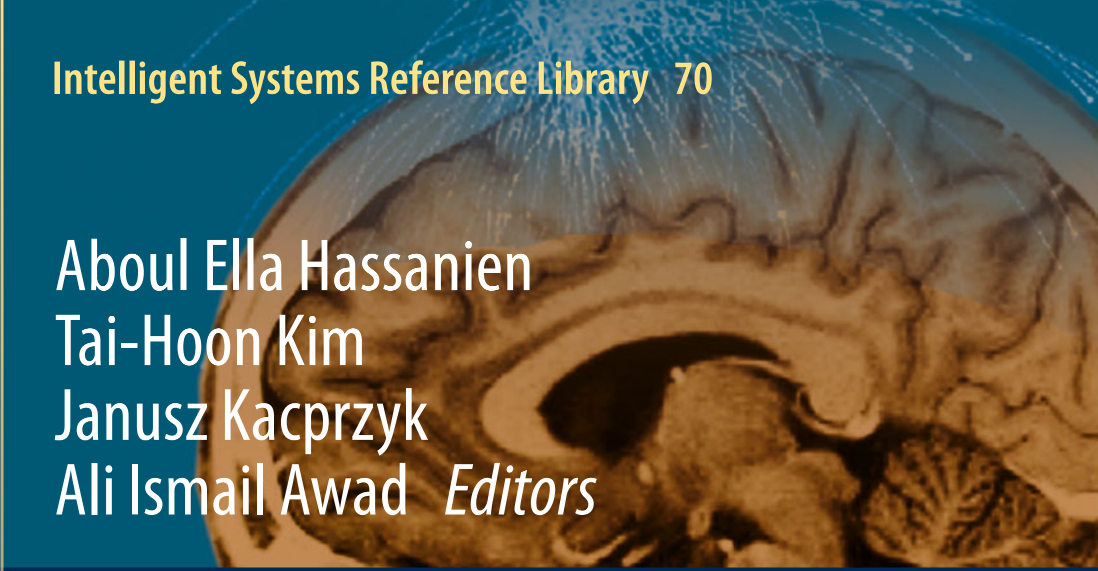


Intelligent Systems Reference Library 70

Aboul Ella Hassanien
Tai-Hoon Kim
Janusz Kacprzyk
Ali Ismail Awad *Editors*

Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations

Springer

Intelligent Systems Reference Library

Volume 70

For further volumes:
http://www.springer.com/series/8578

About the Series

The aim of this series is to publish a Reference Library, including novel advances and developments in all aspects of Intelligent Systems in an easily accessible and well structured form. The series includes reference works, handbooks, compendia, textbooks, well-structured monographs, dictionaries, and encyclopedias. It contains well integrated knowledge and current information in the field of Intelligent Systems. The series covers the theory, applications, and design methods of Intelligent Systems. Virtually all disciplines such as engineering, computer science, avionics, business, e-commerce, environment, healthcare, physics and life science are included.

Aboul Ella Hassanien · Tai-Hoon Kim
Janusz Kacprzyk · Ali Ismail Awad
Editors

Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations

Springer

Editors

Aboul Ella Hassanien
Faculty of Computers and Information
Cairo University
Cairo
Egypt

Tai-Hoon Kim
Multimedia Control and Assurance Laboratory
Hannam University
Taejeon
Korea, Republic of (South Korea)

Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

Ali Ismail Awad
Faculty of Engineering
Al Azhar University
Qena
Egypt

ISSN 1868-4394 ISSN 1868-4408 (electronic)
ISBN 978-3-662-43615-8 ISBN 978-3-662-43616-5 (eBook)
DOI 10.1007/978-3-662-43616-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014942527

Printed on acid-free paper

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

Preface

Recently, and due to the large diversity of the hackers and attacks, information security algorithms and application have become a crucial need for protecting almost all information transaction applications. Cyber security is an important branch of information security that is applied in computers and networks. This volume covers the cyber security branch, and aims to ensemble the up-to-date advances in various topics in cyber security, and reports how organizations can gain competitive advantages by applying the different security techniques in the real-world scenarios. The volume will provide reviews of the cutting-edge technologies, algorithms, applications, and insights for cyber security-based systems. In particular, the book will be a valuable companion and comprehensive reference for both postgraduate and senior undergraduate students who are taking a course in cyber security. The volume will be organized in self-contained chapters to provide greatest reading flexibility.

This edited volume comprises 22 chapters, including an overview chapter, which provides an up-to-date and state-of-the art research on Cyber Security and Cloud Services.

The book is divided into four main parts:

- ***Part-I:*** Bio-inspiring System in Cyber Security
- ***Part-II:*** Mobile Ad Hoc Networks and Key Managements
- ***Part-III:*** Biometrics Technology and Applications
- ***Part-IV:*** Cloud Security and Data Services

Part I entitled *Bio-inspiring System in Cyber Security* contains five chapters that deal with several techniques of bio-inspiring in application to watermarking.

In Chap. 1, *A Bio-inspired Comprehensive Distributed Correlation Approach for Intrusion Detection Alerts and Events in Complex Computer Networks*, Ayman M. Bahaa-Eldin presents a Distributed Agent Correlation Model (DACM) providing a scalable alert correlation for large-scale networks. The proposed model utilizes multiple distributed agents to provide an integrated correlation solution. The model can be extended by creating new correlation agents, and can be tailored to a protected network by selecting what agents to use and configuring each individual agent's parameters. DACM correlates alerts from IDSs with other information source such as INFOSEC tools and system and application log files.

The results show that DACM enhances both the accuracy and completeness of intrusion detection by reducing both false positive and false negative alerts; it also enhances the early detection new threats.

Chapter 2, *Bio-inspired Evolutionary Sensory System for Cyber-Physical System Security*, by Mohamed Azab and Mohamed Eltoweissy discusses the current challenges to Cyber-Physical Systems security, survey relevant solutions, and present a novel system, CyPhyMASC, towards realizing pervasive Monitoring, Analysis, Sharing, and Control (MASC) for enhanced CPS security. CyPhyMASC is a bio-inspired intrinsically resilient, situation-aware system utilized by Security Service Providers (SSPs) to provision MASC services to the Target-of-Security system (ToSs) comprising numerous heterogeneous CPS components.

In Chap. 3, *An Optimized Approach for Medical Image Watermarking,* Mona M. Soliman et al. present an authentication scheme which enhances the security, confidentiality, and integrity of medical images transmitted through the Internet by invoking particle swarm optimization with its modifications including particle swarm optimization, quantum particle swarm optimization, weighted quantum particle swarm optimization techniques in adaptive quantization index modulation, and singular value decomposition in conjunction with discrete wavelet transform (DWT) and discrete cosine transform (DCT). The experimental results show that the proposed approach yields a watermark which is invisible to human eyes, robust against a wide variety of common attacks, and reliable enough for tracing colluders.

In Chap. 4, *Bio-inspiring Techniques in Watermarking Medical Images: A Review* authored by Mona et al. is reviewed and employed different bio-inspiring and computational intelligence techniques such as restricted Boltzmann machine, deep belief network, rough sets, swarm intelligence, artificial immune systems, and support vector machines to tackle various problems in watermark embedding and extraction procedure of medical image watermarking.

In Chap. 5, *Efficient Image Authentication and Tamper Localization Algorithm Using Active Watermarking*, Sajjad et al. propose an image authentication system with accurate tamper localization ability. In the proposed algorithm a 16-bit watermark key has been created from each block of pixels in a host image; various types of tampering attacks were performed in order to evaluate the proposed tamper detection method. The proposed tamper localization algorithm is evaluated by calculating False positive (FP) and False Negative (FN), and the results are compared to several current image authentication techniques, and the comparison results clearly proved the high level of our tamper detection rate.

Part II entitled *Mobile Ad Hoc Networks and Key Managements* contains six chapters that describe several cyber security techniques including trusted Ant Colony Multi Agent-Based Routing Algorithm for Mobile Ad Hoc Networks, Cybercrime Investigation Challenges: Middle East and North Africa, Multilayer Artificial Immune System for Network Security, and Key Pre-distribution Techniques for WSN Security Services.

In Chap. 6, *TARA: Trusted Ant Colony Multi Agent Based Routing Algorithm for Mobile Ad-Hoc Networks*, Ayman M. Bahaa-Eldin presents a model for using

Multi Intelligent Agents and Swarm Intelligence for trusted routing in mobile ad hoc networks (MANETs) is presented. The new algorithm called TARA is proved more efficient in different ways than the existing protocols. TARA uses a methodology to establish an objective trust value for each node based on self-monitoring. The protocol is tuned to minimize the number of messages to be exchanged and trust value propagation is eliminated. Ant Colony Optimization is used to find the optimal route resulting in a better performance and lower latency. The advantages and complexity of TARA are examined and it is compared with other protocols. Simulation results show that TARA works better than standard and trusted routing protocols in the presence of malicious nodes.

In Chap. 7, *An Overview of Self-Protection and Self-Healing in Wireless Sensor Networks*, Tarek Gaber and Aboul Ella Hassanien review the autonomic computing paradigm which is the key concept of self-protection and self-healing. It then describes the need of these features for WSN application. It also gives an overview of the self-protection and self-healing of WSNs. It finally highlights a number of open issues about the self-protection and self-healing in the WSN environment.

Chapter 8, *Cybercrime Investigation Challenges: Middle East and North Africa* by Mohamed Sarrab et al. focuses mainly on highlighting the main challenges of Middle East and North Africa's countries in cybercrime investigation system by considering the recent developments in the continents Internet infrastructure and the need for information security laws in these particular countries. In addition, it focuses on the Internet infrastructure development, particularly to show how they might become vulnerable to cybercrime attacks.

In Chap. 9, *Multilayer Machine Learning-Based Intrusion Detection System*, Amira Sayed A. Aziz and Aboul Ella Hassanien present a basic model of a multi-layer system, along with the basics of artificial immune systems and network intrusion detection. An actual experiment is included, which involved a layer for data preprocessing and feature selection using Principal Component Analysis, a layer for detectors generation and anomaly detection using Genetic Algorithm with Negative Selection Approach. Finally, a layer for detected anomalies classification using decision tree classifiers is presented. The principal interest of this chapter is to benchmark the performance of the proposed multi-layer IDS system by using NSL-KDD benchmark data set used by IDS researchers.

In Chap. 10, *An Improved Key Management Scheme with High Security in Wireless Sensor Networks*, Satish kumar et al. present solutions to the security issue in relay network which were developed with validation and privacy in mind using the Incorporated Network Topological control and Key management scheme.

Chapter 11, *Key Pre-distribution Techniques for WSN Security Services* by Mohamed Mostafa M. Fouad and Aboul Ella Hassanien, gives an overview of the desired security services required for the WSNs, their threats model, and finally the details of different pairwise key distribution security techniques for distribution WSNs.

Biometrics Technology and Applications is the **Third Part** of this volume. It contains five chapters discussing some biometrics approaches.

Chapter 12, *Fusion of Multiple Biometric Traits: Fingerprint, Palmprint and Iris* by Manasa et al., explores multiple biometric fusion using two architectures: Parallel architecture and Hierarchical-cascade architecture. Multi-biometric recognition systems designed with hierarchical architecture not only are robust, fast, and highly secure but also mitigate problems.

Chapter 13, *Biometric Recognition Systems Using Multispectral Imaging* by Abdallah Meraoumia et al., describes the design and development of a multimodal biometric personal identification system based on features extracted from multispectral palmprint. The experimental results, obtained on a database of 400 users, show very high identification accuracy. They also demonstrate that combining different spectral bands does significantly reduce the accuracy of the system.

Chapter 14, *Electrocardiogram (ECG): A New Burgeoning Utility for Biometric Recognition* by Manal et al., is a comprehensive survey on the employment of ECG in biometric systems. An overview of the ECG, its benefits and challenges, followed by a series of case studies are presented. Based on the survey, ECG-based biometric systems can be fiducial or non-fiducial according to the utilized features.

Chapter 15, *Image Pre-processing Techniques for Enhancing the Performance of Real-Time Face Recognition System Using PCA*, by Behzad Nazarbakhsh and Azizah Abd Manaf proposed a recognition framework for human face recognition. The comprehensive proposed frameworks, include different processes, techniques, and algorithms that are used for human face authentication.

Chapter 16, *Biometric and Traditional Mobile Authentication Techniques: Overviews and Open Issues* by Reham Amin et al., reviews the current mobile authentication mechanisms: traditional and biometric, and their most commonly used techniques in the mobile authentication environment. In addition, the pro and cons of these techniques are highlighted. Moreover, a comparison among these techniques is conducted. The chapter also discusses the other techniques that could be much suitable for the current environment in mobile applications. Furthermore, it discusses a number of open issues of mobile authentication which needs further research in the future to improve the adoption of biometric authentication in the smart phones environment.

The **Final Part** of the book deals with *Cloud Security and Data Services*. It contains six chapters, which discusses Cloud Services Discovery and Selection, dynamic Self-Organizing Maps, and Hybrid Learning approach as well as Cloud Customers Self-Monitoring and Availability-Monitoring.

Chapter 17, *Cloud Services Discovery and Selection: Survey and New Semantic-Based System* by Yasmine Afify, presents a comprehensive survey on cloud services discovery and selection work. The survey highlights the need for a complete Cloud Services Discovery and Selection 25 system that assists SaaS users in finding the service that meets their functional and non-functional requirements efficiently. In response to this finding, a semantic-based SaaS cloud service publication, discovery, and selection system is proposed, which facilitates the process of mapping the user specified service request to real cloud offerings.

Chapter 18, *Data and Application Security in Cloud* by Nirnay Ghosh et al., reviews the recent works reported specifically in the area of *data and application security* relevant to cloud computing. Some works which use biologically inspired phenomenon to manage load balancing in cloud environment have also been studied. It provides an insight into the present state-of-the-art cloud security problems, proposed solutions, and identifies future research directions as well as scopes in various security issues.

Chapter 19, *Security Issues on Cloud Data Services* by Nour Zawawi et al., examines recent research related to data security and addresses possible solutions. Research in employing uncommon security schemes into Cloud environments has received increasing interest in the literature, although these schemes are neither mature nor rigid yet. This work aspires to promote the use of security protocols due to their ability to reduce security risks that affect users of data Cloud services.

Chapter 20, *A Reputation Trust Management System for Ad-Hoc Mobile Clouds* by Ahmed Hammam and Samah Senbel, proposes a reputation trust management system (TMC) for mobile ad hoc clouds. TMC system considers availability, neighbors evaluation, and response quality and task completeness in calculating the trust value for a node. The trust management system is built over Planet Cloud which introduced the term of ubiquitous computing. Eigen Trust algorithm is used to calculate the reputation trust value for nodes. Finally, performance tests were executed to prove the efficiency of the proposed TMC in terms of execution time, and detecting node behavior.

Chapter 21, *Secured and Networked Emergency Notification Without GPS Enabled Devices* by Qurban A. Memon, reviews various well-known systems in Location-Based Service Applications, followed by existing location detection methods and technologies. The privacy issues are also highlighted. The comparative study is done to highlight the strengths and weaknesses of various location detection approaches. Furthermore, various position equation methods are derived to estimate position location accuracy. In order to address privacy, a location detection system is developed for a limited geographical area as a case study. The simplicity and security of data transfer are also addressed by encryption using special-purpose microcontrollers. The standardization efforts for location identification are also summarized. Finally, the implemented system boosts privacy by (a) no data transfer except during emergencies and (b) encrypting data during transfer. Such a deployed system facilitates the safety management departments to address personal safety and security of the user in an operational area.

Chapter 22, *Towards Cloud Customers Self-Monitoring and Availability-Monitoring* by Sameh Hussein and Nashwa Abdelbaki, provides advices and guidelines for Cloud layers which can be under Cloud Customer control, to allow Cloud Customer contributes in Cloud infrastructure monitoring and controlling. In addition, they produce their developed monitoring tool to allow Cloud Customer contributes in service monitoring. It is for Cloud Customers to self-monitor the Availability as a metric of the outsourced IT service.

Finally, we are very much grateful to the authors of this volume and to the reviewers for their great effort by reviewing and providing useful feedback to the authors. The editors would like to express thanks to Dr. Thomas Ditzinger (Springer Engineering In house Editor) and the editorial assistant at Springer Verlag, Heidelberg, for the editorial assistance and excellent collaboration to produce this important scientific work. We hope that the reader will share our joy and will find the volume useful.

Egypt, April 2014 Aboul Ella Hassanien
Korea, Republic of (South Korea) Tai-hoon Kim
Poland Janusz Kacprzyk
Egypt Ali Ismail Awad

Contents

Part I Bio-inspiring Systems in Cyber Security

Part II Mobile Ad Hoc Networks and Key Managements

Part III Biometrics Technology and Applications

Part IV Cloud Security and Data Services

Part I
Bio-inspiring Systems in Cyber Security

Chapter 1
A Bio-inspired Comprehensive Distributed Correlation Approach for Intrusion Detection Alerts and Events

Ayman M. Bahaa-Eldin

Abstract In a complex network with intrusion detection and logging, a huge number of alerts and logs are generated to report the status of the network, servers, systems, and applications running on this network. The administrator(s) are required to analyze these pieces of information to generate an overview about the network, hacking attempts and vulnerable points within the network. Unfortunately, with the enormous number of alerts and recorded events that grows as the network grows, this task is almost impossible without an analysis and reporting model. Alerts and events correlation is a process in which the alerts produced by one or more intrusion detection systems and events generated from different systems and security tools are analyzed and correlated to provide a more succinct and high-level view of occurring or attempted intrusions and attacks. While the existing correlation techniques improve the intrusion detection results and reduce the huge number of alerts in a summarized report, they still have some drawbacks. This article presents a modular framework for a Distributed Agent Correlation Model (DACM) for intrusion detection alerts and events in computer networks. The framework supports the integration of multiple correlation techniques. It introduces a multi-agent distributed model in a hierarchical organization; correlates alerts from the IDS with attack signatures from information security tools and either system or application log files as other sources of information. The agent model is inspired by bio-distribution of cooperating members of a society to achieve a common goal. Each local agent aggregates/correlates events from its source according to a specific pattern matching. Correlation between multiple sources of information and the integration of these correlation agents together forms a complete integrated correlation system and reduces both false negative and false positive alerts, enhancing intrusion detection accuracy and completeness. The model has been implemented and tested using a set of datasets. Agents proposed models and algorithms have been implemented, analyzed, and evaluated to measure detection and correlation rates and the reduction rate of false positive and false negative

A. M. Bahaa-Eldin (✉)
Computer and Systems Engineering Department, Ain Shams University, Cairo, Egypt
e-mail: ayman.bahaa@eng.asu.edu.eg

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_1, © Springer-Verlag Berlin Heidelberg 2014

alerts. The results showed that DACM enhances both the accuracy and completeness of intrusion detection by reducing both false positive and false negative alerts; it also enhances the early detection new threats.

1.1 Introduction

Alert correlation is a promising intrusion detection technique that significantly improves security effectiveness by analyzing alerts from one or more IDSs and providing a high-level view of the attempted intrusions. Correlation components are procedures that aggregate alerts according to certain criteria; the aggregated alerts could have common features or could represent the steps of pre-defined scenario attacks. Correlation approaches are composed of a single component or a comprehensive set of components. The Correlation process is performed through several different stages including normalization, aggregation, verification, and correlation. Agents have been widely used in IDSs because they can be added and removed without having to restart the IDS, thereby providing flexible scalability. Agents are capable of performing simple functions on their own; a group of agents working together are able to derive complex results by exchanging information. Use of many agents reduces system overhead and avoids single point of failure. Finally, agents provide a multi-point detection and knowledge sharing capability.

1.1.1 IDS Correlation Problem Definition

Existing correlation techniques do not enable the integration of correlation from multiple information sources and are limited to operate in IDSs alerts and still have some limitations. Examples of these limitation are a high false detection rate; missing alerts in a multi-step attack correlation; alert verifications are still limited; Zero Day attacks still have low rates of detection; Low and Slow attacks and Advanced Persistent Threats (APTs) cannot be detected; and some attacks have evasion techniques against IDSs.

1.1.2 Article Organization

The remainder of this chapter is structured as follows. Section 1.2 presents a survey of intrusion detection correlation and related work, describing and giving an introduction to current correlation systems and distributed correlation techniques. Section 1.3 introduces a description of our distributed agent correlation model and its components. Detailed implementation of this model components and algorithms is presented in Sect. 1.4. In Sect. 1.5, detailed experimental results of applying the proposed model on the gathered dataset are presented. Finally, Sect. 1.5 presents conclusions and outlines future work.

1.2 Related Work

1.2.1 Distributed and Bio Inspired Intrusion Detection

Recently, many researchers [1–9] focused on the distributed agent models for intrusion detection and alert correlation. These models try to use a distributed computational model to avoid the need for a high performance computation to perform the detection and correlation tasks.

Other approaches use time series analysis [10, 11], while data mining is used in other approaches [12, 13].

1.2.2 Distributed Correlation

The model proposed in [14] describes a mission-impact-based approach to the analysis of security alerts produced by spatially distributed heterogeneous information security (INFOSEC) devices.

The model in [15] is proposed to achieve alert correlation which supplies information about the vulnerabilities. The proposal has a relational database that implements parts and the corresponding tables are automatically generated from data sources. IDS and vulnerability scanner fill the database with events.

In [16] it is analyzed how the control and estimation methods can be applied to correlate distributed events for network security. Based on those methods, a Process Query System has been implemented which can scan and correlate distributed network events according to users' high-level description of dynamic processes.

1.2.3 Agents in IDS and Correlation

In [17–22], agents have been used in IDSs, they were used as static or mobile adaptive agents for distributed IDS, moreover they were used for both host-based and network based IDS.

In [23, 24], distributed agent approach for alarm correlation was proposed to identify the root causes of network failures and fault identification. The proposed model presented a new distributed alarm correlation approach that effectively tackles the aforementioned data deficiencies.

1.2.4 Comprehensive Approach Model for IDS Alert Correlation

Comprehensive approach model for real-time alert correlation [25–27] has been produced as integrated solution. It consists of a set of correlation components which

Table 1.1 CAM components reduction rate for different datasets

Component	Data set							
	MIT/ LL1999	MIT/ LL2000	CTV	Defco9	Rome AFRL	Honeypot	Treasure hunt	Average
AF	6.38	0.01	0.04	28.43	0	0	0.09	4.99
AV	0	0	0	0	0	97.1	0	13.9
TR	77.1	6.61	31.5	60.25	69.8	71.8	99.9	59.5
ASR	0	0	0	0	0	0	2.27	0.32
FR	10.9	49.6	89.9	88.65	70.8	2.26	50.6	51.8
MSA	0	0.16	0.63	1.24	0	1.01	2.2	0.7
Count of used components	3	4	4	4	2	4	5	3.7

cover different correlation techniques. The study and analysis of the components reduction rate of the model and the effectiveness of each component on the different analyzed datasets are shown in Table 1.1.

The performance of IDS correlation is measured by reduction rate and correlation time. The correlation time for each component is calculated by the count of input alerts and the correlation time for each alert. The sequence order of correlation components affects the correlation process performance; the total time needed for the whole process depends on the number of processed alerts in each component.

Different reduction rates of each component simply affect the following component input of alert stream, i.e. the arrangement order of the components is a primary concern for each dataset to obtain faster correlation process. Different RR of a single component in different dataset is varying because of the difference of attack scenario, and target networks used for each dataset.

1.2.5 Distributed Agent Correlation Model

Distributed Agent Correlation model (DACM) is a multi-agent distributed correlation model in a hierarchical organization. It correlates alerts from IDS's and from other sources of information. Data sources for correlation are IDSs, system and application log files for different services provided by the system, and security tools. Examples of security tools are Firewalls, Vulnerability Scanners, and Performance monitors, while examples of application and system log files are audit system logs, FTP logs, SSH logs, http/https logs, and OS logs files.

