
Rainbow	2	675.5899	647.9544	675.5899	675.5899
	3	709.4733	669.5988	705.4287	706.1001
	4	720.1576	677.7450	724.9927	722.2881
Twig	2	1233.8178	1200.1675	1233.8178	1233.8178
	3	1306.6195	1254.4280	1306.6213	1306.6213
	4	1340.4432	1304.8110	1343.7712	1343.7734

Table 3. Thresholds value obtained by aggressive PSO and others

Images	T_m	AgPSO	*K*PSO	*w*PSO	PSO
Mud	2	86,159	95,160	86,159	86,159
	3	72,135,175	76,141,176	72,135,175	72,135,175
	4	60,115,153,183	61,126,153,182	60,115,153,183	60,115,153,183
Rainbow	2	126,158	123,160	126,158	126,158
	3	113,135,161	116,148,167	113,135,161	113,135,161
	4	109,128,148,167	31,113,132,156	109,128,148,167	109,128,148,167
Twig	2	115,171	127,179	115,171	115,171
	3	94,145,179	98,155,189	94,145,179	94,145,179
	4	81,128,164,187	89,122,158,184	81,128,164,187	81,128,164,187

4.2 Stability Comparison of Algorithms

The stability of AgPSO and other algorithm were examined based on the standard deviation. The lower standard deviation means particular algorithm is more stable and vice versa. The standard deviation is represented as follows.

$$std = \sqrt{\sum_{i=1}^{\text{run}} \frac{(\sigma_i - \bar{\sigma})}{\text{run}}} \tag{15}$$

where
std = standard deviation
run = repeated times of each algorithm
σ_i = best objective value at i^{th} run
$\bar{\sigma}$ = average value of σ

In this paper, each algorithm runs 10 times. Table 4 showed the AgPSO, *w*PSO and PSO stabilities were same for Mud test image. The AgPSO was better for Rainbow test image at $T_m > 2$. However, for Twig image, *w*PSO and PSO stabilities were better than AgPSO and *K*PSO. Similar to objective value results, *K*PSO showed the lowest performance for all problems.

Table 4. Stability value obtained by aggressive PSO and others

Images	T_m	AgPSO	*K*PSO	*w*PSO	PSO
Mud	2	0	35.6076	0	0
	3	4.8E-13	32.0982	4.8E-13	4.8E-13
	4	4.8E-13	36.9389	4.8E-13	4.8E-13
Rainbow	2	1.2E-13	26.7822	1.2E-13	1.2E-13
	3	0	16.2350	8.7983	8.6657
	4	0.9260	41.1504	1.0051	6.1755
Twig	2	2.4E-13	17.8389	2.4E-13	2.4E-13
	3	0.0056	26.9735	2.3E-13	2.4E-13
	4	0.0439	24.6109	0.0067	0.0031

4.3 Efficiency Comparison of Algorithms

The computation time for the algorithms is given in Table 5. It is clearly for all problems and T_m, the AgPSO were faster in comparison to all PSO variants.

Table 5. CPU time comparison

Images	T_m	AgPSO	*K*PSO	*w*PSO	PSO
Mud	2	0.4511	0.6888	0.7382	0.7526
	3	0.6055	0.8319	0.8632	0.8661
	4	0.7484	0.9557	0.9788	0.9823
Rainbow	2	0.4157	0.6897	0.7499	0.7300
	3	0.5371	0.8402	0.8697	0.8984
	4	0.6872	0.9705	1.0035	1.0032
Twig	2	0.4194	0.7151	0.7768	0.7652
	3	0.5967	0.8651	0.8765	0.9023
	4	0.6295	0.9919	1.0053	1.0132

The image output using thresholds values obtained by AgPSO algorithm is presented in Fig. 3.

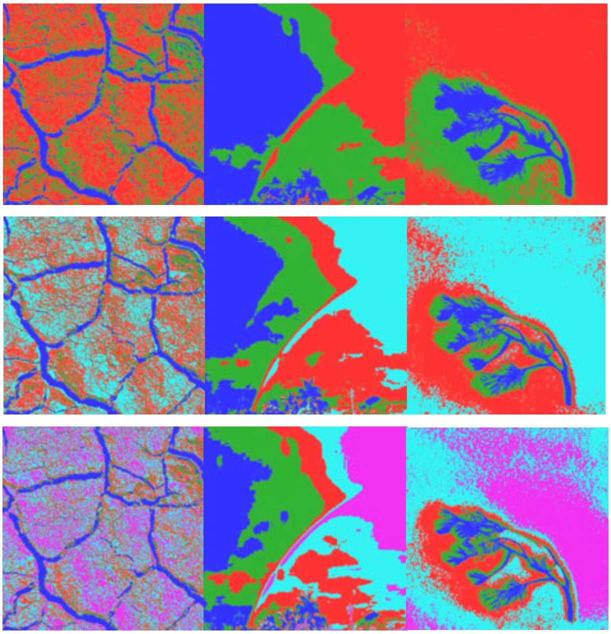


Fig. 3. Output images of AgPSO (top) represents T_m = 2 level, (middle) represents T_m = 3 level, and (bottom) represents T_m = 4 level