Figure 1.1 shows the block diagram of DACM. DACM has its inputs from different information sources. DACM core correlates these input data using a set of local and central agents depending on the learning capability as well as knowledge base and security policy.

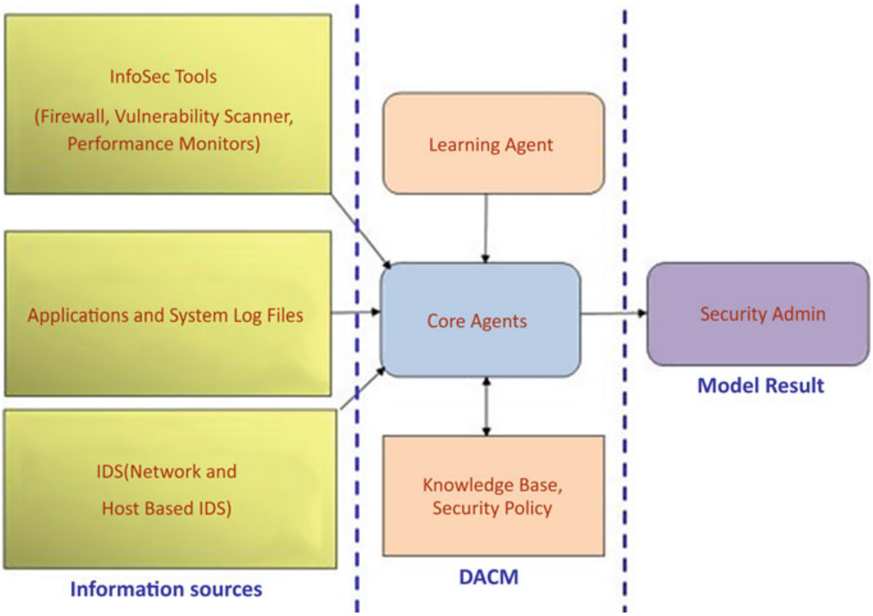

InfoSec Tools
(Firewall, Vulnerability Scanner, Performance Monitors)
Applications and System Log Files
IDS(Network and Host Based IDS)
Information sources
Learning Agent
Core Agents
Knowledge Base, Security Policy
DACM
Security Admin
Model Result

Fig. 1.1 DACM block diagram

Finally, DACM produces the output as a report for security administrator or automated response capability. DACM core agents consist of a set of correlation agents for different sources of information. These agents grouped into three main classifications: IDSs correlation agents for both network based and host based IDSs, INFOSEC tools agents for different available security tools in the system, and system and application logs agents for different auditing logs of operating system and available serves and application.

1.2.6 IDSs Correlation Agents

IDS alert correlation consists of a set of correlation components within certain structure. IDS sensors could be network based IDS or host based IDS, each of them produce its own alerts. Network based IDS has local correlation agent to correlate its alerts, in the other hand Host based IDS has its own correlation agent. The main IDS correlation agent correlates the output of network based correlation agent and host based correlation agent together and produce IDS's correlated alerts.

1.2.7 INFOSEC Tools Agents

Information Security tools are software which concern and analyze the information exchange within network traffic to determine which of this traffic trying to access resources as illegitimate behavior. Local agent to correlate/aggregate Firewall [28] router log files entries and group these entries for each blocked IP per day. Correlation is based on grouping the same blocked IP into a single record, a number of attempts field is added as an aggregation attribute.

Other local agent to correlate/aggregate Vulnerability Scanner [29] outputs according to scanner type. Mainly vulnerability that are related to the same target port and grouped together and the IP/Port combination are used to identify this alert.

Local agent to correlate/aggregate Performance monitor outputs according to specific use profile matching. This agent type analyzes the network performance monitoring tool outputs. Normal network performance thresholds values are defined in different intervals during the day resulting in a performance profile. This profile is constructed by supervised learning and stored in the knowledge base. If the monitored performance exceeds the threshold, an alert is generated. Multiple alerts for the same performance metric and period are grouped together, performance metric are used for traffic rate, usage ratio, congestion rate and so on.

1.2.8 System and Application Log Agents

System log consists of audit log files and application log files. These logs may be either access log or error log within each running application or service. DACM includes local agents to correlate each system log file contents according to specific pattern matching and comparing attack profiles which previously generated during learning period. These patterns and profiles are generated from a supervised learning process where normal and abnormal log patterns are identified by an operator or by the learning agent (LA).

Service and application logs have formal description in which they represent mathematical relation to determine the attack signature in their log files, FTP Agent as an example of these agents can be formally described as follows:

```
$\mathrm{A_{FTP}} \in$ FTP alerts
$\mathrm{FTP_{Entry}}$ (IP, Date, Time, Command, User) $\in$ FTP Log
$\{\mathbf{S}\}$: set of unauthorized FTP commands; $\{\mathbf{U}\}$: set of unauthorized users
$\forall$ Entry $\in$ FTP Log
***If*** command ($\mathrm{FTP_{Entry}}$) $\in \{\mathbf{S}\}$ or user ($\mathrm{FTP_{Entry}}$) $\in \{\mathbf{U}\}$
***Then*** FTP Entry is malicious, Produces AFTP
$\mathrm{FTP_{Entry}}$ (IP, Date, Time, Command, User) $\xrightarrow{\mathrm{FTPAttackTable}}$
**Else**
***Read*** next $\mathrm{FTP_{Entry}}$
```

1.2.9 DACM Central Agent

Main central agent correlates alerts from IDS's with outputs from other local agents from other information sources. Each local agent aggregates/correlates events from its source and modifies it to standard alert format and stores these results in its own table for main agent. Each agent has specific function and data to extract depending on its source of information and taking into account the network nature like impact analysis and prioritization. The output correlation of the central agent represents the final intrusion reports provided to security admin. These reports include even summary results or detailed intrusion attempts. Correlation between alerts from different sources is done based on a similarity function for the source of attack, attack type, and near time stamps.

1.2.9.1 Formal Description for Central Agent

Formal description is a method of presenting software systems in a way to facilitate further analysis for several metrics as completeness and correctness. In this section, a formal description for the central agent is given as an example to show the mathematical formula used in the agent.

$A_i \in \mathit{IDS\ \ alerts}$, $A_f \in \mathit{Firewall\ alerts}$, $A_L \in \mathit{log\ alerts}$;

A_i (source, time, destination, type) $\in$ *IDS alerts*
A_F (source, time, destination) $\in$ *Firewall alerts*
A_L (source, time, destination, type) $\in$ *Logs alerts*
$\forall\, alerts\ A_i$
***A_i* Is verified** alerts w.r.t. A_F
If source (A_i) = source (A_F) And Destination (A_i) = Destination (A_F)
And |Time (A_i) – Time (A_F)| <= $T_{threshold}$
Where $T_{threshold}$ is the minimum allowed difference time

***A_i* Is verified** alerts w.r.t. A_L
If source (A_i) = source (A_L) And Destination (A_i) = Destination (A_L)
And |Time (A_i) – Time (A_L)| <= $T_{threshold}$
Where $T_{threshold}$ is the minimum allowed difference time

***A_i* Is IDS only**
If attributes (A_i) <> attributes (A_L) OR
Attributes (A_i) <> attributes (A_F)

***A_i* Is Low and Slow attack**
A_i is IDS only and Count (source [A_i]) = 1 per day
And days (source [A_i]) > 3
$\forall\, alert\ A_L$,

A_L is negative alert w.r.t. A_I
If source (A_i) = source (A_L) and
Time (Ai) <> Time (A_L)
Or attributes (A_L) <> attributes (A_i)

A_L is reconnaissance
If count (A_L) > A_{Th}
Where Al is access count of specific IP / day and ; A_{TH} : allowed threshold access per day

1.2.10 Response Agent

Response agent is responsible for the suitable action against the attacker. The response agent interacts with the main central agent to respond depending on the final report. The final report contains summary of correlated attacks and the response agent suggests suitable response against these attacks. The attack response matches are included in specific tables according to the knowledge base in the system depending on historical behavior or learning systems. The implementation of the response agent is not included in this thesis and could be considered as important topic for future work.

1.2.11 Learning Agent

The proposed model has learning capability through LA which learns the precondition and post condition of new attacks as well as needed learning from other sources, in log files which of the log contents could be considered as attack signatures and which is considered normal signature. The model support adaptive learning by providing the contents which not previously indicated as either attack or normal signature. These contents are classified into three different types; similar to attack, similar to normal and unknown, later the system administrator can convert the type of these to either normal or attack signature. To enhance learning capability and trace attacker behavior, honey pot [30] agent could be used to learn new attacks and build attack profiles for more accurate knowledge base. In addition, learning capability could be extended to include learned attacks through sharing information with other external knowledge bases of similar systems.

1.2.12 The Knowledge Base and Security Policy

Knowledge base and security policy information represent the network nature and the needed authorization and behavioral profiles information. This information could be

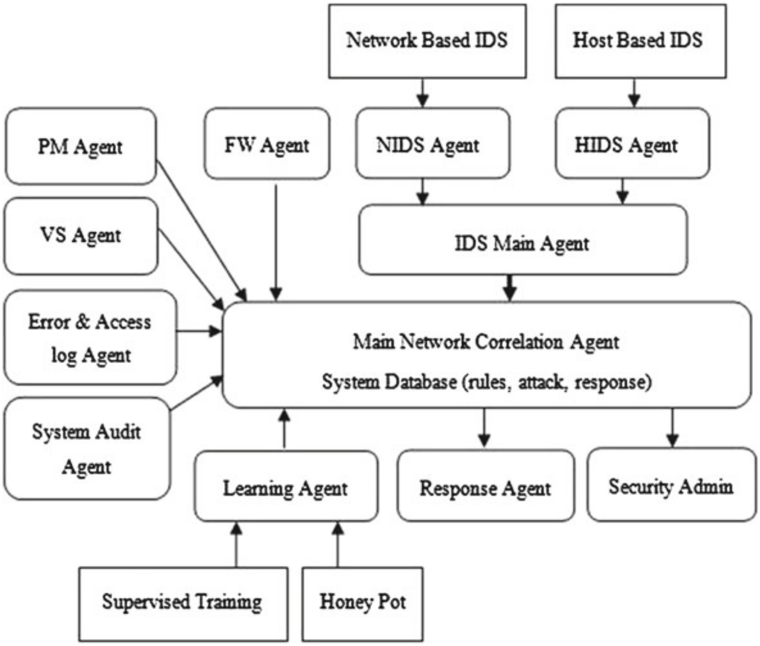


Fig. 1.2 DACM components structure

used by the individual local agents and central agent to discover the related attacks. These information are saved in database tables which include preconditions and post conditions for multi-step attacks, specific learning parameters, normal and attack profiles, attacks response matching, access control list which mention system users and their privileges. Some threshold values for profile matching such as performance measure, time of use, and network reconnaissance measures are also saved in the knowledge base.

1.2.13 DACM Components

DACM components structure is shown in Fig. 1.2, it consists of two levels, in the first level a set of agents which represent the model components, some of these agents represent local correlation component within IDS or other INFOSEC or system log files, and LA represents the learning capability in the mode. Each correlation agent read data from its source and matches it according to a specific template. A template is a particular pattern used in pattern recognition; it could be a characteristic pattern of attack by an individual or group of attackers. DACM agent’s algorithms are smart to avoid correlating important alerts; the new unknown alerts will be moved to second phase for further correlation and more analysis.

In second level, main correlation agent is considered as the central agent of the model which correlates the outputs of other agents to provide the whole picture of the network to the security administrator and/or provides the response agent with the suitable automated response action against the detected attacks according to predefined rules.

1.2.14 Implementation Scope and Performance Enhancement

The implemented model does not include all the previous described agents; it includes set of agents representing the different type of correlation because of the nature of collected data and the scope of this research. The implemented model includes the required component to prove the research concept. Network based IDS correlation agent have been implemented as an example of IDS correlation agent while host based IDS correlation agent was not implemented, Firewall agent have been implemented as an example for INFOSEC tools agent while vulnerability scanner and performance monitor agents were not implemented.

Correlation agents for error and access log files for different services within the network have been implemented as an example for system log files, while the system audit files correlation agent was not implemented. Supervised training with support of system admin has been implemented as an example learning capability, while learning using honey pot was not implemented. Finally response agent was considered out of the current scope for this research, it will be interesting research topic for future work. Analysis of packet dump of the network during the period of collecting data was performed manually with Wire Shark tool [31]; automation of capture data packets correlation was considered out of scope for current model implementation.

Several algorithms, parallelization, and enhancement are presented in detail in Sect. 1.3 for the sake of performance enhancement.

1.3 DACM Design and Algorithms

In this section several different individual agents and central agent implementation will be demonstrated. Different agent's algorithms for alerts and events correlation are presented.

1.3.1 IDS Alert Correlation

In this section, two IDS alerts correlation techniques are presented to enhance the correlation process presented in Comprehensive Approach Model (CAM) [25–27]. CAM results showed that the average time used to process one alert by different components varies depending on used dataset. Some components needs higher time to process one alert in a dataset while it needs shorter time to process one alert in another different dataset.

1.3.1.1 IDS Alert Correlation Performance Analysis

Results Analysis of CAM [25–27] reduction rate for each component against different datasets was presented in Table 1.1. This analysis showed that the sequence order of the correlation is not ideal and many components have not been used for most of the datasets which increases the correlation time needed to obtain effective correlation report for security administrator.

The correlation performance is measured by reduction rate and correlation time, the optimum correlation process has highest reduction rate in lowest correlation time.

Consider a N input alerts and O output alerts as a result of the correlation process, the reduction rate is defined as:

$$\mathrm{Reduction\ Rate\ (RR)} = 1 - \frac{\mathrm{O}}{\mathrm{N}}$$

For component (i), $\mathrm{RR_i}$: Reduction rate by Component i is defined as:

$\mathrm{RR_i} = 1 - (\mathrm{O_i}/\mathrm{N_i})$

The Total Reduction Rate:

$$\mathrm{RR} = \prod_{i=1}^{n} \mathrm{RR_i} \tag{1.1}$$

Equation 1.1 represents the total reduction rate of the model components.

For the *i*th component, $\mathrm{T_i}$: is the total time taken by component i to perform correlation and is a function of the count of input alert and time taken to analyze each alert.

$$\mathrm{T_i} = \mathrm{f}(\mathrm{c_i}, \mathrm{N_i})$$

$$\text{Total correlation time: } \mathrm{T} = \sum_{i=1}^{n} \mathrm{T_i} \tag{1.2}$$

The correlation time used in CAM model is represented in Eq. 1.2 which represent the sum of correlation time of all components even if they do not have effective reduction rate value.

1.3.2 Modified CAM Time

To eliminate the use of components with zero reduction rate affect, and to have optimum order of correlation components such that the components with higher reduction rates can be used before other components with lower reduction rate, we will assume the activity variable Xi is a Boolean variable that could be zero or one as follows

$$\mathrm{X_i} = \begin{cases} 0, & \mathrm{RR_i} = 0 \\ 1, & \mathrm{RR_i} > 0 \end{cases}$$

Giving the condition that $\mathrm{RR_i} > \mathrm{RR_{i+1,}}$

The above condition determines the sequence of correlation components, with the minimum total correlation time. This sequence require that components which have higher reduction rate to be used first before others which have lower reduction rate values. Modifying each component time Ti by its activity variable Xi eliminates component with zero reduction rate.

$$\mathrm{T_{opt}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{T_i X_i} \tag{1.3}$$

Equation 1.3 represents optimum total correlation time depending on the used components and datasets. The correlation time will be calculated for effective components which have reduction rate greater than zero value. The enhancing of the correlation process can be obtained by calculating the reduced time. It can be represented by the difference time between calculated T in Eq. 1.2 and calculated $\mathrm{T_{opt}}$ in Eq. 1.3 as follows:

$$\mathrm{T_{diff}} = \mathrm{T} - \mathrm{T_{opt}}. \tag{1.4}$$

1.3.2.1 Agent Based Correlation Model

The ABCM [32, 33] works through a learning phase and correlation phase. During the learning phase, LA learns the nature of the alert datasets and effective correlation components and their Reduction Rate (RR) and builds an Active Correlation Component List (ACCL). The ACCL contains the effective correlation components in descending order of their RR. Depending on the learning phase, the agent controls the correlation process during the correlation phase using the implemented ACCL. The order of correlation starts with components with higher RRs in ACCL followed by lower RRs until correlation by the last component in ACCL. The output of the first component will be the input of the second one which has the second highest RR, and so on till they reach the last component in ACCL.

The total correlation time by ABCM is calculated as follows:

$\mathrm{T_{ABCM}} = \mathrm{T_{learning}} + \mathrm{T_{correlation}}$ where $\mathrm{T_{correlation}}$ of ACCL components as optimal serial sequence without unneeded components and in proper order is as follow:

$$\mathrm{T_{correlation}} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{t_j}$$

where n is the count of ACCL components.

Total T_{ABCM} is much lower than total correlation time by CAM.

1.3.2.2 Dynamic Parallel Correlation Model

Dynamic Parallel Correlation Model DPCM [34] has parallel processing correlation to assure using the suitable component and its order. It consists of correlation stages with each stage consisting of a set of correlation components. The proposed model dynamically selects the optimum order of the needed correlation components depending on the working environment. The input to each stage is the output of the correlation component with the highest RR in the previous stage. In the next stage, the higher RR component and components which have zero value RRs will be disabled. The optimal components order minimize the number of processed alerts in each stage by starting from higher to lower reduction rate components. This model dynamically selects optimum correlation components arrangement order and provides minimum correlation for different datasets. DPCM correlation result of collected dataset will be discussed in Sect. 1.5. The total correlation time by DPCM is calculated as follows:

$$\mathrm{TDPCM} = \sum_{i=1}^{n} T_i X_i$$

where X_i represents the active correlation stages, these stages contain only effective correlation components and have dynamic descending order of its reduction rates. Total T_{DPM} correlation time is optimum compared with total correlation time by CAM.

1.3.3 DACM Individual Agents

In this section DACM correlation agents design and algorithm will be presented, block diagram of DACM components shown in Fig. 1.3.

DACM [1] is composed from a set of correlation agents; each agent correlates alerts or events from its information source. Different agent's sources of information including IDS alerts, firewall log file, other services log files. Log files may be error log files or access log files, and finally a set of knowledge base which include network security policy and needed threshold values to determine behavioral profiles and attacks signatures.

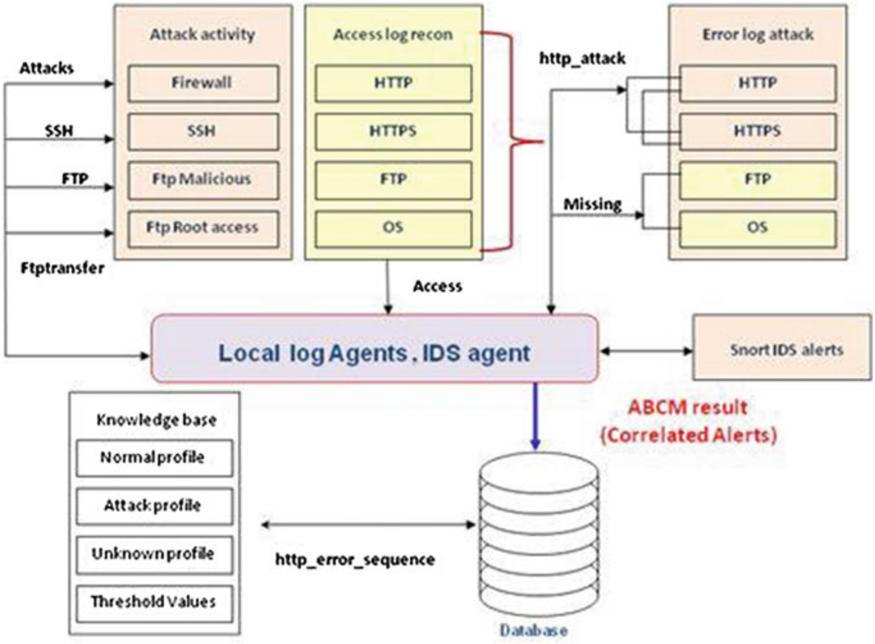


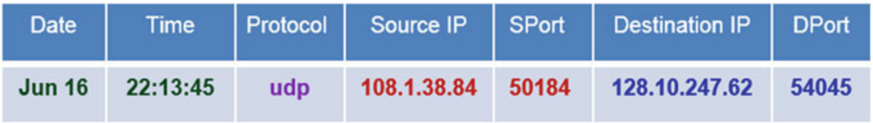
Fig. 1.3 DACM individual agents

1.3.3.1 IP Address Normalization

Different source of information have been used for retrieving attack signatures, the detected IP address has different format in each source. IDS alerts include decimal format for source and destination IPs, while INFOSEC tools and other application log files include standard 32 bit representation “Standard IPs”. Normalizing IPs together is a necessary process for correlation such information together, this process indicates that every IP address has a unique ID in the system.

1.3.3.2 Firewall Agent

Firewall agent read the contents of router log file, this log file contains list of blocked IP which tried to attack the network, and it extracts the data indicating the attack such as date and time of attack trial, the destination protocol, source IP and source port, and destination IP and port.



Date	Time	Protocol	Source IP	SPort	Destination IP	DPort
Jun 16	22:13:45	udp	108.1.38.84	50184	128.10.247.62	54045

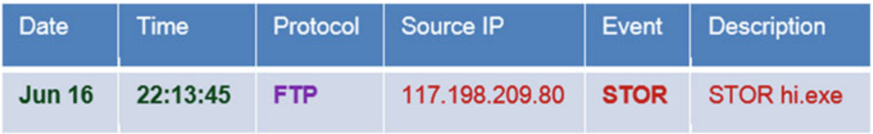
Fig. 1.4 Firewall attack entry

Algorithm 1 shows the agent process to read the log contents and convert it to a record in the attack table. The input is the router log file, and the output is the attack table record shown in Fig. 1.4.

```
Algorithm1 Firewall Agent
Input: router log file., Output: fill data to database table 'attacks'
Initialization:  set server name = "cisco3.cerias.purdue.edu"
                 Set SP1 = "%SEC-6-IPACCESSLOGP: " ,  Set SP2 = " -> ""
Begin
    While not EOF
        For each line
          Read line contents; Split line with static variables name;
        read date and time;  read source IP and port;  read protocol type; read destination IP and port;
        ignore unwanted variables;
        Store to attack table (date, time, source IP, source Port, protocol, destination IP , destination Port );.
        End for;
    End While;
Return 'attack' table.
```

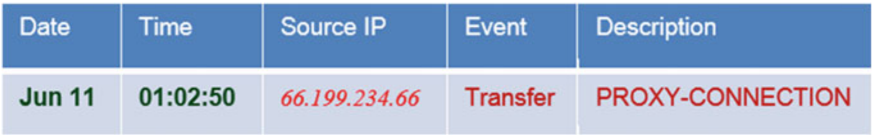
1.3.3.3 FTP Local Agents

The first FTP agent algorithm reads the contents of the log file which contains ftp service logs and error messages and requests associated with FTP commands. Local FTP agent reads complete session for the user activity in the log file to check the command type tried by the user. Algorithm 2 shows the FTP agent process to read the log contents and check if it contains any FTP command which violates the network security policy. In case of finding an evidence of that violation, it extracts this log entry and inserts it to a record in FTP attack table shown in Fig. 1.5.



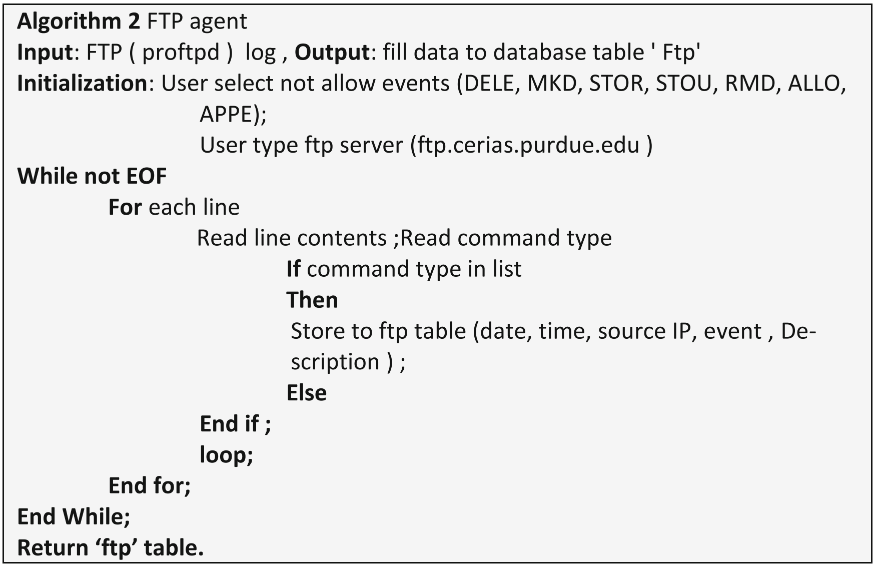
Date	Time	Protocol	Source IP	Event	Description
Jun 16	22:13:45	FTP	117.198.209.80	STOR	STOR hi.exe

Fig. 1.5 FTP attack entry



Date	Time	Source IP	Event	Description
Jun 11	01:02:50	*66.199.234.66*	Transfer	PROXY-CONNECTION

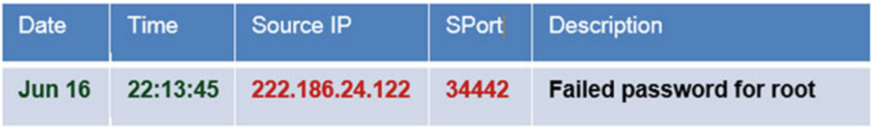
Fig. 1.6 FTP transfer attack entry



```
Algorithm 2 FTP agent
Input: FTP ( proftpd ) log , Output: fill data to database table ' Ftp'
Initialization: User select not allow events (DELE, MKD, STOR, STOU, RMD, ALLO,
                APPE);
                User type ftp server (ftp.cerias.purdue.edu )
While not EOF
      For each line
            Read line contents ;Read command type
                  If command type in list
                  Then
                   Store to ftp table (date, time, source IP, event , De-
                   scription ) ;
                  Else
            End if ;
            loop;
      End for;
End While;
Return 'ftp' table.
```

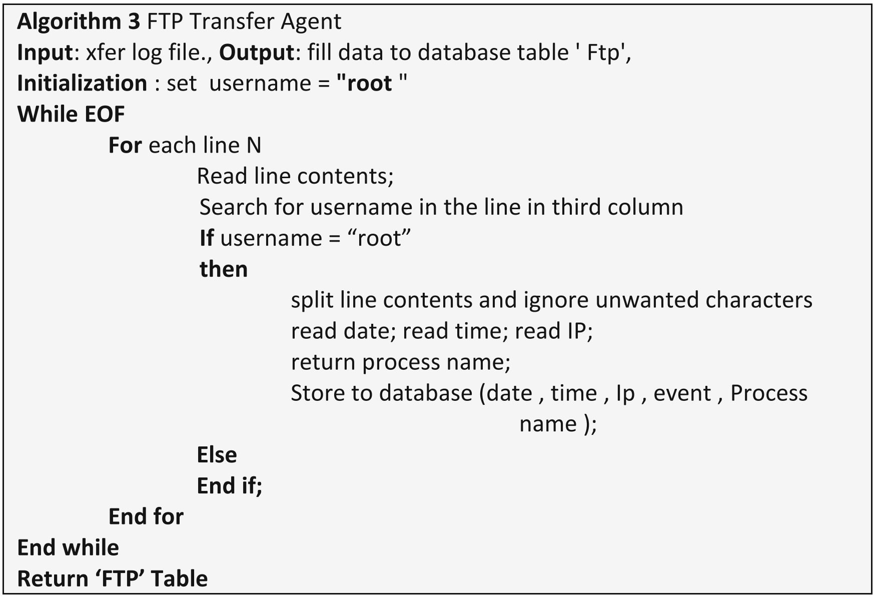
FTP transfer agent algorithm detects malicious behavior during transfer FTP files. The log file contains a listing of files transferred over FTP, normally the third column should say ftp for all users which indicates FTP user trying to transfer file, while if we have a “root” username and the commands do not appear to be standard, it indicates a malicious trial to FTP transfer using root access.

Algorithm 3 shows FTP transfer algorithm, it reads the log file as an input while it stores its output in FTP transfer attack table as shown in Fig. 1.6.



Date	Time	Source IP	SPort	Description
Jun 16	22:13:45	222.186.24.122	34442	Failed password for root

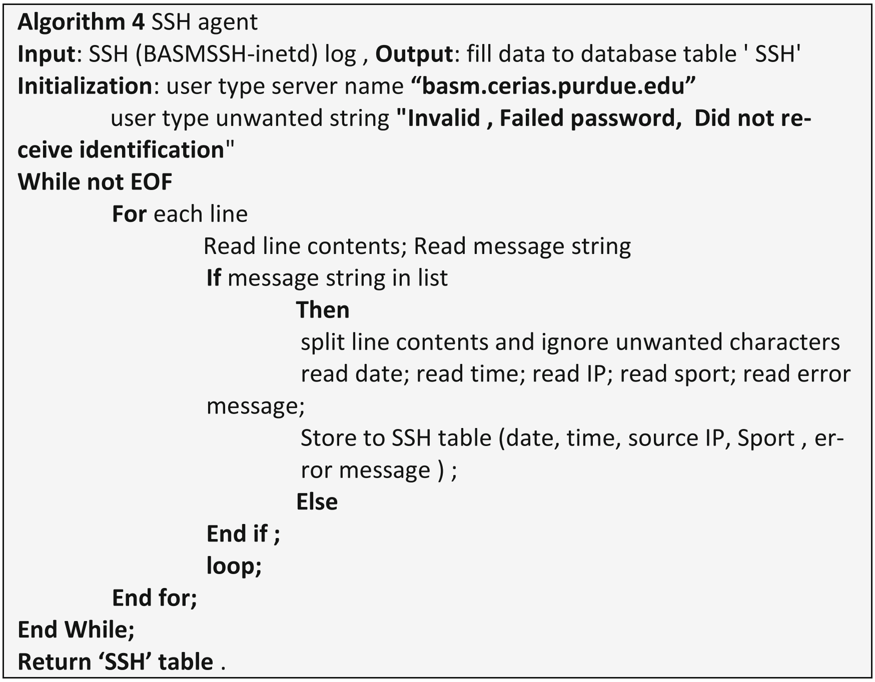
Fig. 1.7 SSH transfer attack entry



```
Algorithm 3 FTP Transfer Agent
Input: xfer log file., Output: fill data to database table ' Ftp',
Initialization : set  username = "root "
While EOF
        For each line N
                Read line contents;
                Search for username in the line in third column
                If username = “root”
                then
                        split line contents and ignore unwanted characters
                        read date; read time; read IP;
                        return process name;
                        Store to database (date , time , Ip , event , Process
                                                    name );
                Else
                End if;
        End for
End while
Return ‘FTP’ Table
```

1.3.3.4 SSH Agent

Secure Shell (SSH) agent is an example of service agents. SSH agent reads the contents of SSH service log file that records service usage and error messages from the service and child processes. It shows the attacker’s attempts to access SSH service with root user or invalid username/password. Algorithm 4 shows the SSH agent process to read the log contents and check if it contains any error messages which violates the network security policy and extracts this log entry and inserts it to a record in SSH attack table as shown in Fig. 1.7.

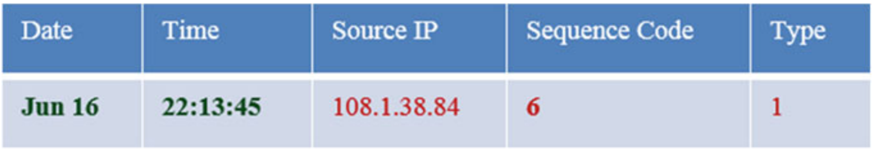


```
Algorithm 4 SSH agent
Input: SSH (BASMSSH-inetd) log , Output: fill data to database table ' SSH'
Initialization: user type server name "basm.cerias.purdue.edu"
          user type unwanted string "Invalid , Failed password,  Did not receive identification"
While not EOF
          For each line
                    Read line contents; Read message string
                    If message string in list
                              Then
                              split line contents and ignore unwanted characters
                              read date; read time; read IP; read sport; read error
                    message;
                              Store to SSH table (date, time, source IP, Sport , error message ) ;
                              Else
                    End if ;
                    loop;
          End for;
End While;
Return 'SSH' table .
```

1.3.3.5 Error Log Agent

Error log agent reads the contents of the file associated with http and https services. The purpose of the error log agent is to identify attack signatures stored in the http or https error log files by reading these files and comparing their contents with previously created attack or normal profiles. In case of detecting a new profile in the log files, error log agent checks the similarity of this profile to one of known profiles and identify the new profile as similar to attack or similar to normal. Later the user administrator can assure this similarity and change the type of profile to attack profile or normal profile.

Algorithm 5 shows the error agent process to read the log contents and check if it contains any attack signatures which violates the network security policy and extracts this log entry and inserts it to a record in http attack table. The input is the "errorlog" log file which contains user's messages in http and https server, and the output is the http attack table, http attack table attributes are shown in Fig. 1.8. Figure 1.8 shows the result of malicious behavior by IP 108.1.38.84. Sequence "6" is the error sequence code stored in http error sequence profile and type "1" represent that the type of this profile is an attack profile.



Date	Time	Source IP	Sequence Code	Type
Jun 16	22:13:45	108.1.38.84	6	1

Fig. 1.8 HTTP attack entry

```
Algorithm 5 Http Agent
Input: http or https error log file, error sequence table. Output: fill data to database table ' http attack'
Initialization: user select type of file type :   1 – http , 2 – https
While not EOF
        For each line
        Read line contents;  Get first error line; Repeat until  error sequence end;
        Read date; Read time; Read ip address; Read error sequence;
        Check the error sequence in  http_error_sequence table ;
        If match
        then
                get  error sequence code ; get sequence type;
        else
                check similarity;
        end if;
End for;
End While ;
Return  ' http_attack ' table
```

The following is the algorithm for the check similarity function in algorithm 5

```
Check similarity
If error sequence subset of other known error sequence
Then
        Error type is similar to type; Insert to http_error_sequence table ;
Else
        Error type is unknown; Insert to http_error_sequence table ;
End if;
example:        A - " {main}(),include() "
                B-"{main}(),include(),PageDef->show(),CPL:: checkInternalIP() "
                If error B stored as normal behaviour the error A  is part of B
                then A will have type “similar to normal “
```

1.3.3.6 Access Log Agent

On the contrary of other individual agents, access log agent does not indicate attack signatures or malicious behavior by external users, it indicates users who are trying to gather information and check the website contents or the operating environment of the network. Access log agent reads the contents of "access log" files about access messages associated with http and https services.

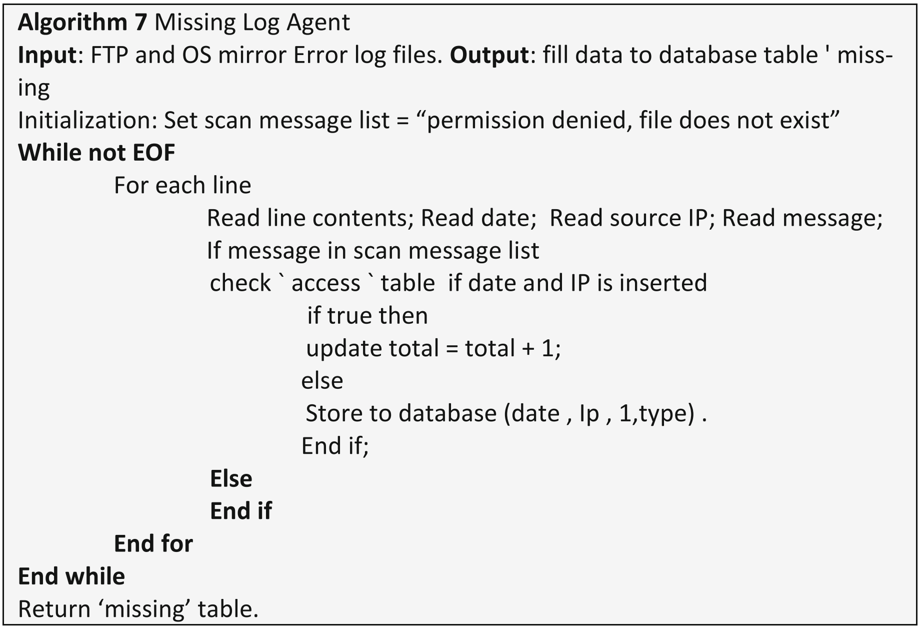
The purpose of the access log agent is to identify reconnaissance activities against the network which allows early detection of expected attacks.

Algorithm 6 shows the access agent process to read the log contents and aggregates different user access within the website link. The input is the "access log" files which contain user's access in http, https, OS, and FTP services, and the output is the access table.

```
Algorithm 6 Access  Agent
Input: Access log files;  Output: fill data to database table 'Access'
Initialization: User select file type : 1 – http , 2 – https , 3 – os mirror , 4 – ftp
While Not EOF
        For each line
                Read line contents; Read date; Read source IP;
                check ` access ` table  if date and IP is inserted
                        if true then update total = total + 1;
                        else
                        Store to database (date , Ip , 1,type) .
                        End if;
        End for
End while
Return ' Access'  table .
```

While access log agent indicates the external user access to the contents of the web site, Missing log agent indicates the external user scan of the network files which represent their trial to identify the server Operating system or looking for the availability of specific services. Missing log agent reads the contents of the OS error log and FTP error log files. The purpose of missing log agent is to identify scan activities against the network which allows early detection of expected attacks.

Algorithm 7 shows the missing agent process to read the log contents and aggregates different user scans within the website files. The input is the "error log" files which contain user's access in http, https, OS, and FTP services, and the output is the missing table. Initial scan messages list include the messages indicating that the users are trying to access unauthorized files or looking to specific files names which show the used operating system.



Algorithm 7 Missing Log Agent

Input: FTP and OS mirror Error log files. **Output**: fill data to database table ' missing

```
Initialization: Set scan message list = “permission denied, file does not exist”
While not EOF
        For each line
                Read line contents; Read date;  Read source IP; Read message;
                If message in scan message list
                check ` access ` table  if date and IP is inserted
                        if true then
                        update total = total + 1;
                        else
                        Store to database (date , Ip , 1,type) .
                        End if;
                Else
                End if
        End for
End while
Return ‘missing’ table.
```

1.3.3.7 DACM Central Agent

DACM central agent has access to the result tables of different individual agents; it aggregates these results together into unified table which includes those results together in relation with the attacker IP. Figure 1.9 show that different individual agents stored their results in central database tables. Central agent gets that database to produce a set of useful reports which summarize different attacks against the network together. Those results shown in the figure includes daily report, IP report, severity alerts, single alerts, false negative alerts, reconnaissance alerts, summary date report, and sever IPs report. By the end of the individual agent results, table IPs contains all different IPs which has been stored in attack tables or access or missing tables with a unique ID representing that IP address.

DACM central agent performs its function through two steps, the first step is to aggregate all different results from individual agent results into daily table which includes different attacks and activities for different IPs. The second step is the analysis of these attacks together to represent the whole picture of the situation in the network and improve the detection rate and to produce the correlation results of such different agents together.

Algorithm 8 show the first step process, the agent loops through IPS table, for each record, it selects the ID for the IP address and select related attacks and activity for that IP from different result tables and insert that record in new table called daily.

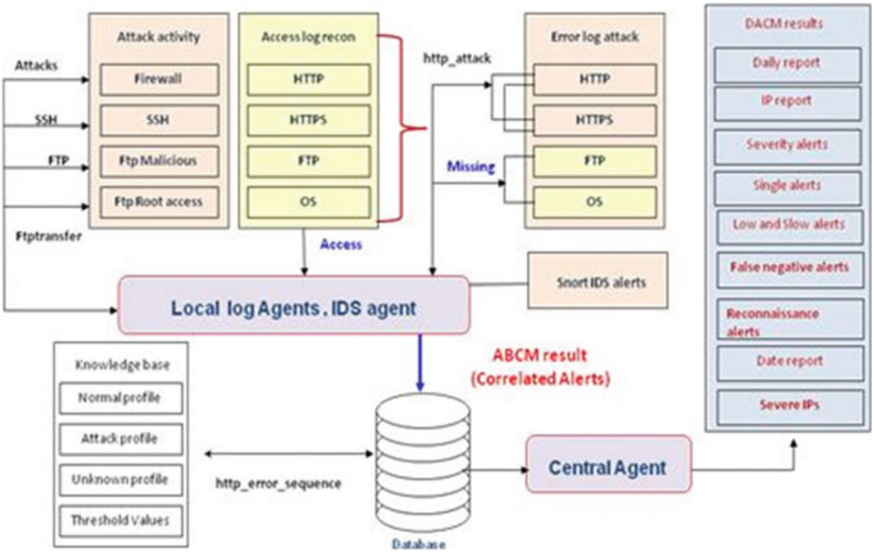
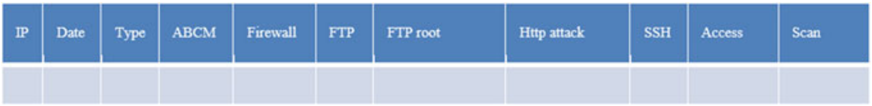


Fig. 1.9 DACM central agent results

```
Algorithm 8 Daily Agent
Input: All other agents' results tables.
Output: fill data to database table ' daily' and ' daily_res'
Step 1 : Loop through IPS table
                    Insert new record in  ' daily' table with
Select   IP from IPS, // where IP = IP in related tables
  Date as current system data, // where date = date in related tables
  Count `abcm_res` , // IDS correlated alerts , Count  `attacks_res`, // firewall contents for IP ,
  Count  `ftp`, // FTP attack table ,  Count `ftptransfer`, // FTP transfer attack table
,
  Count `http_attack`,// http attack table, Count  `ssh`, // ssh attack table,
  Sum `access`, // acces count for that IP  Where count > access threshold value  ,
sum `missing`, // scan count for that IP Where count > scan threshold value
End loop
Step 2 : For IP in daily table select IP, Date, Count `abcm_res` , Sum `access`,
          Count `attacks_res`, Count `ftp`,Count `ftptransfer`, Count `http_attack`,
Sum `missing`, Count `ssh`, Alert type
```

Figure 1.10 shows the daily table fields. These fields are Date, IP, count of alerts for this IP from IDS correlated alerts in that day, count of IP in firewall blocked IPs log, related FTP and FTP transfer attacks, related http attacks, related SSH attacks, and related access or scan activities count for that IP which exceed the allowed threshold values for normal access or scan activities. In addition an analysis result field called type to show the IP behavior and conclusion in the system.



IP	Date	Type	ABCM	Firewall	FTP	FTP root	Http attack	SSH	Access	Scan

Fig. 1.10 Daily report attributes

Daily report algorithm aggregates that related attacks together for better understanding of current situation for different attacks, and summarizes wide range of activity for each attacker IP.

Algorithm 8 performs the second step to show the conclusion and better analysis of that aggregated alerts in Step 1. Step 2 determines the behavior type for each IP. Step 2 loops through daily table to determine the alert type by comparing the count of alerts from different sources to indicate the alert type according to a set of rules shown in the following pseudo code for each set of types. The alert type for each IP will be determined according to different alert and attacks from that IP.

The alert type could be false negative in case that the IP attacks are not detected in the IDS alerts while they were detected from other sources of attacks such as FTP, SSH, and HTTP attacks. It could be considered false negative alerts where the IP attack is detected in IDS alert and detected in other sources but in different times. False negative and verified alerts conditions are shown in false negative and verified alerts pseudo code; if the alerts were detected in IDS correlated alerts and detected in the same window time from other sources of attacks such as FTP, SSH, and HTTP attacks, then the alerts are considered verified alerts or severe alerts. Verified alerts assure that the alerts were detected using different sources of information which reduce false positive alerts.

```
False negative and verified alerts pseudo code
If Count `abcm_res` = 0 and {   Count `attacks_res`      >= 1
   OR    Count `ftp` >= 1  OR count `ftptransfer`    >= 1 OR count `http_attack`
         >= 1
   OR    Count `ssh` >= 1 }
Then  Alert type = False Negative
Else  For each alerts
   If time (alert) <> time (alerts from other attacks)
  Then  Alert type = False Negative
Else Alert type= Verified alerts
  End if
End if;
```

Alert type is a single IDS alert if only one alert was detected from IDS and have not been detected by other sources of information in that window time. Such kind of an alert could be stored to indicate the probability of low and slow attack.

```
Case : single alert // single alerts for Low and Slow attacks
                                  If Count `abcm_res` = 1   and
                                             Count  `attacks_res`= 0  and  Count  `ftp` = 0
                                   and   count `ftptransfer`   =0   and  count `http_attack`
                                             = 0   and  Count  `ssh`= 0  ;
                                  Then  Alert type = Single alert;
                                   End if;
```

Alert type is Reconnaissance alertif there is no alerts detected either from IDS or from other sources of information in that time, but there is system access and/or scan which exceed the threshold allowed values. Such kinds of reconnaissance indicate it is gathering data about the system before being attacked. Such kinds of alert type enable the early detection of those trials of attacks.

```
Case : Reconnaissance // no attacks, gathering data or systemscan
If Count `abcm_res` = 0   and  {
   Count  `attacks_res`       = 0   and Count  `ftp` = 0   and count `ftptransfer`= 0
and count `http_attack` = 0    and   Count  `ssh` = 0   } and  {sum `missing`         >=
minimum value   OR Sum `access`>= minimum value }
then Alert type = Reconnaissance alert.
Else
End if
```

Alert type is IDS only when many alerts were only detected by IDS and have not been detected by other sources of information in that time. Such kinds of alerts indicate that no other sources were able to detect that attack which may help in improving the logging capability.

```
Case: IDS alerts only
If Count of `abcm_res` > 1   and
{Count  `attacks_res`        = 0   and  Count  `ftp`        =    0       and
count `ftptransfer`= 0   and  count `http_attack` = 0   and Count  `ssh`  = 0   } and
                       { OR      sum `missing`      <= minimum value   OR     Sum `ac-
                       cess`      <= minimum value    }
Then Alert type = IDS only alert;
 Else
End if ;
Return ' daily' and ' daily_res' tables
```

1.3.4 Implementation Environment

DACM have been implemented using Mysql database for storing alerts tables and correlated alerts results as well as result tables of different agents, it also used to store knowledge base and learning criteria. Borland6 C++ programming language have been used to implement IDS alert correlation agent, while Microsoft Visual Basic 6 used to implement central agent and other correlation agents for security tools and log files. Finally, we used Windows7 Ultimate 64-bit operating system over Dell studio laptop with Intel Core2Duo CPU-9300-2.5GHz – 6MB cache processor, and 2GB RAM for testing the implemented model.

1.4 DACM Results and Analysis

In this section, detailed DACM results will be presented in its individual agents and central agent, it also includes description of the CERIAS dataset which have been gathered and tested to implement DACM.