5 Conclusion

In this paper, a multilevel thresholding method based on new variant of Particle Swarm Organization namely AgPSO is presented. The proposed algorithm adds a new velocity component which is the aggressive behavior to the movement rule. The Otsu criterion was used as the objective function. Three natural images were employed to evaluate the proposed PSO variant for multilvel thresholding problems. The experimental results showed that the proposed PSO variant was faster in comparison to all PSO variants. However, the effectiveness and stability of the proposed algorithms has similar performance to *w*PSO and PSO. This means, the proposed PSO variant was not able to improve the PSO significantly. Similarly to other PSO variants, the proposed algorithm was easily being trapped into local convergence. Future exploration on the aggressiveness behavior is highly suggested to improve the effectiveness as well as the efficiency.

Acknowledgments. The authors thank for the support from Universiti Teknikal Malaysia Melaka for the Zamalah Scheme.

References

1. Pal, N.R., Pal, S.K.: A Review on Image Segmentation Techniques. Pattern Recognition 26(9), 1277–1294 (1993)
2. Xu, X., Xu, S., Jin, L., Song, E.: Characteristic Analysis of Otsu Threshold and its Applications. Pattern Recognition Letters 32(7), 956–961 (2011)
3. Otsu, N.: A Threshold Selection Method for Grey Level Histograms. IEEE Transactions on System, Man and Cybernetics 9(1), 62–66 (1979)
4. Hammouche, K., Diaf, M., Siarry, P.: A Comparative Study of Various Meta-Heuristic Techniques Applied to the Multilevel Thresholding problem. Engineering Applications of Artificial Intelligence 23(5), 676–688 (2010)
5. Zhao, L., Yang, Y.: PSO-based Single Multiplicative Neuron Model for Time Series Prediction. Expert Systems with Applications 36(2), 2805–2812 (2009)
6. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: IEEE International Conference on Neural Networks, Perth, pp. 1942–1948 (1995)
7. Shi, Y., Eberhart, R.C.: Empirical Study of Particle Swarm Optimization. In: Congress on Evolutionary Computation, Washington, pp. 1945–1950 (1999)
8. Grimaccia, F., Mussetta, M., Zich, R.E.: Genetical Swarm Optimization: Self-Adaptive Hybrid Evolutionary Algorithm for Electromagnetics. Transactions on Antennas and Propagation 55(3), 781–785 (2007)
9. Shi, Y., Eberhart, R.: A Modified Particle Swarm Optimizer. In: International Conference on Evolutionary Computation, Anchorage, pp. 69–73 (1998)
10. Eberhart, R.C., Shi, Y.: Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization. In: Congress on Evolutionary Computation, La Jolla, pp. 84–88 (2000)
11. Clerc, M.: The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization. In: Congress on Evolutionary Computation, Washington, pp. 1951–1957 (1999)

12. Eberhart, R., Shi, Y.: Special Issue on Particle Swarm Optimization. IEEE Trans. Evol. Comput. 8(3), 201–228 (2004)
13. Clerc, M., Kennedy, J.: The Particle Swarm - Explosion, Stability, and Convergence in a Multidimensional Complex Space. IEEE Trans. Evolutionary Computation 6(1), 58–73 (2002)
14. Fink, G.: Encyclopedia of Stress 1(A-D). Academic Press, San Diego (2000)
15. Omran, M.G.H.: Particle Swarm Optimization Methods for Pattern Recognition and Image Processing, PhD thesis. University of Pretoria (2005)
16. Castro, L.N.D., Zuben, F.J.V.: Learning and Optimization using the Clonal Selection Principle. IEEE Trans. Evol. Comput. 6(3), 239–251 (2002)
17. Free Nature Pictures, `http://www.freenaturepictures.com`
18. Chander, A., Chatterjee, A., Siarry, P.: A New Social and Momentum Component Adaptive PSO Algorithm for Image Segmentation. Expert Systems with Applications 38(5), 4998–5004 (2011)
19. Zhao, B., Chen, Y., Mao, W., Zhang, X.: Image Segmentation to HSI Model Based on Improved Particle Swarm Optimization. In: Huang, D.-S., Jo, K.-H., Lee, H.-H., Kang, H.-J., Bevilacqua, V. (eds.) ICIC 2009. LNCS (LNAI), vol. 5755, pp. 757–765. Springer, Heidelberg (2009)

Author Index