1.4.1 CRIAS Data Set

The dataset needed for DACM implementation was collected during scholar visit through the period of April–July 2010. The visit location was the Center for Education and Research in Information Assurance and Security (CERIAS) [35]. Data was collected through the period of June 11–28 2010. It consists of three main sources: Snort [36] alerts, network packet data, and application and system log messages. The snort sensors monitored network traffic on the 128.10.254.0/24 and the 128.10.252.0/24 subnets.

Snort captured almost 800,000 alerts in this period. The alerts were stored in database tables within a MySQL database and are accessed through the ACIDBASE [37] web interface. The environment was Ubuntu 10.04 Linux-based OS.

We used Wireshark [31] to capture network packets traffic for CERIAS website (kargad) which contains detailed packet information. We also retrieved the output of Nessus [38], a network vulnerability scanner, and Nmap [39], a network mapping utility, to check for known network vulnerabilities and network port status.

Finally, we collected log files for CERIAS services as follows: FTP log messages, error messages and requests associated with FTP, a listing of files transferred over FTP, HTTP log messages, HTTP requests, errors associated with HTTP requests, HTTPS (SSL) requests, HTTPS (SSL) errors associated with HTTPS, SSH logs which includes error messages associated with SSH, and Firewall Router log messages which contain a list of blocked network packets from outside world. While our model includes performance data, they were not collected during the data collection process (experiment) as necessary performance monitoring tools were not available.

Table 1.2 SNORT IDS alert attributes

Sid	Cid	Sig_Name	Timestamp	IP_src	IP_dst	Proto	Sport	Dport
6	16	ICMP PING speedera	6/11/2010 9:14	3460811837	2148204039	1		
6	20	EXPLOIT ntpdx overflow	6/11/2010 9:21	1656885345	2148204039	17	123	123
7	28	WEB-MISC /doc/ access	6/11/2010 21:42	1131319090	2148203530	6	59285	80
7	30	WEB-MISC robots.txt access	6/11/2010 21:47	3475949512	2148203529	6	19427	80

Some attacks were tried during the capture data period added to the normal attack behavior, our simulated attacks was for the purpose to assure the captured data contains low and slow attacks and is as follows:

Nmap: port scan of web server, we used the version check option to determine the name and version of the service “nmap –sV”, we looked to port 80 and 443 (http and https) service.

Nikto: Configuration scan of the web server, we attempted to evade IDS detection by slowing the scan speed down, “nikto.pl -Tuning 3b -Pause 5 –evasion”.

Using the Firefox plug-in called tamper data; we attempted to send bad data to a form in CERIAS web site in order to exploit vulnerabilities in the form processing script.

1.4.2 IDS Alerts Correlation Results

Snort collected 858000 alerts in CERIAS dataset within 18 days; alerts were divided to be correlated through those days. We correlated the alerts using CAM [25–27], ABCM [32, 33], and DPCM [34]. Table 1.2 show samples of these alerts; alerts attributes includes sensor id, alert id, signature, timestamp, source IP, destination IP, protocol, source port, and destination port.

Integrated interface to correlate the alerts using the three different techniques was implemented to compare their reduction rate and their correlation time. We run the three models for 18 days alerts and get the reduction rate and correlation time for each model.

1.4.2.1 IDS Alert Correlation Techniques Performance

This section presents processing results comparison of gathered alerts in CERIAS dataset using the three different IDS correlation techniques. Table 1.3 summarizes reduction rate comparison for CAM, DPCM, and ABCM.

Table 1.3 Alert correlation reduction rates comparison

Day	I/P alerts	CAM		DPCM		Alert learned	ABCM	
		O/P	RR	O/P	RR		O/P	RR
11	28664	725	97.47	625	97.82	2866	619	97.60
12	46703	1076	97.70	979	97.90	4670	961	97.71
13	54759	1060	98.06	935	98.29	5475	909	98.16
14	34303	1184	96.55	1178	96.57	3430	1050	96.60
15	51823	944	98.18	870	98.32	5182	832	98.22
16	49609	1095	97.79	1010	97.96	4960	997	97.77
17	34879	982	97.18	905	97.41	3487	881	97.19
18	152240	1083	99.29	1077	99.29	15224	1044	99.24
19	15175	531	96.50	513	96.62	1517	497	96.36
20	11788	354	97.00	346	97.06	1178	312	97.06
21	49336	1164	97.64	1143	97.68	4933	1118	97.48
22	39786	1146	97.12	1139	97.14	3978	1035	97.11
23	27753	1214	95.63	1171	95.78	2775	1067	95.73
24	70686	1192	98.31	1174	98.34	7086	1140	98.21
25	36072	1117	96.90	1106	96.93	3607	1003	96.91
26	57598	1055	98.17	1036	98.20	5759	1018	98.04
27	52035	1083	97.92	1075	97.93	5203	1039	97.78
28	25139	736	97.07	722	97.13	2513	696	96.92

Figure 1.11 shows a graph chart of daily reduction rate for CAM, DPCM, and ABCM. The results showed that the reduction rates of the three models were very close to each other. Table 1.4 summarizes the correlation time comparison for CAM, DPCM, and ABCM. It shows the daily alerts ordered by date of alerts, alerts count and correlation time for each techniques.

Figure 1.12 shows a graph chart of daily correlation time for CAM, DPCM, and ABCM. The results shows that the correlation time increase linearly with increasing alerts count till the range of 70,000 alerts, while it increased exponentially in case of 152240 alerts. Detailed correlation time for alerts count less than 70000 alerts (dashed red rectangular in Fig. 1.12) will be showed in individual graph chart to represent it.

Figure 1.13 shows a graph chart for correlation time for different IDS correlation techniques with excluding of maximum number of alerts of 152240 alerts to show the detailed difference in correlation time for each technique for alerts less than 70000 alerts.

Finally ABCM has lowest correlation time because it uses only effective components in ACCL, while CAM has higher correlation time because it consumes time in ineffective components; DPCM has higher correlation time because the use of stages increased every stage time compared with use of components in the single processor used environment.

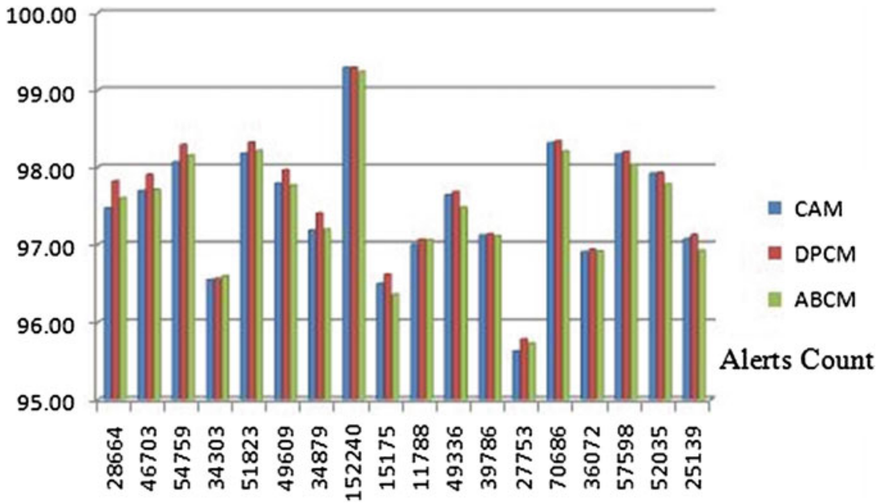


Fig. 1.11 Reduction rates comparison of IDS correlation techniques

Table 1.4 Comparison of correlation time for IDS alert correlation models

Day	Alerts count	Correlation time			Day	Alerts count	Correlation time		
		CAM	ABCM	DPCM			CAM	ABCM	DPCM
11	28664	195	10	367	20	11788	27	3	50
12	46703	436	28	842	19	15175	52	9	102
13	54759	620	35	1159	28	25139	141	15	272
14	34303	233	23	434	23	27753	171	18	281
15	51823	447	35	982	11	28664	195	10	367
16	49609	504	39	1130	14	34303	233	23	434
17	34879	325	24	526	17	34879	325	24	526
18	152240	6048	103	10082	25	36072	278	20	543
19	15175	52	9	102	22	39786	375	34	738
20	11788	27	3	50	12	46703	436	28	842
21	49336	457	32	889	21	49336	457	32	889
22	39786	375	34	738	16	49609	504	39	1130
23	27753	171	18	281	15	51823	447	35	982
24	70686	940	44	1863	27	52035	564	27	1120
25	36072	278	20	543	13	54759	620	35	1159
26	57598	652	39	1260	26	57598	652	39	1260
27	52035	564	27	1120	24	70686	940	44	1863
28	25139	141	15	272	18	152240	6048	103	10082

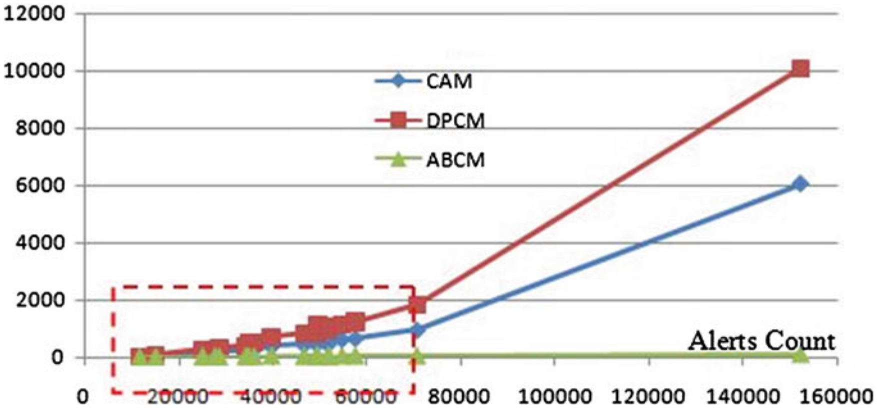


Fig. 1.12 Comparison of IDS correlation techniques

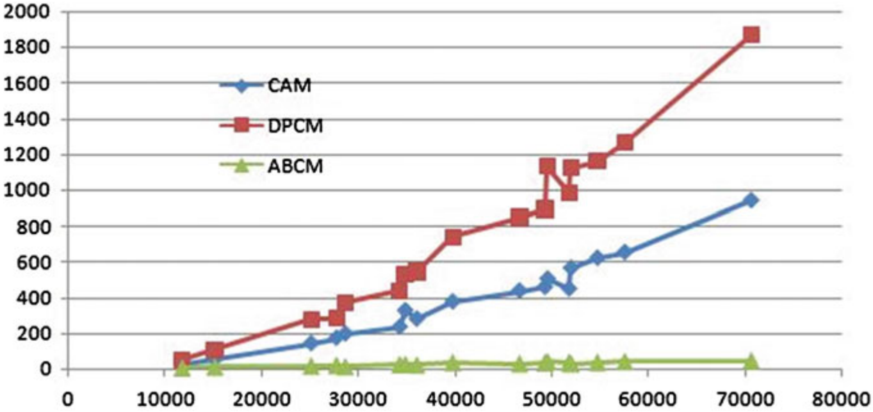


Fig. 1.13 Correlation times comparison of IDS correlation techniques-2

1.4.3 DACM Components Results

Different individual agent's results for IDS agent, INFOSEC tools, and system log files could be obtained; detailed attack result tables include such attacks. The agent's results include SSH agent result which includes detected malicious activity within SSH service. SSH agent detects users who are trying to guess username or password for SSH services or trying to hide their browser information to prevent system from identifying them. Other similar agent's results for different services such as Firewall, FTP, FTP transfer, error log attack, access log, and system scan could be presented in such way. ABCM correlation for IDS alerts is included as IDS agent result to integrate with other agents results. ABCM correlated alerts include source IP, date, time, alert signature, and destination. Each agent result includes ID, Date, and Time, Source

Table 1.5 Attack type summary of central agent

False negative	Severity alerts	IDS only	Single alerts	Firewall	Reconnaissance
4819	337	1375	4578	74378	1273

IP, and attack description. Three common attributes of all alerts from different agent result date, time, and attacker or source IP, these attributes could be used to integrate attacks done by same IP in same time. Integrating such kind of alerts together with IDS correlated alerts conclude the current situation of attempted intrusion to the system.

1.4.4 DACM Central Agent Results

DACM central agent has rich valuable information from different individual agent's result. Using this information together, the central agent provides set of reports summarizing the improvement in IDS capability.

Daily report of DACM concludes alerts and shows their total classification. The alert type is driven from the integration of different agent's result for the same IP. Table 1.5 shows summary of different type of discovered alerts during the whole test period. False Negative alerts indicate the alerts were not detected by IDS through its correlated alerts and discovered by other agent's results. Severity alerts indicate the IDS correlated alerts which also discovered by other agents result. IDS only alerts are such alerts were detected more than one time for the same IP by only IDS through its correlated alerts and not discovered by any of the other agent's results. Single alerts are such alerts were detected by only IDS through its correlated alerts just one time during the whole period of test and never repeated and not discovered by any of the other agent's result. Single alerts are stored for a while to be analyzed for detection of low and slow attack. Firewall alerts represent summarized information in daily result report. Reconnaissance alerts conclude the trial of gathering data about the network.

DACM agent has different reports to trace a specific IP address to conclude the trial and attack signature for this IP in different source of information. IP "108.1.38.84" was detected by ABCM correlation as IDS alerts and the same IP appeared in the firewall agent result as blocked IP, also it was detected by http attack in the same time window in one case. In other cases it was detected in http attack while not detected in IDS because of the use of IDS evasion technique. In other cases it was detected only in IDS and was not detected by http attack because of the attack nature. DACM determines the most repeated IPs as sources for different attacks. DACM produces maximum priority report, which displays the most common IPs which was detected by different agent's results or by DACM central agent. The report shows the top n IPs which has the highest count of repeated alerts for different alerts type and/or reconnaissance activity.

Table 1.6 DACM summary result

Day	Alerts count	Detection of correlated alerts			Detection Percentage		
		IDS	Other Logs	DACM	IDS	Other logs	DACM
11	28664	277	339	616	45	55	100
12	46703	393	241	634	62	38	100
13	54759	380	290	670	57	43	100
14	34303	449	304	753	60	40	100
15	51823	449	300	749	60	40	100
16	49609	458	322	780	59	41	100
17	34879	373	357	730	51	49	100
18	152240	469	342	811	58	42	100
19	15175	248	248	496	50	50	100
20	11788	166	288	454	37	63	100
21	49336	444	306	750	59	41	100
22	39786	414	354	768	54	46	100
23	27753	427	370	797	54	46	100
24	70686	451	317	768	59	41	100
25	36072	411	322	733	56	44	100
26	57598	421	266	687	61	39	100
27	52035	389	224	613	63	37	100
28	25139	334	248	582	57	43	100

DACM summarizes the total daily alerts and classify them depending on their source of detection. Table 1.6 summarizes daily alerts count and number of IDS detection and missed alerts from IDS which have been detected by other log agents, and the total of them.

The average percentage alerts for IDS alerts percentage was 56 % and was 44 % percentages for missed alerts in case of average daily alerts of 46574 alerts. Figure 1.14 shows a graph chart of number of IDS detection as IDS and missed alerts from IDS which have been detected by other log agents as other log, and the total for both of them as DACM. The graph shows that the use of other agents in DACM enhances the detection rate of missed alerts.

Figure 1.15 shows a graph chart of percentages of IDS detection and missed alerts from IDS which have been detected by other log agents and the total for both of them. Total detected alerts by DACM represent the complete unit for both kinds of detection. The graph shows that DACM enhances the detection rate of missed alerts by 44 % compared with the case of using just IDS correlation.

1.4.5 DACM Evaluation and Assessment

DACM improves IDS capability through the use of different sources of information. False positive alerts are reduced because of the verification of alerts detected by IDS

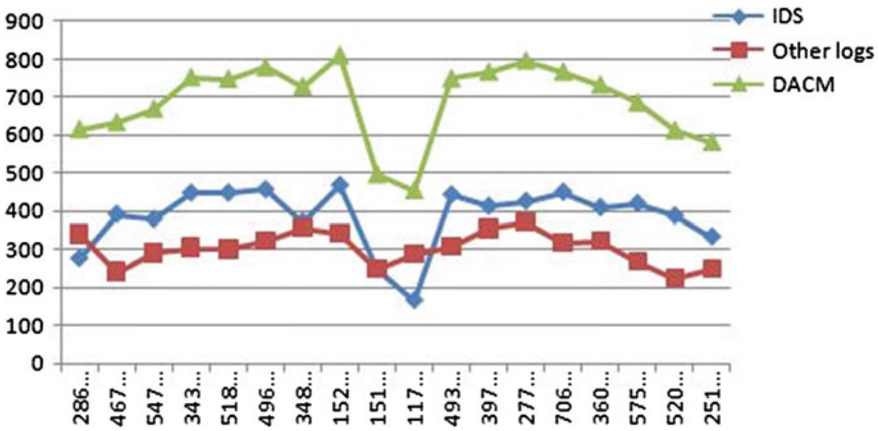


Fig. 1.14 DACM summary results chart

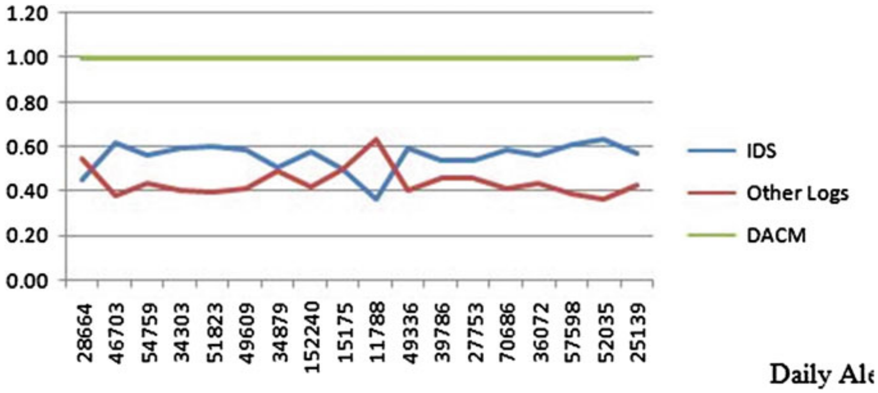


Fig. 1.15 DACM percentage summary results chart

from other sources like firewalls, different log attacks, or http attacks. False negative alerts are reduced because missed alerts from IDSs are detected from other logs such as SSH, FTP, and http attacks. DACM enables early detection of trials to gather data which represent the first phase of advanced persistent threat and individual alerts for Low and Slow attack.

Previous correlation techniques were limited to the use of IDS alerts for correlation and enhancing correlation component performance. Few techniques [40, 41] used vulnerability scanners to assure alert verification.

Using the simplicity of the relationship between individual agents, it is an easy and simple task for each individual agent to correlate its alerts and shares its output with other agents. This approach reduces the overhead and enables the ideal use of system resources such as memory and CPU. DACM enables minimum correlation time of ABCM as IDS alerts correlation technique and allows continuous adaptive learning to update ACCL, assuring the use of suitable correlation components for different datasets.

The DACM central agent accesses the results tables of other agents from central database, reducing network traffic compared with the case of accessing them from multiple machines or accessing the information source itself. DACM has the capability for real time operation with minor modification in agent programs; the current proposed prototype was implemented after collecting the dataset, so it was not possible to run it as a real time model.

Additional DACM learning is needed to build more accurate behavioral profiles which determine attack signatures in system and application log files. Multi-step attack scenarios also need more learning to build pre-condition and post-condition tables for such attacks. DACM alerts are considered on an equal footing, and aren't considered the influencing factors of different alerts on the same information system. Research is needed to distinguish between such alerts and assign weights for each alert type depending on its source of information and influence of the provided service.

1.5 Conclusions and Future Work

This article presented a Distributed Agent Correlation Model (DACM) providing a scalable alert correlation for large scale networks. The model utilizes multiple distributed agents to provide an integrated correlation solution. The model can be extended by creating new correlation agents, and can be tailored to a protected network by selecting what agents to use and configuring each individual agent's parameters. DACM correlates alerts from IDSs with other information source such as INFOSEC tools and system and application log files.

This article compared two alternative models to enhance the IDS alert correlation model in CAM: an Agent Based Correlation Model (ABCM) and a Dynamic Parallel Correlation Model (DPCM). The results showed that ABCM and DPCM have similar RRs as CAM, while ABCM has the lowest correlation time and DPCM has the highest correlation time.

Firewall log file was used as INFOSEC tools information example, firewall agent reads router log files and summarizes the blocked IP in firewall tables. SSH, FTP, access log, OS log, and error logs were used as system and application log files information example, these logs agents read the related log files, and extract the attack signature or reconnaissance trials in each file to its output tables according to previous learning capability.

The DACM central agent correlates the output of ABCM as IDS alert correlation with other agent's output. The results show that DACM enhances both the accuracy and the completeness of intrusion detections by reducing false positive and false negative alerts through the integration of these alerts from multiple information sources. DACM supports an adaptive continuous learning capability by providing profiles which have never been learned as normal or attack behavior to the system administrator to classify these profiles.

The results show that DACM provides 44 % better intrusion detection than other IDS techniques through the detection of new attacks which were not detected by IDSs. It also showed that DACM detected low and slow attacks and reconnaissance trials by external users. These reconnaissance trials are a signature of early detection of Advanced Persistent Threats. DACM can be used to detect Zero Day Attacks through detection of any malicious behavior compared with normal network behavior. Finally, DACM could be used as a real time system with minor modifications to the current implementation to allow continuous online correlation for individual and central agents.

Future research including extending the model by implementing other agents for network security tools, system audit logs, and host based IDS; enhancing the learning capability with more accurate behavioral profiles for detecting coordinated attacks and multi-step attacks. Studying distributed wide area networks and worldwide correlation would improve the intrusion detection and early detection of new attacks. Expanding the model to include automated responses would address the need for immediate responses to attacks. Finally, measuring the performance, trustworthiness, and assurance of distributed agents is a challenge to the problem of the probability of the existence of fake agents.

Acknowledgments I would like to thank the Center for Education and Research of Information Assurance and Security (CERIAS), Purdue University, USA. I appreciate the valuable support of the CERIAS executive director Prof Eugene Spafford, the generous effort of his staff especially-Information Assurance Research Engineer Keith Watson, for their cooperation during the scholar visit to the Center. They provided great resources to capture and collect the data needed for this work.

References

1. Taha, A.E.: Intrusion detection correlation in computer network using multi-agent system. Ph.D. Thesis, University of Ain Shams, Cairo, Egypt, 2011
2. Tran, Q.A., Jiang, F., Ha, Q.M.: Evolving block-based neural network and field programmable gate arrays for host-based intrusion detection system. In: 2012 Fourth International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2012
3. Elshoush, H.T., Osman, I.M.: An improved framework for intrusion alert correlation. Proceedings of the World Congress on Engineering, Vol I, pp. 1–6, 4–6 July. London, U.K (2012)
4. Tran, Q.A., Jiang, F., Hu, J.: A real-time netflow-based intrusion detection system with improved BBNN and high-frequency field programmable gate arrays. In: 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012
5. Spathoulas, Georgios, Katsikas, Sokratis: Methods for post-processing of alerts in intrusion detection: a survey. Int. J. Inf.Secur. Sci. **2**(2), 64–80 (2013)
6. Jiang, F., Michael F., Hu, J.:A bio-inspired host-based multi-engine detection system with sequential pattern recognition. In: IEEE Ninth International Conference on Dependable, Autonomic and Secure Computing (DASC), 2011

7. Shittu, R. et al.: Visual analytic agent-based framework for intrusion alert analysis. In: IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012
8. Elshoush, H.T., Osman, I.M.: Intrusion alert correlation framework: an innovative approach. In: IAENG Transactions on Engineering Technologies, pp. 405–420. Springer, The Netherlands (2013).
9. Jiang, F., Ling, S.S.H., Agbinya, J.I.: A nature inspired anomaly detection system using multiple detection engines. In: IEEE 2011 6th International Conference on Broadband and Biomedical Communications (IB2Com), 2011
10. Bahaa-Eldin, A.M.: Time series analysis based models for network abnormal traffic detection. In: 2011 International Conference on Computer Engineering & Systems (ICCES), pp. 64–70, 29 Nov–1 Dec 2011. doi:10.1109/ICCES.2011.6141013
11. Tucker, C.J.: Performance Metrics for Network Intrusion Systems (2013)
12. Gabra, H.N., Bahaa-Eldin, A.M., Korashy H.:Classification of ids alerts with data mining techniques . In: 2012 International Conference on Internet Study (NETs2012), Bangkok, Thailand, 2012
13. Gabra, H.N., Bahaa-Eldin, A.M., Korashy HM.: Data mining based technique for IDS alerts classification. Int. J. Electron. Commer. Stud. **5**(1), 1–6 (2014) (Academy of Taiwan Information Systems Research)
14. Porras, P., Fong, M., Valdes, A.: A mission-impact-based approach to INFOSEC alarm correlation. In: Proceedings of the. International Symposium. The Recent Advances in Intrusion Detection, pp. 95–114. Zurich, Switzerland, Oct 2002
15. Long, W., Xin, Y., Yang, Y.: ‘Vulnerabilities analyzing model for alert correlation in distributed environment. In: 2009 IITA International Conference on Services Science, Management and Engineering, pp. 408–411. Nov 2009
16. Jiang,G., Member., Cybenko, G.: Temporal and spatial distributed event correlation for network security. In: Proceedings of the American Control Conference, 30 June–2 July 2004
17. Eid, M., Artail, H., Kayssi, A., Chehab, A.: A lightweight adaptive mobile agent-based intrusion detection system LAMAIDS. Int. J. Netw. Secur. **6**(2), 145–157 (2008)
18. Dastjerdi, A.V., Bakar, K.A.: A novel hybrid mobile agent based distributed intrusion detection system. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 35. ISSN 2070–3740, Nov 2008
19. Liu, J., Li, L.: A distributed intrusion detection system based on agents. In: 2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 553–557, Dec 2008
20. Crosbie, M., Spafford, G.: Active defense of computer system using autonomous agent. Technical report no 95–008, COAST group, computer science department, Purdue University, February, 1995
21. Balasubramaniyan, J.S., Spafford, E., Zamboniy, D.: An architecture for intrusion detection using autonomous agents. COAST technical report 98/05, COAST Laboratory, Purdue University, 11 June 1998
22. Ktata, F.B., El-Kadhi, N., Ghedira, K.: Distributed agent architecture for intrusion detection based on new metrics. In: Proceeding 2009 Third International Conference on Network and System, Security, pp. 321–327, Oct 2009
23. Mohamed, A.A., Basir, O.: Fusion based approach for distributed alarm correlation in computer networks. In: 2010 Second International Conference on Communication Software and Networks, pp. 318–324, Feb 2010
24. Mohamed, A.A., Basir, O.: An adaptive multi-agent approach for distributed alarm correlation and fault identification. In: Proceedings of the Ninth IASTED International Conference on Parallel and Distributed Computing and Networks, Feb 2010
25. Valeur, F., Vigna, G., Kruegel, C., Kemmerer, R.A.: Comprehensive approach to intrusion detection alert correlation. IEEE Trans. Dependable Secure Comput. **1**, 146–69 (2004)
26. Valeur, F.: Real-time intrusion detection alert correlation, Ph.D. Thesis, University of California Santa Barbara, Santa Barbara, California, USA, (2006)

27. Kruegel, C., Valeur, F., Vigna, G.: Intrusion Detection and Correlation Challenges and Solutions. Springer, New York (2005). ISBN: 0-387-23398-9
28. David W Chadwick, "Network Firewall Technologies", Technical Report, IS Institute, University of Salford, Salford, M5 4WT, England.
29. Kak, A.: Port and Vulnerability Scanning, Packet Sniffing, Intrusion Detection, and Penetration Testing, Lecture Notes on Computer and Network Security, April 15, Purdue University (2014). https://engineering.purdue.edu/kak/compsec/NewLectures/Lecture23.pdf
30. Veysset, F., Butti, L.: Honey pot technologies. First Conference, France Télécom R&D, June 2006
31. Wireshark, Network Protocol Analyzer. http://www.wireshark.org, June 2010
32. Taha, A.E., Ghaffar, I.A., Bahaa-Eldin, A.M., Mahdi, H.M.K.: Agent based correlation model for intrusion detection alerts. In: Proceeding of IEEE International Conference on Intelligence and Security Informatics (ISI 2010), pp. 89–94. Vancouver, Canada May 2010
33. Ghaffar, I.A.,Taha, A.E., Bahaa-Eldin, A.M., Mahdi, H.M.K.: Towards implementing agent based correlation model for real-time intrusion detection alerts. In: Proceeding of 7th International Conference on Electrical Engineering, ICEENG 2010, MTC, Cairo, Egypt, May 2010
34. Bahaa-Eldin, A.M., Mahdi, H.M.K., Taha, A.E., Ghaffar, I.A.: Dynamic Parallel correlation Model for intrusion detection alerts, posterIn. In: Annual Information Security Symposium of Center of Education and Research of Information Assurance and Security (CERIAS), Purdue University, West Lafayette. Indiana, USA, March 2010
35. Center of Education and Research for Information Assurance and Security (CERIAS). http://www.cerias.purdue.edu, June 2011
36. Snort—the open source network intrusion prevention and detection system. http://www.snort.org (2010)
37. Basic Analysis and Security Engine (BASE). http://base.securei-deas.net/about.php. June 2010
38. Nessus Vulnerabilty Scanner. http://www.nessus.org. June 2010
39. Nmap- Network Mapper, Security Scanner For Network Exploration & Hacking. http://nmap.org June, 2010
40. Templeton, S., Levitt, K.: A requires/provides model for computer attacks. In: Proceedings of New Security Paradigms Workshop, pp. 31–38. ACM Press, Sept 2000
41. Ning, P., Cui, Y., Reeves, D.S.: Constructing attack scenarios through correlation of intrusion alerts. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 245–254. Washington, D.C., Nov 2002

Chapter 2
Bio-inspired Evolutionary Sensory System for Cyber-Physical System Security

Mohamed Azab and Mohamed Eltoweissy

Abstract Cyber-Physical Systems (CPS) promise advances towards smarter infrastructure systems and services, significantly enhancing their reliability, performance and safety. Current CPS Monitoring, Analysis, Sharing, and Control (MASC) technologies offer disparate and largely inadequate services for the realization of effective and efficient CPS security. Most current technologies did not consider that cyber and physical convergence would need a new paradigm that treats cyber and physical components seamlessly and pervasively. Further, information sharing was severely curtailed by enforcing parameter defense to preserve the privacy of the system to be secured, the Target-of-Security system (ToS). These limitations negatively impact the quality, reliability, survivability, and promptness of security services. In this chapter, we discuss the current challenges to CPS security, survey relebant solutions, and present a novel system, CyPhyMASC, towards realizing pervasive MASC for enhanced CPS security. CyPhyMASC is a bio-inspired intrinsically-resilient, situation-aware system utilized by Security Service Providers (SSPs) to provision MASC services to ToSs comprising numerous heterogeneous CPS components. CyPhyMASC is unique in that it acts as a generic middle layer between the SSPs and the ToSs creating a uniform interface that isolates ToS scale and heterogeneity aspects from control and management aspects. Such isolation and uniform representation facilitate interoperable security services. CyPhyMASC intelligently mixes and matches heterogeneous tools and control logic from various sources towards dynamic security missions. CyPhyMASC is also elastic where situation-driven MASC solutions can be dispatched using dynamic sets of sensor and effector software capsules

M. Azab (✉)
The City of Scientific Research and Technological Applications, Alexandria, Egypt
e-mail: Mohamed.m.azab@gmail.com

M. Eltoweissy
Bradley Department of Electrical and Computer Engineering,Virginia Tech,
Blacksburg, Virginia, USA
e-mail: toweissy@vt.edu

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_2, © Springer-Verlag Berlin Heidelberg 2014

circulating through the ToS rather than using pre-deployed MASC components. Such approach provides evolvable, pervasive and scalable MASC services.

2.1 Introduction

Major physical infrastructure systems such as the water distribution systems and the electric power grid are large-scale complex systems that are expected to be highly reliable and trustworthy. Modern versions of these infrastructure systems go far beyond simple measures to integrate intelligence and automated control into the system through tightly coordinated and integrated cyber components constructing large-scale Cyber-Physical Systems (CPS).

CPS safety and security are prerequisites to assure stability, reliability, and survivability of such mission-critical systems. Security services for CPS are highly dependent on the promptness and accuracy of the Monitoring and Analysis (M&A) mechanisms employed. Traditional M&A approaches do not treat sensing and effecting for cyber components and physical components seamlessly. The current M&A mechanisms were designed based on a set of assumptions that unintentionally neglect the real-time interaction and the tight coupling between these converging components. The assumption was that physical components were protected by isolation and parameter defense while real-time response was not a primary factor for cyber components. Further, they assumed that there is no need to employ privacy preservation techniques as the Target of Security (ToS) privacy is implicitly protected by cyber and physical parameter defense. Additionally, they assumed that resource heterogeneity and scale could still be resolved by a distributed set of heterogeneous, pre-deployed platform-dependent defense tools with fixed resource profiles.

Research works in [1, 2] as well as our own have disputed the validity and correctness of such assumptions as they lead to drastic problems and limitations negatively impacting the quality and promptness of the CPS security service provisioning. Current CPS Security Service Providers (CPS-SSPs) fail to provision trustworthy robust and reliable monitoring and evaluation of the ToS components due to the use of scattered, uncoordinated, uncooperative, unaware, isolated and heterogeneous monitoring tools, and reporting mechanisms. Such limitations increase the use of resources due to redundancy, increase the risk of conflicts, and failures due to limited awareness and coordination, lower the defense quality due to the poor, and boundary limited feedback, increase the latency in security provisioning and in detecting attacks giving the attacker the advantage to spread the attacks through multiple networks, the tool heterogeneity and uncooperative nature massively complicates automating its management, the static nature of such tools complicates attempts to autonomously adapting to changes in the surroundings.

The problem is not only at the monitoring and evaluation phase of the defense provision process, but also at the analysis and investigation phase where the collected feedback gets analyzed searching for attack signs and indications. When the network scale grows exponentially, it becomes almost impossible to analyze the feedback from all the sensors, compile that feedback together, and extract valuable information from it efficiently, and promptly. Additionally, due to the isolation between the monitoring technology and the analysis technology, and the lack of computational power needed to expand the sources of feedback, the current analysis technology does not have enough data to be fully aware of what is globally happening in the network under investigation. Having most of the conventional analysis mechanisms designed to share the same host/host-network for the sake of protecting their privacy, lead to serious limitations. The limited investigation search space, being easy to be targeted by attackers, Can be used to cause a DoS attack, and cannot cooperate in analyzing feedback or share information with out-of-perimeter nodes are examples of such limitations.

In addition to the presented set of limitations in the field of monitoring and evaluation, and analysis of feedback, the control phase has another set of serious limitations too. Control phase represents the stage where the defense system takes actions regarding detected threats face a serious set of limitations.

Control related limitations are mainly the result of lack of cooperation and awareness that limit the defense tools capability to resolve or even contain persistent fast spreading attacks. For example, it is too hard for such uncoordinated, scattered tools to marshal and coordinate task force to hunt down the attacks spreading all over the network or a set of interconnected networks. The reason behind such complexity is the difficulty in autonomously and promptly controls and coordinates both the SSP, and the ToS tools and equipment to block attack access given the current centralized management technology. Further, without appropriate global control, and situational awareness it is too hard to block the source of dynamic fast-changing remote attacks. Such limitations can be utilized to cause DoS attack by keeping the SSP busy treating infected files and strike more and more files. This chapter presents an Evolutionary Sensory System (CyPhyMASC), designed to induce a new paradigm for security service provisioning that intrinsically and comprehensively addresses the aforementioned challenges facing conventional techniques. CyPhyMASC is a biologically-inspired, intrinsically-resilient, intelligent, situation-aware sense and response system to effect biological-immune-system-like security provisioning. We address ToS heterogeneity and scale by enabling dynamic security resource elasticity.

CyPhyMASC is designed to separate the main security provisioning concerns; the tool logic, management and control, delivery mechanism, and physical resources. CyPhyMASC Utilize our smart, biologically inspired, resilient, adaptable, self and situational aware, elastic, and autonomously managed building blocks (the Cell) to construct mobile, dynamic, and runtime-reprogrammable defense carriers to pervasively distribute accurate, trustworthy, and prompt security services. CyPhyMASC acts as a middle layer between the SSPs and the ToS creating a uniform security interface that hides ToSs scale and heterogeneity concerns from control and management.

This uniform representation enables interoperable and cooperative defense. Further, such isolation maintains security provisioning survivability in case of ToS failure and DoS attacks. Additionally, CyPhyMASC autonomously and dynamically profile ToS hosts and direct security services based on the host dynamic behavior and attachments. CyPhyMASC works on a biologically-inspired, intrinsically-resilient, adaptable foundation based on the Cell Oriented Architecture (COA). The COA provides intrinsic dynamic, distributed, resilient resource management and allocation needed to support CyPhyMASC pervasive M&A.

CyPhyMASC manages a vast number of elastic and intelligent containers (Cells) to host/abstract cyber/physical sensing and effecting tools. The COA Cell (the containers) can be described as a micro virtual machine capable of hosting the code logic of a single task. CyPhyMASC mimics the human blood stream circulation effect by utilizing its adaptable infrastructure to circulate these context-driven, functionally customizable sensor and effector Cells into the ToS body to pervasively monitor, analyze and control the TOS components. CyPhyMASC sensors and effectors are used to execute defense missions provisioned by SSP. A defense mission is a mixture of sensing and effecting tasks involving information gathering, partial analysis, control, and manipulation of the ToS elements.

CyPhyMASC can alternate/mix different defense/control missions from different SSPs to provision security services to the same ToS in a process called vaccination. The vaccination process involves sharing security experience and tools between SSPs in terms of abstract missions, and sensing and effecting packages. Vaccines are autonomously checked for privacy violations and maliciousness before utilization or storage. It is exactly like in biological systems where antibodies can be extracted from one immune body to another to create a healthy up-to-date defense community.

CyPhyMASCs main contributions presented in this chapter can be outlined as follows: Enable pervasive autonomously managed monitoring and analysis; Uniform security service provisioning for heterogeneously-composed multi-enclave CPS systems; Enable trustworthy, interoperable multi organization cooperative, dynamic, autonomous security; and Facilitate early failure/attack detection and resolution.

2.2 Attack Scenario

To further motivate our research and to illustrate the effectiveness of CyPhyMASC in achieving its mission we utilize the following working scenario depicting a hypothetical CPS attack named the BlackWidow attack [3]. The players are two competing companies ABC which is the victim, and XYZ that recruited a resourceful attacker to attack ABC.

The BlackWidow (BW) is designed to split into a set of code parts and spread in different directions and locations to decrease the probability of detection. The distribution of parts and the interconnection between the parts in different hosts weave a large web. This web is bi-directionally traversed to send any harvested data from the attacked target and to update the malware with new tools and missions. The BW is designed to be as generic as possible; it is not oriented to any specific application. By constructing the BW web the attacker can start to task the BW towards its designated mission based on the attackers target. The tasks might be remotely assigned through the Internet or preprogrammed in Internet-inaccessible locations (Fig. 2.1).

Fig. 2.1 Commercial security example

The attack is designed to be stealthy by hiding from the security system sensors searching for attack signatures. The attack will target an intermediate host machine that will contain the BW and command and control channel communications. In order to do so, the BW is designed to not harm the host or change any of its settings that might raise the Anti-Malware (AM) alerts. The malware will use minimal resources and will work in a very slow fashion not to alert the network security systems by its existence. The BW uses stolen digital certificates to authenticate its existence in the host machine in the form of drivers. The only way to detect this malware is through deep analysis of the logs of all the communicating nodes, which is computationally very costly to the current systems that share the same host machines. This sharing guarantees host privacy preservation, but limits the capabilities and the awareness of the security tools. The attacker utilizes these limitations to his advantage as illustrated later.

The malware is intended to be targeted, but due to the intentionally random deployment method, the code works in two modes as follows: (1) Benign mode where the malware infects other machines that do not belong to the target space. The machines might be used later in case of target change, or as a base for future attacks; and (2) Malicious mode, where the BW works only on the target host systems. The attacker feedback can determine the mode. The default will be benign unless the attacker changes that or predetermined targets have been programmed.

Attacker assumptions

1. The security and management system shares the same network or host with the target of attack/security system. [Note: security system might be exposed to attack by compromising the ToS. Additionally, Stolen passwords can simply be used to modify rules of IDS, routers, switches, firewalls, proxies, etc.]
2. The attack target security system, or major parts of it, uses COTS and signature based security products 3. The system is computationally incapable of being fully situation aware of all its components in a massive-scale network, in real-time.

3. Security systems are not resilient against attacks, and have weak recovery mechanisms. [Note: most of them assume that they will not be the target of an attack as long as they were able to secure their ToS.]
4. Cyber security is oblivious of and is not coordinated with physical security to protect the target cyber-physical system. Human intervention is need to facilitate such coordination. [Note: the attack can make them conflict with each other to bypass both of them]

Attack procedure (on Air-gapped Target). The attacker uses phishing attack that targets users emails and social network personal pages. The attacker uses social networks as a source of information to generate more convincing phishing emails. These emails will be directed from one of the closely related contacts to the victim. The attacker selects a group of employees working in different branches of ABC. These branches are distributed in various geographical locations, and the victims that will be the malware couriers have no direct relation with each other. The BW is programmed to search the user network for connected computers then it starts using one of the zero days exploits to clone itself into these computers. The attack victims will receive parts of the malware. Each of these parts will contain a fraction of the designated mission and a simple communication module. The communications module will be used to open a direct channel with the attacker and to search and establish communication with other parts. Directions to other parts locations might be sent by the attacker to minimize the search time. The attacker uses malware fractions to construct logical executable entities in the form of mobile software agents targeting different objectives. The first objective will be to search and infiltrate the network for data stores. The malware will sniff network traffic searching for predetermined signatures for such locations. The second objective will be to attack such data stores using the zero day exploits and the stolen certificates to locate targeted industrial secrets, and any available access keys to the protected area behind the air gap. The malware will frequently update the attacker of its findings based on a predetermined update methodology. After successful reception of this data, the attacker will use it to generate legitimate keys to access the air gap. The attacker will use the malware to locate the workstations controlling the surveillance cameras. The malware will record periodic video feeds to be sent to the attacker. These videos with the help of the attacker generated access keys will guide a recruited insider into infecting the air gap with the BW. The malware controlling the cameras will make sure that this process wont be recorded to protect the recruited insider. The air gap malware is programmed to increase the operational hours of certain machines that use specific raw materials manufactured by XYZ to increase XYZ profits. The malware can easily identify such machines by searching a predetermined fixed identifier that must be added to all the programming files targeting such machines. Further, the attacker will use the stolen secrets and designs to equip the malware with the needed logic to randomly manipulate the operational motors frequency in the production machines to induce random defects in the output products to lower its quality. XYZ shall benefit from ABCs loss due to its low quality products. Additionally XYZ will maliciously gain both financially and more control over ABCs production lines by, for example, carefully adjusting the amount of consumed and supplied raw materials. We will revisit this scenario to illustrate how CyPhyMASC secures the ToS against such attack. Additionally we illustrate how CyPhyMASC invalidates the BW design invariants and attacker assumptions.

2.3 CPS Attack Detection and Resolution Techniques

2.3.1 Malware Detection

A malware is malicious software designed to infiltrate or damage a Cyber system or Cyber Physical System (CPS) without the owners informed consent [4]. There are many malware types with different shapes and entry points. Most of these software objects share similar purposes while they are expected to behave differently at time of infection. Viruses, worms, botnets, wabbits, Trojan-horses, exploits backdoors, spyware scumware, stealware, parasiteware, adware, rootkits, blended threats, evolving threats, keyloggers, hoaxes are examples of the different malware types. Figure lists the different types of attacks and the usability ration of each one of them [33].

Each malware group has its own way of being undetected. Modern malware detection tools utilize multiple detection mechanisms to be able to detect multiple malware categories as presented in Fig. 2.2. Malware especially viruses are either memory resident or non-memory resident. Non memory resident are simple attacks that can easily be detected an entry point with a cleaver detection tool.

The memory resident attacks are more complex and efficient that stays in memory and hides their presence from detection tools. These attacks are either fast infectious aiming to infect as much files as possible locally within the infected host or remotely through the host network, and network shares. The second category of memory resident attacks is the slow infectors. Slow infectors are the most dangerous type of malware as it uses stealth and encryption techniques to stay undetected as long as it can. They are powerful attacks that can be a combination of multiple processes working together towards certain objective. Malware detectors use signature based detection techniques to detect known attacks. Signature based detection became very efficient way of detecting known threats [5]. Finding a specific signature in one of the executable codes can accurately identify any enclosed threats within such code. Attack signatures are frequently updated and stored on the local anti-malware database. Unfortunately this technique is inefficient if the attack has a malformed signature either by the programmer or by a mutation engine.

Heuristic techniques are one the most efficient ways to detect such mutated attacks. Heuristic and metaheuristic techniques are used to spot unknown or known attacks with polymorphic behavior. By definition, heuristic technique is an informal technique to solve problems efficiently and in a way close to the optimal path [5]. Heuristic techniques are commonly used to rapidly reach a solution that is somehow close to the best possible solution. The metaheuristic technique is a heuristic method for

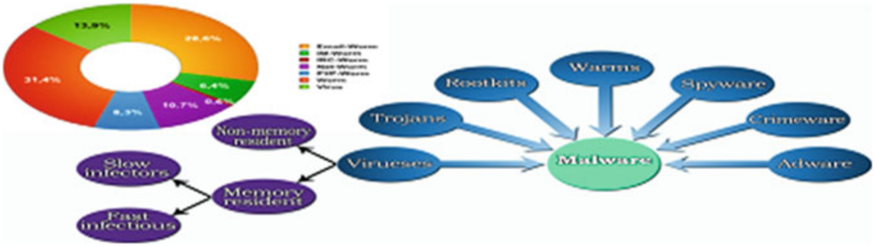


Fig. 2.2 Classification of malware related attacks

solving many of the computational problems by combining user-given black-box procedures in a hopefully efficient way [5].

Most of the modern malware detection techniques that use metaheuristics to detect attacks utilize a set of isolated tools utilizing different techniques hoping in detecting one of the attacks that there is no specific way to detect it. Most of these tools utilize one of the following mechanisms, Pattern matching, automatic learning, environment emulation, neural networks, data mining, byes networks, and hidden markov models. There are other metaheuristics techniques but most of them are built based on one or more of the aforementioned mechanisms.

The main concept of heuristic based detection techniques is to detect attacks without knowing too much about its internal structure. Heuristic techniques mainly focus on examining the behavior and the characteristics of the executing software to anticipate whether it is acting maliciously or not. The most successful heuristic based detection technique named as The Heuristic Scanning Technique utilizes a mixture of multiple metaheuristic techniques such as pattern matching, automatic learning, and environment emulation.

Heuristic scanning in the common sense uses pattern matching to examine the assembly language instruction execution sequence, and qualifies them by their potential dangerousness. Heuristic scanning usually follows a set of built-in rules with preassigned weight on each rule. In case of violation of any of any of the rules the weight of the violated rule is added to the total violated rule by the same program or process. The program is flagged as malicious only if the total sum of added weights exceeds certain threshold. Figure 2.3 illustrated the idea of a single layer classifier with predetermined threshold. The feedbacks from the different scanners are fed into global summarizing point that follows a certain metaheuristic mechanism as illustrated in Fig. 2.4. The overall result will decide whether to flag the scanned object or not. As the detection techniques gets more cleaver, the modern attacks or malware also emerge to more complicated attacks utilizing more sophisticated stealth techniques. Such techniques give them the advantage of being invisible to traditional scanners. Moreover the use of real-time encryption, and anti-heuristic sequences made them looks totally harmless to traditional malware scanners.

Heuristic scanners that use single metaheuristic mechanism that focuses only on monitoring the execution flow of the instructions of a certain program are deceivable

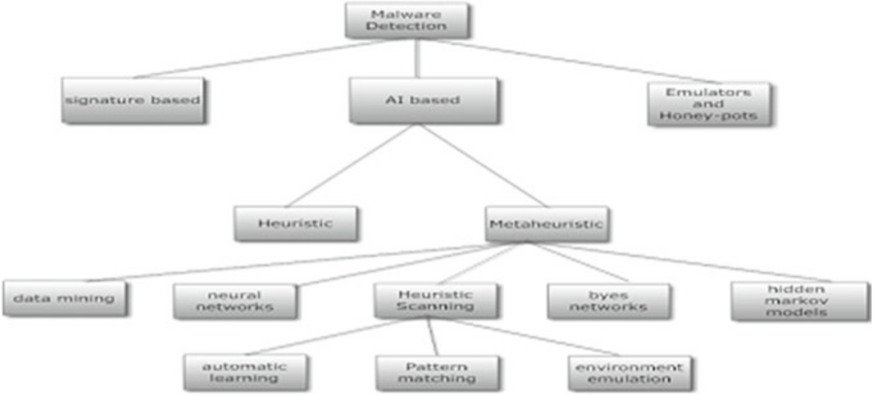


Fig. 2.3 Classification of malware related attacks

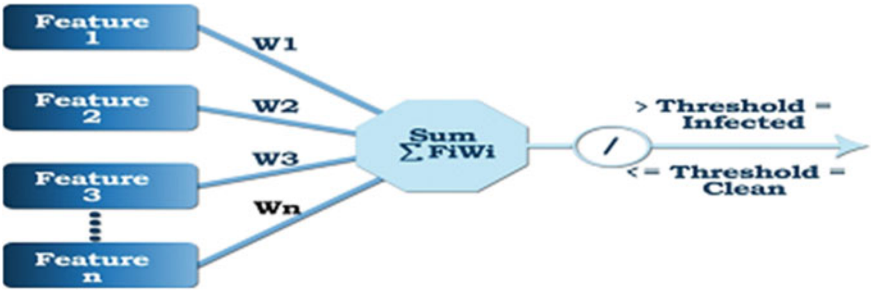


Fig. 2.4 Single layer classifier

by code obfuscation. Code obfuscation occurs by embedding some meaningless instructions within a malicious code. The same technique deceives detectors utilizing heuristic and signature scanning combined together. One of the successful mechanisms to resolve the aforementioned problem is the use of artificial runtime environment emulation. However, it is not a light weight detection mechanism, but it has high success rates in detecting unknown attacks. Environment emulation utilizes the idea of virtual machines; the malware detection tool provides a virtual machine with independent and isolated operating system and allows malware to perform its routines freely within the virtual environment. The execution behavior of the suspicious application is being continuously examined while the malware is not aware. Most of the stealth and anti-heuristic techniques are irrelevant in this case, as the detection tools scan the behavior from outside the box with a clear vision of what is really happening inside.

The main problem facing such technique is the massive resource consumption and the expected delay needed to construct the virtualization environment, and infiltrate the harmful instructions from being executed on the real machine. Another problem that arises with using heuristic methods for detecting malwares is the possibility of false positives. A false positive event occurs when a benign program gets flagged as malicious by the heuristic scanner. The problem occurs frequently specially with noncommercial programs having suspicious routines through their encryption functionalities.

The use of automatic learning is a good resolution of such problem, where the detector learns from its mistakes. The main issue with this technique is it requires an advanced user. In order to resolve such problem autonomically, detection scanners have to increase their scanning depth, and combine feedback from multiple heuristic mechanisms. Also external consultation is one of the most efficient techniques, where an external resourceful node gets consulted for guidance related to suspicious programs with weights that parley cross the threshold line. The only issue with that solution is the possibility of privacy violation due to sending specifics about the suspicious events.

Recently more complicated attacks were introduced that depends on infecting and controlling multiple hosts creating an automated taskforce targeting multiple objectives. Such attacks usually have dynamic objectives, and construction components. Additionally, they are frequently and autonomically get updated using a dynamic up/down link between the attacker and the malware itself. Detecting such attacks is a very complicated task given the uncooperative nature of the conventional modern detection tools, and the fact that they share the same host, or host network with their ToS.

Sharing the same network or host with the ToS makes them an easy target for attackers to deceive, or destroy [6, 7]. Additionally, the successfulness of the malware detector depends mostly on the fast real-time, and deep analysis of the scanners feedback. Such process, especially when it involved creating a runtime emulated execution environment is a computationally costly process for a tool that shares the ToS resources.

2.3.2 Standalone and Distributed Monitoring and Evaluation Solutions

Defense services for CPS are highly dependent on the promptness and accuracy of the Monitoring and Analysis (M&A) mechanisms employed. Traditional M&A approaches do not treat sensing and effecting for cyber components and physical components seamlessly. The current M&A mechanisms were designed based on a set of assumptions that unintentionally neglect the real-time interaction and the tight coupling between these converging components. The assumption was that physical components were protected by isolation and parameter defense while real-time response was not a primary factor for cyber components. Further, they assumed that there is no need to employ privacy preservation techniques as the Target of Defense (ToS) privacy is implicitly protected by cyber and physical parameter defense. Additionally, they assumed that resource heterogeneity and scale could still be resolved by a distributed set of heterogeneous, pre-deployed platform-dependent defense tools with fixed resource profiles.

Research works in [1, 2] as well as our own have disputed the validity and correctness of such assumptions as they lead to drastic problems and limitations negatively impacting the quality and promptness of the CPS security service provisioning. Current CPS Defense Service Providers (CPS-SSPs) fail to provision trustworthy robust and reliable monitoring and evaluation of the ToS components due to the use of scattered, uncoordinated, uncooperative, unaware, isolated and heterogeneous monitoring tools, and reporting mechanisms. Such limitations increase the use of resources due to redundancy, increase the risk of conflicts, and failures due to limited awareness and coordination, lower the defense quality due to the poor, and boundary limited feedback, increase the latency in security provisioning and in detecting attacks giving the attacker the advantage to spread the attacks through multiple networks, the tool heterogeneity and uncooperative nature massively complicates automating its management, the static nature of such tools complicates attempts to autonomously adapting to changes in the surroundings.

Research presented in [8–11] attempted to resolve some of the problems resulting from such assumptions using more flexible sensing and control elements. They devised a mobile multi-agent based attack detection system. The presented solutions were situation unaware and offered limited defense-tools pervasiveness and coordination. Generally speaking, provisioning security services while sharing the same host with the ToS exposes the ToS to DoS attacks, and limit the systems scalability and interoperability.

Works in [11, 12] utilized a multidisciplinary approach to intelligently resolve some of the presented limitations. They combined multiple artificial intelligence techniques to build a complex smart attack detection system. Unfortunately, these techniques were bounded by the available technology constraints; they were designed to provision dedicated security service while sharing the ToS host or host network. They were unable to overcome the curse of complex systems dimensionality. With the increase of system complexity and numerousness of input features, the processing time involved with clustering system events might badly affect system, and attack detection timeliness. Time constraints may sometimes force the system to prune less important features (dimensionality reduction) to maintain system timelines. However, the pruning approach is not always possible as it might compromise the detection accuracy.

All the above mentioned approaches were mainly concerned with security service provisioning for cyber components. The work presented in [13, 14] is a hardware based static detection system capable of supporting the requirements of both cyber and physical components. Using hardware based detection and analysis techniques guarantee prompt, and resource efficient response for quickly spreading attacks. A major disadvantage of technology is its limited flexibility, adaptability, interoperability, and maintainability. These systems are designed to work for specific target and cannot seamlessly adapt to match different targets. Multiple attack detection solutions were presented utilizing mixtures of the abovementioned methodologies employing different M&A techniques [15, 16], Unfortunately, none of these systems where capable of presenting a comprehensive, autonomous, interoperable, globally situational aware and scalable solution that can guarantee adequate security provisioning quality and promptness while maintain the ToS survivability, operability, and privacy. Up to our knowledge CyPhyMASC is the first solution that can provide such features comprehensively and pervasively with low overhead.

2.3.3 CPS Related Control Solutions

In addition to the limitations presented in the previous two sections, in regards to monitoring and evaluation, and analysis of feedback, the control phase; where the defense system takes actions regarding detected threats face a serious set of limitations [7]. The limitations are mainly due to the lack of cooperation and awareness that limit the defense tools capability to resolve or even contain persistent fast spreading attacks.

For example, it is too hard for such uncoordinated, scattered tools to marshal and coordinate task force to hunt down the attacks spreading all over the network or a set of interconnected networks as it is hard to control the SSP, and the ToS tools and equipment to block attack access to the shared network. Further, without appropriate global control, and situational awareness too hard to block the source of dynamic remote attacks. Such limitations can be utilized to cause DoS attack by keeping the SSP busy treating infected files and strike more and more files.

Research work has been focusing on presenting a resolution for some of the control problems in CPS environments. Researchers in [17] presented what is called Autonomous Multi-agent Cooperative Problem Solving (TEAM-CPS), and successfully applied it on one of the critical CPS, the public telephone networks. They used multi intelligent agents that were designed to work together to provide distributed control for such system. Unfortunately, the system was not scalable enough to suit large scale systems. The limitations against this approach and other agent passed approaches like the work presented in [18, 19] is the high resource consumption nature of the agents, and the fact that they are designed to share the host resources. These limitations limit the approach capability to scale.

From another perspective, the use of intelligent agents lacks the support of the physical part of the network. The used agents are not aware of the interactions between the cyber and the physical parts of the system. Such unawareness increases the chance of conflicts, errors, and failures.

A more advanced version of this line of research was resented by the work of [7, 18] as they used multiple AI techniques to control a pool of mobile agents performing control tasks. The use of AI guided the management platform towards smarter decisions. Unfortunately, they shared the same problem of their insisters, the lack of situational awareness, and the inconsideration of isolating the control platform from the host under control. Such limitations limited the scalability of such systems, and their ability to suit CPS applications.

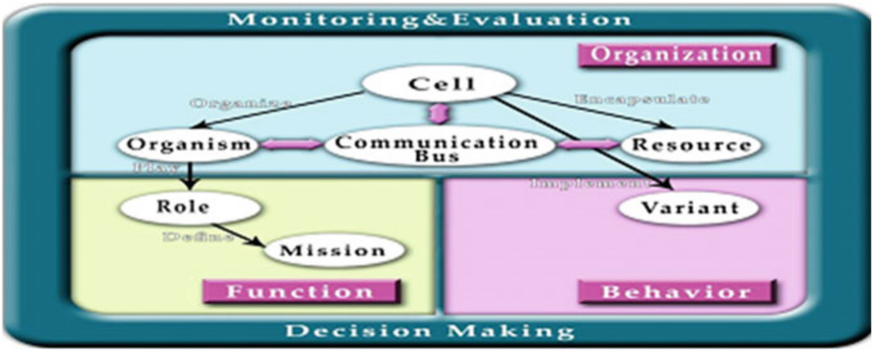


Fig. 2.5 Components of the COA

2.4 Evolutionary Sensory System (CyPhyMASC)

2.4.1 Overview of the Cell-Oriented Architecture

The COA employs a mission-oriented application design and inline code distribution to enable adaptability, dynamic re-tasking, and re-programmability. The Cell, is the basic building block in COA, it is an abstraction of a mission-oriented autonomously active resource. Generic Cells (Stem Cells) are generated by the host middleware termed COA-Cell-DNA (CC-DNA for short), then they participate in varying tasks through a process called specialization. Cells are intelligent, independent, autonomous, single-application capsules that acquire, on the fly, application specific functionality in the form of an executable code variant through the specialization process. Cells are also dynamically composable into larger structures Organisms representing complex multi-tasking applications. An Organism is a dynamic structure of single or multiple Cells working together to accomplice a certain mission. Figure 2.5 illustrates the different aspects and components of the COA. Applications built over this COA are constructed as a group of cooperating roles representing a set of objectives. An Organism represents a role player that performs specific mission tasks. Organisms are bound to functional roles at runtime. Note that different Organism variants can be bound to the same functional role at the same time leading to heterogeneous redundancy to support enhanced resilience.

2.4.1.1 The Cell

Conceptually, the Cell is the basic atomic active resource in a distributed computing platform. Cells act as a simple, single application module capsules (virtualization environment, or sandbox) isolating the executable logic from the underlying physical resources. Cells acquire, on the fly, application specific functionality in the form of an

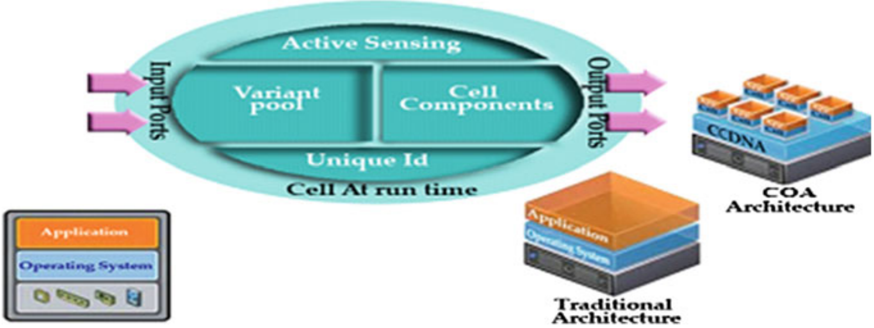


Fig. 2.6 COA cell at runtime

executable code variant. COA Cells are equipped with multiple tools supporting their intrinsic features regarding resilience, dynamic performance optimization, Cell self and situation awareness, and online programmability and adaptability. Figure 2.6 illustrates an abstract view of a COA Cell at runtime. A single workstation can host one or more Cells, providing a flexible way to share the physical resources among multiple applications. The COA Cell encapsulates the running application via elastic dynamic emulation of a suitable execution environment. Such encapsulation isolates the running application from the host platform decoupling the executable logic from the underlying physical resources. The COA Cell is designed to work on a redundant copy of the applications critical data, while all the application critical data are always saved in a remote safe warehouse. The intelligent COA management platform described in [20] works on maintaining all the aspects related to these operations seamlessly.

Decoupling data, logic, and physical resources enable the COA Cells to seamlessly move at runtime between different hosts regardless of their configuration heterogeneity. A successful utilization of such mobilization for security objectives is presented in [21]. CyPhyMASC also utilizes these features by employing the COA Cells as security-service carriers. CyPhyMASC circulate these carriers through the ToS network for host data collection and security service delivery as illustrated in the next sections.

2.4.2 The Foundation

CyPhyMASC is an evolutionary sensory system designed to enable real-time pervasive monitoring and analysis towards autonomous context aware security service provisioning. CyPhyMASC security provisioning platform is composed of three main layers, the management layer, sensor and effector abstraction layer, and sensor and effector tools layer as presented in Fig. 2.8. The three layers are founded over a COS based foundation. The management layer rules are played by a set of organisms composed of Cells. The abstraction layer uses COA Cells to encapsulate attack investigation and resolution tools defined as binary code variants (APIs) constructing a set of platform independent sensing and effecting capsules. CyPhyMASC constructs a biological immune system like, defense environment by circulating generic streams of such capsules into the (Target of Defense) ToS body to induce a blood stream like effect. The following section illustrates and provides more technical details about the defense-capsules creation process, and the security provisioning methodology of CyPhyMASC.

2.4.2.1 CyPhyMASC Organisms and Capsules Composition

CyPhyMASC leverages the COA ability to abstract, encapsulate, and virtualize heterogeneous physical and logical resources into unified programmable objects Cells. Cells are sandboxes internally construct a suitable working environment for heterogeneous tools. Externally, it is capable of changing its characteristics to work with many targeted architectures. Regardless whether the sensing target was a computer in a network, or a physical sensor with COA-ready digital interface COA Cells will hide these differences from the enclosed sensing/effecting API.

CyPhyMASC uses the middleware (CC-DNA) installed on the ToS host machines to instantiate, deploy, and host sensor and effector Cells. CyPhyMASC sensors and effectors are a set of precompiled APIs with specific sensing or effecting tasks. Sensors and effectors come with a detailed specification file describing the targeted platform, estimated computational Wight, needed libraries to support it, possible conflicts, ... etc.

CyPhyMASC defense mission (organism role) is defined using a custom-made programming language used to generate scripts defining the structure, workflow, and the set of tasks for the sensing and effecting organisms. Additionally, it also defines the type of sensors and/or effectors needed to execute that mission.

Organism creation starts when the host CC-DNA receives the logic script. CC-DNA interprets this logic to construct the organism sensor or effector Cells. In COA, resources can easily be acquired when needed. CC-DNA might ask the local or remote logic reservoir for any sensor or effector APIs that are required to execute the designated sensing or effecting mission in case they were not already available on the targeted host.

2.4.3 CyPhyMASC Security Provisioning Methodology

CyPhyMASC manages and controls sensing and effecting Cells based on specific mission objectives provided by the SSP. CyPhyMASC works as a middle layer between the SSP and the ToS as illustrated in Fig. 2.8. CyPhyMASC leverages the uniform abstract representation of sensing and effecting Cells to circulate sensors and effectors intelligently throughout the ToSs. Additionally, CyPhyMASC leverages

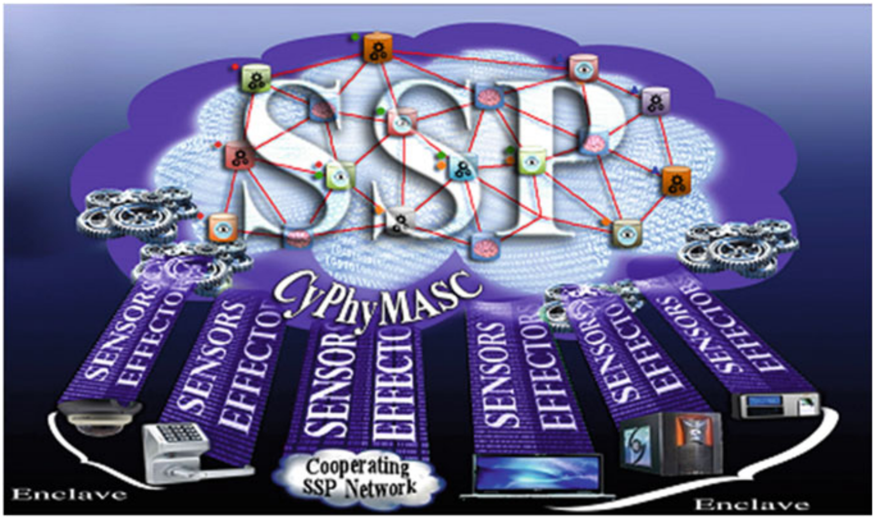


Fig. 2.7 CyPhyMASC abstract view

such features to enable security service interoperability; where different SSPs can share in security provisioning for the same ToS in a privacy-preserving manner as will be explained shortly. CyPhyMASC security provisioning includes two main modes, a SSP-guided mode where CyPhyMASC blindly execute predetermined defense missions provided by the SSP; and an evolutionary-mode that involves evolutionary sensing, and effecting. The SSP-guided modes use CyPhyMASC as a delivery platform that executes certain commands blindly without being involved in the details of the process. CyPhyMASC collects runtime commands and deliver real-time feedback to the SSP. This mode is highly un-scalable and not recommended for large systems. As this is a limited version of our Evolutionary sensory system, we will move forward and illustrate more about the more generalized mode the evolutionary-mode (Fig. 2.7).

2.4.3.1 Overview and Initial Configuration

Evolutionary sensing aims to detect malicious abnormal behaviors without prior knowledge of that behavior. In both modes, CyPhyMASC maintains minimum level of security by maintaining the normal work mode of the currently-being-used tools on the ToS. CyPhyMASC treats such tools as part of its sensing and effecting arsenal. The Evolutionary sensing process involves analyzing and correlating different information feeds from multiple sources to magnify up-normal behavior deviation identifying possible attack indications. Evolutionary effecting involves utilizing the pervasive control feature of CyPhyMASC to autonomously deploy safe-resolution tools “that doesn’t conflict with the running applications,” or to contain such attacks within certain perimeter while waiting for administrators to provide clear resolution procedure to execute. The deployment or containment mechanism works based on an intelligent and dynamic profiling mechanism.

The profiling mechanism works on the fact that attacks can be directed attacks working towards certain objective, or undirected attacks that seek maximizing the victim losses. Even for undirected attacks they can be considered as directed attacks at certain levels. For example, at the operating system levels, windows based attacks cannot infect Unix operating hosts, Java based attacks cannot infect C based software packages, etc. The working environment of a certain host can significantly limit the type of attacks infecting this host even for undirected attacks. Based on that, we can easily classify attacks into groups based on certain classification protocol. Such classification can be attributed with a set of parameters determining the likelihood of having an attack under this class/group in one of the hosts within certain boundaries.

CyPhyMASC has a set of pre-deployed manually/automatically generated profiles used to direct the sensor circulation and the basic deployment package for each host. CyPhyMASC adapt such profiles all the time to maximize the efficiency and accuracy of security provisioning. The dynamic adaptation of profiles adjusts the definitions defining the needed type and density of defense missions within each profile. Each profile is configured to match the host, host network, attached organization or enclave settings.

2.4.3.2 Joining CyPhyMASC Network

Joining an CyPhyMASC-equipped SSP network procedure starts by installing the CC-DNA on the host machines, registering the hosts' physical IPs, hosts configurations, and their security and privacy policies to the SSP host database, and classifying the host based on one of CyPhyMASC Profiles. Usually the host follows the general profile of the enclave/organization that it belongs to. However, CyPhyMASC can assign a more fine-grained profile to a certain set of hosts within the same network if they are supposed to behave differently at runtime. The configuration of such fine grained profiles can change over time if the behavior of the host or the surrounding changes.

The host configuration profile illustrates all the details regarding the host platform, computational capabilities, the organization /enclave id(s) for that host if any, and any special consideration regarding the applications running on it. The security and privacy policy defines the needed security level, the scope of cooperation, and the type of allowable sharing materials.

Upon registration of a new cyber/physical host, CyPhyMASC is notified to start the initial evaluation of the host to determine the profile that the host will follow, identifying the basic sensor deployment-package composition-profile. CyPhyMASC interprets the host record in the SSP database to identify the appropriate types of sensors and effectors APIs that match the host configuration-profile. CyPhyMASC will deploy the evaluation-sensors package to initiate the initial checkup to verify that the host the minimum requirements needed to join the SSP network. In case of any problems, CyPhyMASC will autonomously deploy the appropriate effectors to resolve it.

CyPhyMASC frequently change the basic deployment-package by circulating new sensors to replace old ones. The process is guided by the global sensing feedback not only at the host but also through the network, and the security provisioning profile that the host is following. CyPhyMASC use a grading system to continuously evaluate sensors on each host based on their success to detect up-normal behaviors. In normal situations, at each evaluation-round, one of the new most successful sensors within each profile replaces the least successful senor in the basic deployment-package of the host. The details about sensor circulation are illustrated later.

2.4.4 Evolutionary Sensing and Effecting Framework

After joining CyPhyMASC Network, the host defense related aspects will be handled by one of CyPhyMASC management units. The management unit works as a part of CyPhyMASC sensing and effecting framework presented in Fig. 2.8. CyPhyMASC sensing and effecting framework is classified into three main layers management layer, sensing and effecting abstraction layer, and the defense delivery tools as illustrated in Fig. 2.8. The sensing and effecting tools layer is a set of logical sensing or effecting APIs stored in the local reservoirs. These tools are autonomously abstracted at runtime into uniform sensing and effecting Cells participating in the construction of organisms playing certain defense missions. CyPhyMASC organisms are anonymously constructed, managed, and controlled at runtime by CyPhyMASC management layer. This layer is responsible for collecting, correlating, and analyzing sensor feedbacks. Additionally, this layer is responsible for taking decisions based on the sensing feedback, previous historical events, and SSP guidelines. Such decisions might involve composing more capable effecting defense missions for resolution or new sensing missions for deeper investigation.

CyPhyMASC management layer is a tree-like hierarchical construction, where hosts are connected to leaf-Brains "decision making organisms" to be monitored and controlled as presented in Fig. 2.5. Based on CyPhyMASC administrator settings, each leaf-Brain manages a specific number of hosts. leaf-brains frequently reports to their parents "Higher-level brains" for more comprehensive guidelines.

2.4.4.1 Feedback Management and Representation

CyPhyMASC Sensors feedback is a score-sheet like report that compares the behavior deviation regarding the sensing target to a predetermined threshold. Sensors are classified into different sets representing their targeted sensing objectives "ex, memory, communications, privacy, .. etc." Thresholds are dynamically adjusted based on the nature of each host, and the number of false negatives/positives reported by the Sensor. Figure 2.9 illustrates the feedback analysis process. Score sheets from

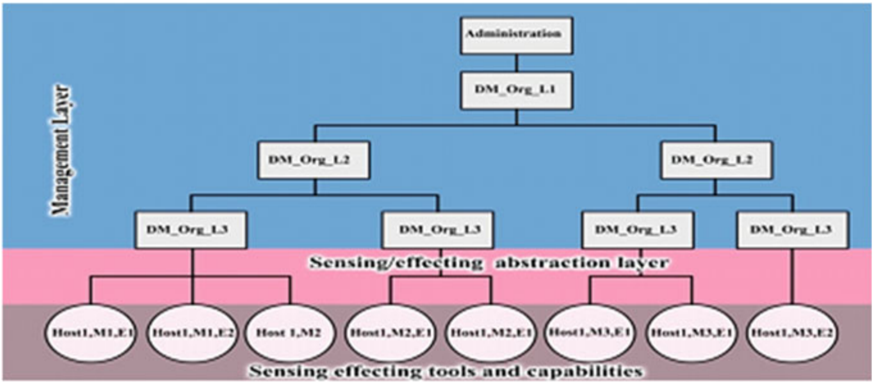


Fig. 2.8 CyPhyMASC abstract view

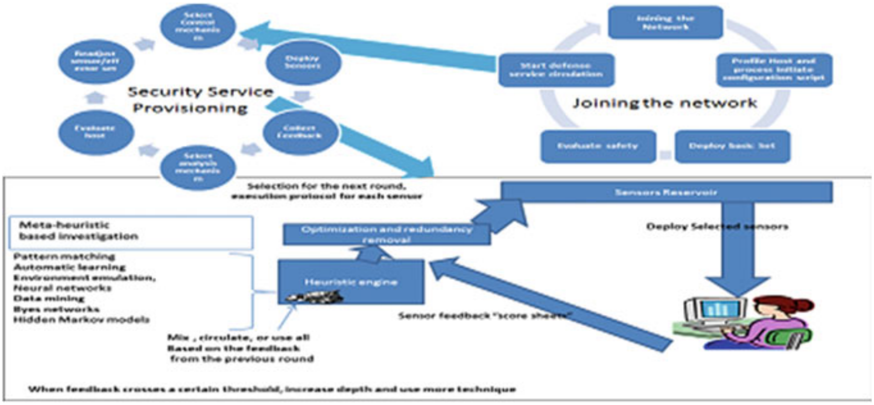

Select Control mechanism
Readjust sensor/effector set
Deploy Sensors
Security Service Provisioning
Evaluate host
Collect Feedback
Select analysis mechanism
Joining the Network
Start defense service circulation
Profile Host and process Initiate configuration script
Joining the network
Evaluate safety
Deploy basic list
Selection for the next round, execution protocol for each sensor
Sensors Reservoir
Meta-heuristic based investigation
Optimization and redundancy removal
Deploy Selected sensors
Pattern matching
Automatic learning
Environment emulation,
Neural networks
Data mining
Byes networks
Hidden Markov models
Heuristic engine
Sensor feedback "score sheets"
Mix , circulate, or use all Based on the feedback from the previous round
When feedback crosses a certain threshold, increase depth and use more technique

Fig. 2.9 The evolutionary security provisioning process

different sensors for the same host are sent to the leaf-brain to compose comprehensive score-sheets to be checked against the defense rules database. Each defense rule has a score-sheet attached to it. Rule-sheets have values for different objects “ex, memory, communication, .. etc” reflecting the behavior patterns “attack signature” of each object in case of infection. Behavior pattern description can be discreet or continuous. Rule description also includes the host sampling procedure. Sampling procedure describes the needed number of samples per object and the duration of each sample, and the sensors needed to takes such samples.

The leaf-brain checks the partial similarity between the sensor feedback “score-sheet” and the existing rules-sheets to allocate the most useful rules for the next deployment round. Based on that selection, the leaf-brain will compose a new organism, holding the list of sensors mentioned in the rule description, with a set of preprogrammed tasks based on the rule sampling procedure. Based on the feedback, threats might be detected, and the resolution mechanism described in the rule will be followed. The leaf-brain will compose a new effector organism with list of effectors mentioned in the rule description and the execution workflow described in the rule.

Rules are under full time update by the parent-brains, and the SSP. Leaf-brain experience is frequently reported to the parent-brains for further guidelines. Parent-brains can construct a more comprehensive view of the whole network by correlating the leaf-brain feedbacks. Such views can magnify certain behavior deviation across the network, which will guide the composition of new defense roles to be executed by the leaf-brains for further investigations.

2.4.4.2 CyPhyMASC Information Sharing (Vaccination)

At the higher levels of the hierarchy "the parent-brains", the collected incident reports with the rules, sensors, and effectors used are archived for sharing. The Clearing-House Organism (CHO) role will be easy in this case, as it will check the ToS privacy policy against sharing of such materials. As described before, the shared materials carry no indications about the specific incident source. It is similar to defense related tips that encourage SSP to apply a specific rule because it might reveal certain threats. The reason that motivated SSP to share such information might be the rule successfulness to detect a threat at one of her ToSs or it is a new rule that was developed with promising results. However, the ToS privacy policy will be checked in case it prophets even that level of information sharing.

The CHO role becomes more significant when the SSP asks cooperating SSPs to provide a solution for a problem she has. In this case, the reported suspicious score-sheet and the sensors used to extract it will be shared with other SSPs. CHO will check that material to make sure that the information enclosed dose not contradict with the ToS privacy policy. CHO rejected authorizations are reported to the administrator to manually override or discard. CHO will also check the SSP feedback regarding such requests, new rules, sensors, and guidelines might be provided as a resolution for the problem. Authorized solutions will be deployed, and rejected ones will be reported to the administrator for further guidelines. The details and the framework of defense mission sharing will be described later.

2.4.5 CyPhyMASC Brain Architecture

Figure 2.10 illustrates CyPhyMASC brain architecture and composition of organisms and the interactions between these organisms to achieve CyPhyMASC goals. I will briefly describe each component and its dedicated task. The Analysis and investigation organism is responsible of analyzing the continuous feedback from the ToS deployed sensors. The analysis and investigation organism sends reports to the Decision making organism to take decisions regarding composing new defense missions, authorizing the use; selecting the type of the effectors if needed. The Role composer organism will use the decision making organism guidance to compose new

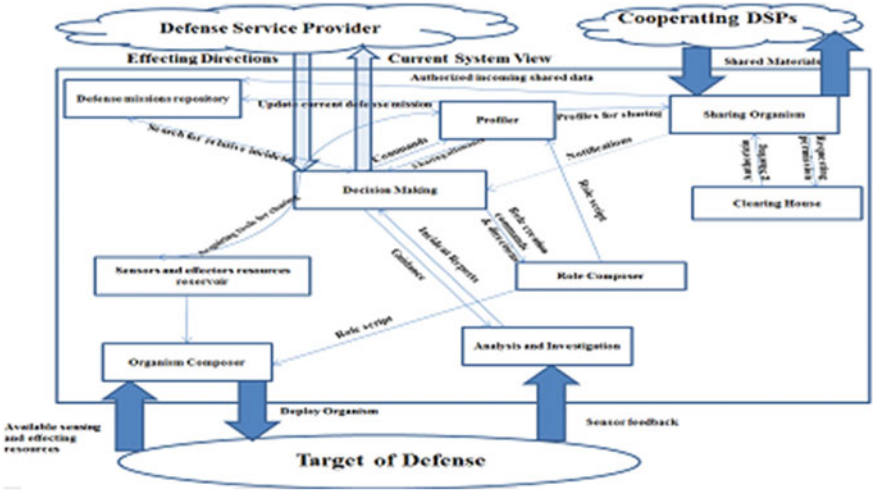


Fig. 2.10 CyPhyMASC architecture

defense roles. The role composition is described using our custom made mission definition language. The role script will be sent to the organism composer. The organism composer is a part of the CC-DNA installed on the host. The organism composer organism is responsible for the resource virtualization process. It will abstract the host resources to compose the requested organism Cells. These resources might be physical resources memory, processor, or logical resources in the form of conventional defense tools, or any locally stored sensors or effectors. If any of the resources needed to compose the cell did not already exist on the host, the cell composer will acquire these resources from the Sensor and effector Reservoir. Upon generating and testing new defense missions the decision making unit might decide to instruct the Profiler to generate a profile for the mission with the used tools to be shared with other cooperating SSPs. The sharing process is handled through the sharing organism that will manage sending and receiving shared materials between SSPs. Sharing or employing shared materials has to be authorized by the Clearing House organism. The Defense missions repository organism will hold history of defense mission usage either locally within the same SSP or globally through feedbacks from other cooperating SSPs with an evaluation for such missions for future reference.

2.4.5.1 Information Sharing and Exchange Protocol Within CyPhyMASC

One of the main contributions of CyPhyMASC is the trustworthy information sharing and exchange. CyPhyMASC share attack events and detection/resolution materials between different management organisms locally within the same organization, and globally between cooperation SSPs. Figure 2.11 illustrates CyPhyMASC defense mission sharing protocol.

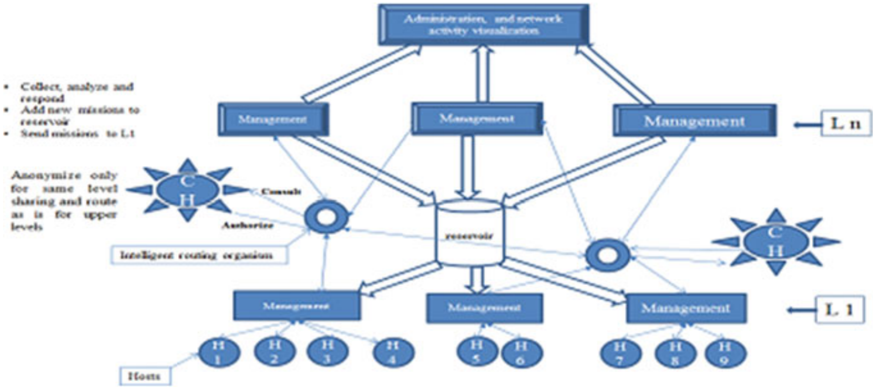

Administration, and network activity visualization
Collect, analyze and respond
Add new missions to reservoir
Send missions to L1
Management
Management
Management
L n
Anonymize only for same level sharing and route as is for upper levels
C H
Consult
Authorize
Intelligent routing organism
reservoir
C H
Management
Management
Management
L 1
H 1
H 2
H 3
H 4
H 5
H 6
H 7
H 8
H 9
Hosts

Fig. 2.11 CyPhyMASC defense mission sharing protocol

As mentioned before, the security provisioning process is managed by a hierarchy of management organisms. At the leaf nodes we have the management layer that manages defense mission circulation and execution on the ToS hosts. Such layer collects the sensor feedback, and the detected list of events that was sent to it through the score sheet technique described before. The collected materials are forwarded to one of the distributed routing organisms.

The routing organisms collect the incoming reports and send them identifying the report-source to the higher layer management. Additionally, all reports that was classified as important events by the report source is anonymized and checked for privacy policy violation by the clearing house organism to be forwarded directly to the others local management organisms. The clearing house organism, ask the local broadcasting organism to handle this task. The broadcasting organism will remove any duplicated reports regarding same event, or same defense mission and instruct the leaf management units to raise the score of the missions related to the reported events based on the severity of the event. Raising the score of a certain mission will increase the chance of applying it in the next mission circulation round. At the upper level of the management hierarchy, the collected reports from the lower level management units are correlated and analyzed to be presented to the network administrator through visualization software. Using such software will make it easier for the administrator to visualize the attack activity, the security provisioning process, and to provision manual override or further guideline if needed.

The higher management units will analyze the collected reports providing further guideline to the lower layers. It is also responsible for generating new missions to be stored in the missions repository. The upper layer management organisms layer provides guidelines to leaf management organisms. The upper management layer organisms can add a new package of sensing or effecting missions to be executed on a specific or a set of ToS hosts at specific time event; or by prioritizing or de-prioritizing in case of many false positives one of the missions already being used by this management unit. SSP level sharing is the responsibility of the higher management layers

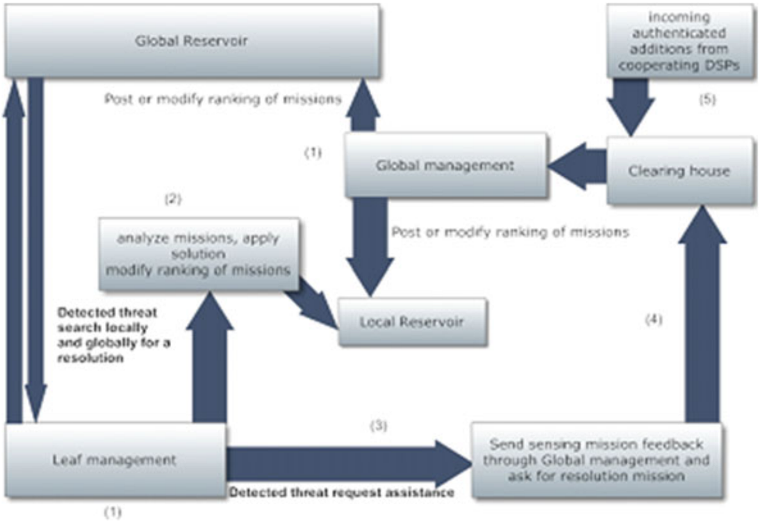

Global Reservoir
incoming authenticated additions from cooperating DSPs
Post or modify ranking of missions
(5)
(1)
Global management
Clearing house
(2)
analyze missions, apply solution modify ranking of missions
Post or modify ranking of missions
Detected threat search locally and globally for a resolution
Local Reservoir
(4)
(3)
Leaf management
Detected threat request assistance
Send sensing mission feedback through Global management and ask for resolution mission
(1)

Fig. 2.12 The defense mission lifecycle

only. This level of sharing occurs upon the reception of a highly suspicious incoming event. The management organisms mark such event and its relative defense mission for sharing. A dedicated organism named as the sharing organism, anonymizes the shared material, and authorizes it for sharing via the clearing house organism. If the shared materials were not in violation of any of the ToS privacy policies, the shared materials are sent to all cooperating SSPs either as an alert or asking for a resolution guidelines. SSP attack alerts should include details about the attack and detection/resolution methodology in terms of defense missions and sensing/effecting tools. If the SSP was asking for an advice related to certain malicious activity, the abstract information related to how such activity is identified and formalized as a sensing only mission, all the tools needed to execute the mission is added to it to be broadcasted to the cooperating SSPs.

Upon reception of such missions, the receptor makes sure that there are no privacy violations. The mission is then tested in a controlled environment before applying it to the SSP attached ToSs. If the request was for a resolution guideline, and the receptor SSP has the resolution methodology, the resolution is composed in a defense mission format attaching all the needed sensing and effecting elements to execute it and sent back to the source. The source test the resolution tools in a controlled environment after making sure that there is no privacy violations then based on the administration opinion it might be deployed to resolve the reported attacks. Figure 2.12 represent the defense mission life Cycle.

2.4.5.2 Intelligent Attack Detection and Resolution

The main objective of any attack detection mechanism is to accurately and properly direct the correct defense tools towards matching attacks. Researchers proved that accurately identifying attacks is an NP-Complete problem [22], while others proved that it might be considered as an NP-hard problem as well [23]. The conclusion is that the problem cannot be solved in realistic time as the problem space expands exponentially over time. The use of Heuristics is always considered to be a good solution for such problems [24]. The most successful malware detection mechanisms uses heuristic scanning and signature based mechanisms to detect attacks. Heuristics scanning in its basic form is an implementation of three metaheuristics mechanism, the pattern matching, automatic learning, and environment emulation. Due to the high computational cost of running such heuristics based techniques, modern anti-malware techniques that are usually shares the same host or the host network of their ToS use a limited set of the available metaheuristics techniques. The reason behind that is to save the computational resources and to speed the process of classifying the executable tasks without interrupting their execution. CyPhyMASC is designed to work in total isolation of the ToS, and to isolate the main design concerns of malware detection and resolution, sensing, effecting, and control logic. Working in isolation from the ToS enabled CyPhyMASC to use more heuristics techniques and increase the depth of learning and investigation of such techniques without any negative impact on the ToS performance, or resources. Most of the workload on the analysis and investigation is waved to the SSP platform, and the ToS participates only in hosted sensing and effecting elements for a limited time frame. As presented before separating the security provisioning design concerns enabled CyPhyMASC to optimize the process of sensing and effecting saving more of the ToS resources. CyPhyMASC intelligent sensor reuse mechanism uses the same sensor to feed multiple heuristic techniques, to save a considerable amount of the computational power of the ToS. The next sections illustrate the sensor circulation, selection and reuse mechanisms of CyPhyMASC.

2.4.5.3 Profile-Guided Sensor Circulation

CyPhyMASC circulates its sensors and effectors to execute defense missions. The sensor circulation protocol depends on the tool requesting sensor deployment. The security provisioning process involves cooperation between multiple detection and resolution tools. Each tool submits a sensor deployment request to the sensor repository to deploy the requested sensors. The requests pass through optimization unit to remove sensors that were recently deployed before. Figure 2.13 illustrates the sensor selection and deployment procedure.

Classification based on profiles

As mentioned before, CyPhyMASC attach hosts to certain profiles based on the host engagement with its organization and enclave, host configuration, behavior,

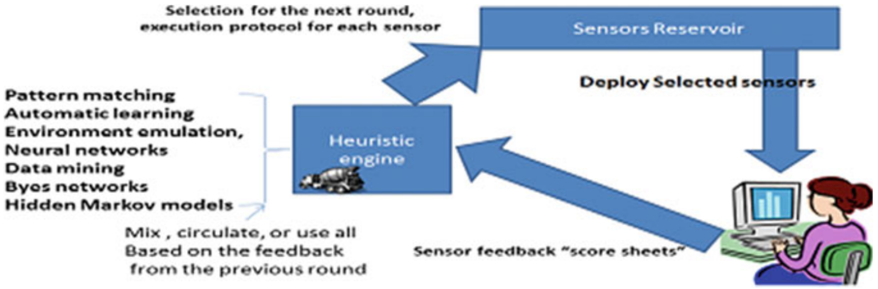
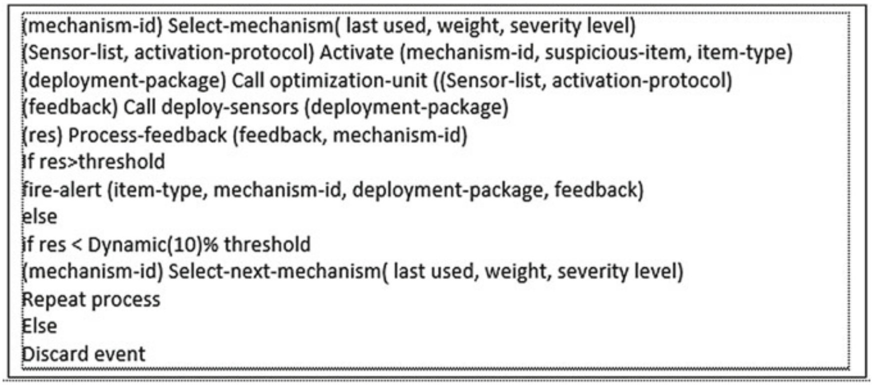


Fig. 2.13 Sensor selection and deployment

usage pattern, etc. some of these profiles are static and preloaded with CyPhyMASC management units, and we call them the coarse grained profiles. Such profiles focusses on static classification aspects, like organization id, enclave id, platform configuration, network protocols, etc. this profiles determine the general security provisioning pattern for the host. The second type of profiles is a fine grained dynamic profile that determines the usage behavior of the host and mostly reflects the user behavior using the host. The type of applications being frequently used, the hours of operation are good examples for the aspects controlling such level of profiling. This profile type is dynamically adjusted based on changes on the usage behavior. CyPhyMASC uses host resident sensors to monitor such changes. The main objective behind using such profiling system is to optimize the utilization of security provisioning tools by directing only tools with high success ratio. Attacks are somehow targeted, either from the application-objective perspective, or from the technical perspective. CyPhyMASC circulates defense missions while favoring the activation of defense tools targeting attacks that match the host profile. CyPhyMASC do not limit tool activation to only those who match the host profile to cover any unexpected out of profile attacks. Using profiles to guide tool activation minimize the search space, enhance the detection accuracy and promptness and optimize resource usage on the host and on the SSP. The detection mechanism relays on multiple control techniques. Signature based technique is used by the resident unit to maintain minimum level of security at all times. The evolutionary sensing mechanism use heuristic technique to guide and control sensor circulation and to analyze sensor feedback to detect unknown attacks, or to identify maliciously acting components.

Identifying unknown threats

The utilization of heuristic / metaheuristics is necessary to enable attack prediction, and detection of unknown attacks. CyPhyMASC utilizes its ability to isolate the main security provisioning concerns sensing, effecting, and control logic to extract abstract, privacy friendly information regarding running processes. The feedback is safely sent to remote analysis units to apply whatever logic is needed to detect attacks. The selected logic metaheuristics technique determines the type of sensors to be used, execution pattern for such sensors, sample collection protocol. Additionally, that logic is the one responsible for processing the feedback coming for such sensor



```
(mechanism-id) Select-mechanism( last used, weight, severity level)
(Sensor-list, activation-protocol) Activate (mechanism-id, suspicious-item, item-type)
(deployment-package) Call optimization-unit ((Sensor-list, activation-protocol)
(feedback) Call deploy-sensors (deployment-package)
(res) Process-feedback (feedback, mechanism-id)
If res>threshold
fire-alert (item-type, mechanism-id, deployment-package, feedback)
else
if res < Dynamic(10)% threshold
(mechanism-id) Select-next-mechanism( last used, weight, severity level)
Repeat process
Else
Discard event
```

Fig. 2.14 Example of the hubristic mechanism selection procedure

to determine whether the host is safe or not. The process starts when the selection mechanism selects the metaheuristics technique to be used for the next detection round.

Metaheuristics selection mechanism

CyPhyMASC can use single metaheuristics technique or multiple metaheuristics technique at the same time to investigate certain issue. The selection of how many techniques to be used and the type of techniques used depend on the severity of the situation under investigation. The default is the use of only one technique, while if the last utilized technique reported high level of attack certainty that is close, but do not cross the required safety threshold, the system use other techniques to enhance the quality of calculation. In CyPhyMASC, metaheuristics techniques are ordered and have weights assigned to each one. The metaheuristics technique selection mechanism always prefers techniques with highest weight value. When an additional mechanism is needed the next highest value technique is selected. The weights for each technique is not fixed, it is dynamic and gets assigned based on the success or failure of each technique to detect attacks, and the number of false positives or negatives of each technique. This evaluation occurs independently at each management unit within CyPhyMASC framework.

Figure 2.14 presents a simple representation of the metaheuristics technique selection process.

2.4.5.4 Sensor Reuse

CyPhyMASC optimize ToS resource usage by minimizing the number of deployed sensors and effectors on the host. CyPhyMASC utilize it ability to separate the main concerns and the availability of abstract sensing and effecting elements to remove any duplication of sensor deployment requests. The separation between the tool and the control behind it enabled CyPhyMASC to reuse previous sensor feedback within a certain time frame to feed multiple control components.

For example, let us assume that the selection protocol at time (t) selected pattern matching metaheuristics technique to investigate events occurring on host (h1). The investigations focused on suspicious memory behavior of a certain process. The sensors are selected based on that objective, and the sampling protocol is automatically generated using the predetermined syntax. The list and the protocol selected are sent to the optimization unit. The optimization unit checks if there was any previous valid match for that request. If any, it removes it and report the recorded feedback directly to the analysis unit. If not, the list is sent to the reservoir to program and deploy the sensors. The feedback is sent back to be analyzed by the pattern matching algorithm to determine whether there was an attack or not.

The optimization unit applies certain aging policy to determine the validity duration for a certain senor feedback to be reused. The expiration date is dynamic and adjusted automatically based on the ToS host workload, nature, types of application, level of changes, amount of data flowing from and into it, number of application reinstalling, deployment, .. etc. If the requested sensor list contain any sensor that was used before with a valid expiration date and compatible deployment protocol measurement sequence, time, sampling rate, etc. then it will not be deployed again, and the old feedback will be reused. Doing so, is expected to save a considerable amount of host resources.

2.5 Example of CyPhyMASC Security Mission

The task of each CyPhyMASC component is further illustrated through a discussion of one of CyPhyMASCs automatically composed security missions that search for memory behavior deviation within a predetermined timeframe, and the dispersion of such deviation through the ToS host networks.

Goals

1. Detect massive deviation in memory usage
2. Locate the area under suspicion
3. Collect information about processes with suspicious memory usage
4. Identify critical and non-critical processes
5. Resolve the problem.

Tools

1. Memory usage monitoring sensors
2. Processes information collection crawler sensors
3. Process killing effectors.

2.5.1 Detection and Resolution Scenario

This synthetic scenario illustrates the event of using X123 to secure organization ABC. ABC is a large organization composed of multiple enclaves. The following incident happened in enclave E1. On the regular inspection round on E1 hosts, with an active heuristic mechanism X, the feedback collected by the deployed sensing organisms and analyzed by the analysis organism indicated an unidentified strange behavior in host A. The decision-making organism calculated the weights from the score sheets and followed the heuristic rules and the comparison between the calculated weights exceeds the threshold. The decision-making organism sent its guidelines to the role-composer based on the analysis reports with the list of high similarity rules to the reported feedback score-sheet. The role-composer composes new security mission that mixes the sensing part of all the similar rules. X123 was the newly generated mission that was built to investigate possible memory-related behavior deviations within E1. X123 have three main roles played by three organisms, sensing, analysis, and effecting. The sensing organism uses the memory scanning Cells to take multiple snapshots of the memory usage within host A for a certain period. The analysis organism applies some predetermined statistics to evaluate the detected behavior deviation that will be evaluated and compared against certain threshold to determine the next step. Based on the result, further investigations might be needed.

These investigations will be handled by the sensing organism that will deploy processes-information-crawler sensor Cells. Crawlers will collect information about processes with high memory usage. The collected data will be sent to the analysis organism that will generate a comprehensive report to the decision-making organism. The decision-making organism will decide whether to discard the incident, or to activate effector organisms on X123 to resolve the situation.

The decision-making organism might decide to share the mission profile with other SSPs asking for external feedback, or deploy X123 locally for further investigations. Based on the sharing command search-scope and the clearinghouse permissions X123 will be re-deployed. The deployment scope might be limited to only the hosts within enclave E1, allover ABC enclaves, or globally between SSPs searching for similar behavior deviation pattern. If the decision was to activate X123 effecting organisms for quick resolution, X123 will be instructed to kill some of the suspicious non-critical processes and re-evaluate the situation. If the redeployment came-out with multiple incoming alerts for the same memory behavior deviation, the decision-making organism will raise the severity level of the situation indicating global wide spreading attack. Based on that, commands will be issued to the role composer to customize a new containment mission based on the attack reported parameter. Meanwhile CyPhyMASC will be applying resolution effectors of X123. The containment effectors will be deployed over host A, other hosts in communication with host A, and the intermediate communication elements routers, switches, etc. to construct a quarantine area around the malicious host. After successful containment and resolution the whole process will be profiled and stored in the security mission repository for future reference. These profiles will hold details about the containment, and resolution methodology, and all the sensors and effectors API used to compose the used missions. Sharing authorizations and scope of these profiles will depend on the local clearing-house decision.

2.5.2 CyPhyMASC Addressing the BW Attacker Assumptions

In this section we discuss the BW attack design invariants and attacker assumptions that were mentioned in earlier illustrating how CyPhyMASC adequately invalidates them. CyPhyMASC is designed to work in total isolation from the ToS, invalidating assumption (1). CyPhyMASC acts as a buffer between the SSP and ToS. Neither CyPhyMASC nor the SSP share the network or the hosts of the ToS. The security services are delivered to the ToS in an isolated network that connects the ToS to CyPhyMASC. The security delivery vehicles are secured using our Moving-target defense (ChameleonSoft) described in [22], which invalidates assumption (4).

CyPhyMASC is an active defense system founded over a smart foundation managed by our automated management system (CyberX) described in [21] and secured by ChameleonSoft [22]. One of the main tasks of ChameleonSoft as is monitor and secure the COA based foundation against threats and attacks. Having ChameleonSoft and CyberX handling such details waves this workload away from CyPhyMASC giving it more space to focus on provisioning security services to the ToS. Additionally, CyPhyMASC is designed to support large scale systems and computationally expensive tasks. CyPhyMASC design supports distributing the tasks over a hierarchy of independent management entities composed of fine grained components managed by CyberX. The fine granularity of such components and the isolation between its logic, data, and physical resources enabled CyberX to fractionize large tasks over multiple hosts constructing a cloud like platform with virtually infinite resources. With that unique feature, CyPhyMASC invalidates assumption (3).

CyPhyMASC uses signature based detection tools as a part of its arsenal, while the major part of that arsenal relay on an evolutionary sensory system. CyPhyMASC utilizes multiple intelligent mechanisms to detect unknown attacks based on monitoring suspicious activities and up normal behavior. Further, CyPhyMASC is not a commercial product available for conventional users. Even though, the foundation of CyPhyMASC is highly dynamic and autonomous inducing high level of dissimilarity between identical copies of the same system, and that invalidates assumption (2).

One of the main objectives of CyPhyMASC is to promote the security system situational awareness of the different ToS components and to isolate the platform composition heterogeneity enabling seamless security provisioning. CyPhyMASC pervasive monitoring and analysis, and the intrinsic trustworthy sharing and cooperative security enhance the situational awareness all ToS components. CyPhyMASC collect events from different entities of the ToS, correlate the collected information to generate a global image of the entire system to be analyzed by high management units. Doing so, enables CyPhyMASC to detect slow moving attacks, and attacks using remote bots to lunch attacks on remote hosts. The aforementioned aspects successfully invalidate assumptions (3 and 5). By invalidating all the attacker assumption, we surmise that it is virtually impossible for the BW attack to succeed in attacking an CyPhyMASC protected system.

2.6 Conclusion

In this chapter, we presented CyPhyMASC whose unique construction and functionalities provide pervasive, seamless and scalable MASC services to enable interoperable and dynamic security employing situation-appropriate tools; early failure/attack detection and resolution; and trustworthy scalable cooperative security amongst heterogeneously composed and interacting targets like CPS. CyPhyMASC is built upon our novel Cell-oriented architecture and acts as a trustworthy elastic intelligent middle layer uniformly interfacing SSPs and ToSs. There are several interesting challenges that remain to be addressed. These include formalizing the sensing and effecting autonomous management and mission generation framework; autonomous inspection and profiling of security vaccines; correlating sensors feedback into comprehensive real-time global views; and optimally adjusting sensors circulation based on the dynamic changes of these views.

References

1. Knight, J.C., Leveson, N.G.: An experimental evaluation of the assumption of independence in multiversion programming. IEEE Trans. Softw. Eng. **12**(1), 96–109 (1986)
2. Jorstad, N., Landgrave, T.S.: Cryptographic algorithm metrics. In: 20th National Information Systems Security Conference, Baltimore (1997)
3. Parra, C.: Towards Dynamic Software Product Lines: Unifying Design and Runtime Adaptation. Universit Lille (2011)
4. Stberg, P.-O., Elmroth, E.: GJMF - A Composable Service-Oriented Grid Job Management Framework (2010). http://www.cs.umu.se/ds
5. Chen, X., Andersen, J., Mao, Z.M., Bailey, M., Nazario, J.: Towards an understanding of anti-virtualization and anti-debugging behavior in modern malware. In: International Conference on Dependable Systems and Networks (2008)
6. Sze, S., Tiong, W.: A Comparison between Heuristic and MetaHeuristic Methods for Solving the Multiple Traveling Salesman Problem. World Academy of Science, Engineering and Technology (2007)
7. Podgrski, W.: Artificial Intelligence Methods in Virus Detection & Recognition—Introduction to Heuristic Scanning (2012). http://podgorski.wordpress.com
8. Haack., J., Fink, G., Fulp, E., Maiden, W.: Cooperative infrastructure defense. In: Workshop on Visualization for Computer Security (VizSec) (2008)
9. Maiden, W.M.: DualTrust, A Trust Management Model for Swarm-Based Autonomic Computing Systems. Washington State University (2010)
10. Maiden, W.M., Dionysiou, I., Frincke, D.A., Fink, G.A., Bakken, D.E.: DualTrust: A Distributed Trust Model for Swarm-Based Autonomic Computing Systems. Data Privacy Management and Autonomous Spontaneous Security (2010)

11. Lee, Y.: A Pre-kernel agent platform for security assurance. In: IEEE Symposium on Intelligent Agent (IA) (2011)
12. Nguyen-Tuong, A., Wang, A., Hiser, J., Knight, J., Davidson, J.: On the effectiveness of the metamorphic shield. In: The Fourth European Conference on Software Architecture ECSA 10, pp. 170–174 (2010)
13. Abraham, A., Jain, R., Thomas, J., Han, S.Y.: D-SCIDS: distributed soft computing intrusion detection system. J. Netw. Comput. Appl. **30**(1), 81–98 (2007)
14. Wu, S., Banzhaf, W.: The use of computational intelligence in intrusion detection systems: A review. Appl. Soft Comput. **10**(1), 1–35 (2010)
15. Mukherjee, S.: FPGA based Network Security Architecture for High Speed Networks, MTech (2001)
16. Otey, M., Parthasarathy, S., Ghoting, A., Li, G., Narravula, S., Panda, D.: Towards nic based intrusion detection. In: The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 723–728 (2003)
17. Santos, I., Penya, Y., Devesa, J., Bringas, P.: N-Grams-based file signatures for malware detection. In: The 11 International Conference on Enterprise Information Systems (ICEIS) (2009)
18. Lemos, R.: White House Network Attack Highlights Need for Stronger Defenses (2012). http://www.eweek.com/security/white-house-network-attack-highlights-need-for-stronger-defenses/
19. Prayurachatuporn, S., Benedicenti, L.: Increasing the reliability of control systems with agent technology. ACM SIGAPP Applied Computing (2001)
20. Box, D.: Essential COM. Addison Wesley, Reading Mass (1998)
21. Knight, J.C., Davidson, J.W., Evans, D., Nguyen-Tuong, A., Wang, C.: Genesis: A Framework for Achieving Software Component Diversity. Technical Report AFRL-IF-RS-TR-2007-9, University of Virginia (2007)
22. Te-Shun, C., Sharon, F., Wei, Z., Jeffrey, F., Asad, D.: Intrusion aware system-on-a-chip design with uncertainty classification. In: The 2008 International Conference on Embedded Software and Systems-ICESS (2008)
23. Azab, M., Hassan, R., Eltoweissy, M.: ChameleonSoft: a moving target defense system. In: 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, (CollaborateCom 11) (2011)
24. Spinellis, D.: Reliable identification of bounded-length viruses is NP-complete. IEEE Trans. Inf. Theory **49**(1), 280–284 (2003)

Chapter 3
An Optimized Approach for Medical Image Watermarking

Mona M. Soliman, Aboul Ella Hassanien and Hoda M. Onsi

Abstract Digital radiological modalities in modern hospitals have led to the producing a variety of a vast amount of digital medical files. Therefore, for the medical image, the authenticity needs to ensure the image belongs to the correct patient, the integrity check to ensure the image has not been modified, and safe transfer are very big challenges. Digital watermarking is being used in broadcast and Internet monitoring, forensic tracking, copy protection, counterfeit deterrence, authentication, copyright communication and e-commerce applications. The basic idea behind digital watermarking is to embed a watermark signal into the host data with the purpose of copyright protection, access control, broadcast monitoring etc. Improvements in performance of watermarking schemes can be obtained by several methods. One way is to make use of computational intelligence techniques by considering image watermarking problem as an optimization problem. Particle swarm optimization is a relatively simple optimization technique, and it is easier to be understood compared with some other evolutionary computation methods. It is widely used in different fields including watermarking technologies. The global convergence of PSO cannot always be guaranteed because the diversity of population is decreased with evolution developed. To deal with this problem, concept of a global convergence guaranteed method called as Quantum behaved Particle Swarm Optimization (QPSO) was developed. Weighted QPSO (WQPSO) is introduced as an improved quantum-behaved particle swarm optimization algorithm. In this chapter we present a secure patient medical images and authentication scheme which enhances the security, confidentiality and integrity of medical images transmitted through the Internet. This chapter proposes a watermarking by invoking par-

M. M. Soliman (✉) · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Faculty of Computers and Information,
Cairo University, Cairo, Egypt
e-mail: monasolyman_it@yahoo.com
URL: http://www.egyptscience.net

H. M. Onsi
Faculty of Computers and Information, Cairo University, Cairo, Egypt

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_3, © Springer-Verlag Berlin Heidelberg 2014

ticle swarm optimization with its modifications(PSO-QPSO-WQPSO) technique in adaptive quantization index modulation and singular value decomposition in conjunction with discrete wavelet transform (DWT) and discrete cosine transform (DCT). The experimental results show that the proposed algorithm yields a watermark which is invisible to human eyes, robust against a wide variety of common attacks and reliable enough for tracing colluders.

3.1 Introduction

The development of multimedia and communication technologies has made medical images act important roles in the field of telediagnosis, telesurgery and so on. At the same time such advances provide new means to share, handle and process medical images, it also increases security issues in terms of: confidentiality, availability and integrity.

Since any person with privilege can access images which are contained in database and can modify them maliciously, the integrity of the images must be protected by using watermarking, which is called integrity watermark. Meanwhile, web-based image database system contains valuable medical image resources for not only research purpose but also commercial purpose. Therefore the copyright and intellectual property of the database should be also protected by a watermark, which is called copyright watermark.

The basic principle of watermarking methods [1–4] is to add copyright information into the original data, by embedding it into the original image. Then if the image is modified in any sense, it can be detected with the watermark. Initially, the watermark could be simply a unique number, such as the patients insurance code but as research moves into new paths, a new role has been given to the watermark: to include (apart from hospital digital signatures or copyright information), the electronic patient record, digital documents with diagnosis, blood test profiles or an electrocardiogram signal. By embedding these files into the original image we increase authenticity, confidentiality of patient data and of the accompanying medical documents, availability, and reduce the overall file size of the patients' records.

There are several types of algorithms for watermarking. Each type of algorithms has its own advantages and limitations. No method can provide fully perfect solution. Each type of solution has robustness to some type of attacks but is less resilient to some other types of attacks. Watermarking techniques can be categorized in different ways [5]. They can be classified according to the type of watermark being used, Watermark in digital watermarking can be either visible or invisible. A visible watermark is a watermark that is visible on the watermarked image. An invisible watermark is a watermark that is not visible on the watermarked image. An invisible watermark is used in applications such as copyright protection systems to prevent unauthorized copying of digital media, or steganography to communicate a hidden message embedded in the digital signal. According to their robustness against attacks, watermarking techniques may be classified as fragile, semi-fragile, and robust. Finally, watermarking systems may be classified according to the embedding domain of the cover media to the spatial domain methods [6–8] and the transform domain methods [9–11].

Singular Value Decomposition (SVD) is one of the most powerful numeric analysis techniques with numerous applications including watermarking, therefore many watermarking schemes that are combine different transforms with SVD have been proposed lately. Recently, many researchers focus on adaptive determination of the quantization parameters for SVD-based watermarking. They consider solving this problem using adopting artificial intelligence techniques or analysing statistical model of each block in the image [12–14].

Improvements in performance of watermarking schemes can be obtained by several methods. One way is to make use of Intelligent Computing techniques by considering image watermarking problem as an optimization problem [5].

In mathematics, or mathematical programming, optimization, refers to choosing the best element from some set of available alternatives. In the simplest case, this means solving problems in which one seeks to minimize or maximize a real function by systematically choosing the values of real or integer variables from within an allowed set. This formulation, using a scalar, real-valued objective function, is probably the simplest example; the generalization of optimization theory and techniques to other formulations comprises a large area of applied mathematics. More generally, it means finding “best available” values of some objective function given a defined domain, including a variety of different types of objective functions and different types of domains. Several mathematical models of optimization take as a starting point biological behaviors and make an abundant use of metaphors and terms originating from genetics, ethology, and even from ethnology or psychology.

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization problems that originally took its inspiration from the biological examples by swarming, flocking and herding phenomena in vertebrates. Particle Swarm Optimization (PSO) incorporates swarming behaviors observed in flocks of birds, schools of fish, or swarms of bees, and even human social behavior.

Although classical PSO converges to the global solution, still for some problems it is not a global optimization technique, since it gets trapped in local minima. Classical PSO has many control parameters. The convergence of the algorithm depends on the value of the control parameters. Tuning a proper value for convergence of PSO algorithm is a tedious work. To avoid this problem a new PSO, which has only one control parameter and in which the movement of particles are inspired by the quantum mechanics is proposed and known as Quantum behaved Particle Swarm Optimization (QPSO) [5]. It provides a good scheme for improving the convergence performance of PSO because it is a theoretically global convergence algorithm [15]. In recent years, some other works such as [16–19] are presented to improve the performance of QPSO. Weighted QPSO (WQPSO) is introduced in [20] as an improved quantum-behaved particle swarm optimization algorithm, in order to balance the global and local searching abilities, a weight parameter in calculating the mean best position in QPSO is introduced to render the importance of particles in population when they are evolving.

In this chapter, a novel SVD watermarking approach based on Swarm Optimization is presented with focusing on adaptive determination of the quantization parameters for singular value decomposition. Two quantization parameters are used: one quantization parameter is determined by exploiting the characteristics of Human Visual System (HVS) [20], the other quantization parameter is optimized through different versions of particle swarm optimization algorithms (Classical PSO, Quantum PSO, Weighted Quantum PSO). These two quantization parameters are combined for ensuring the final adaptive quantization steps are optimal for all embedding blocks and reaching better trade off between the imperceptibility and robustness of the digital watermarking system.

3.2 Preliminaries

3.2.1 Swarm Intelligent Optimization

Particle Swarm Optimization (PSO) was introduced for continuous optimization in the mid-1990s, inspired by bird flocking. Swarm intelligence methods have been very successful in the area of optimization, which is of great importance for industry and science.

PSO is a global optimization algorithm for dealing with problems in which a best solution can be represented as a point or surface in an n-dimensional space. PSO simulates the behaviors of bird flocking. In PSO, each single solution is a "bird" in the search space. We call it "particle". All of particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. PSO is initialized with a group of random particles (solutions) and then searches for optima by updating generations. In every iteration, each particle is updated by following two "best" values. The first one is the best solution (fitness) it has achieved so far. (The fitness value is also stored). This value is called pbest. Another "best" value that is tracked by the particle swarm optimizer is the best value, obtained so far by any particle in the population. This best value is a global best and called gbest. When a particle takes part of the population as its topological neighbors, the best value is a local best and is called lbest (Kennedy).

3.2.1.1 Classical Particle Swarm

The concept of particle swarms, although initially introduced for simulating human social behaviors, has become very popular these days as an efficient search and optimization technique. Particle swarm optimization (PSO) [21], does not require any gradient information of the function to be optimized, uses only primitive mathematical operators and is conceptually very simple. PSO has attracted the attention of a lot of researchers resulting into a large number of variants of the basic algorithm as well as many parameter automation strategies. The canonical PSO model consists of a swarm of particles, which are initialized with a population of random candidate solutions [22]. They move iteratively through the d-dimension problem space to search the new solutions, where the fitness, f, can be calculated as the certain qualities measure. Each particle has a position represented by a position-vector $\boldsymbol{x}_i$ (i is the index of the particle), and a velocity represented by a velocity-vector $\boldsymbol{v}_i$. Each particle remembers its own best position so far in a vector $\boldsymbol{x}_i^{\#}$, and its j-th dimensional value is $x_{ij}^{\#}$. The best position-vector among the swarm so far is then stored in a vector $\boldsymbol{x}^*$, and its j-th dimensional value is x_j^*. During the iteration time t, the update of the velocity from the previous velocity to the new velocity is determined by Eq. (3.1). The new position is then determined by the sum of the previous position and the new velocity by Eq. (3.2).

$$v_{ij}(t+1) = \begin{cases} wv_{ij}(t) + c_1 r_1 (x_{ij}^{\#}(t) - x_{ij}(t)) \\ + c_2 r_2 (x_j^*(t) - x_{ij}(t)) \end{cases} \tag{3.1}$$

$$x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1). \tag{3.2}$$

where w is called as the inertia factor which governs how much the pervious velocity should be retained from the previous time step, r_1 and r_2 are the random numbers, which are used to maintain the diversity of the population, and are uniformly distributed in the interval [0,1] for the j-th dimension of the i-th particle. c_1 is a positive constant, called as coefficient of the self-recognition component, c_2 is a positive constant, called as coefficient of the social component. From Eq. (3.1), a particle decides where to move next, considering its own experience, which is the memory of its best past position, and the experience of its most successful particle in the swarm. In the particle swarm model, the particle searches the solutions in the problem space with a range $[-s, s]$ (if the range is not symmetrical, it can be translated to the corresponding symmetrical range). In order to guide the particles effectively in the search space, the maximum moving distance during one iteration must be clamped in between the maximum velocity $[-v_{max}, v_{max}]$ given in Eq. (3.3):

$$v_{ij} = sign(v_{ij}) min(\left|v_{ij}\right|, v_{max}). \tag{3.3}$$

The value of v_{max} is $p \times s$, with $0.1 \leq p \leq 1.0$ and is usually chosen to be s, i.e. $p = 1$. The end criteria are usually one of the following: maximum number of iterations, number of iterations without improvement, or minimum objective function error.

Although classical PSO converges to the global solution, still for some problems it is not a global optimization technique, since it gets trapped in local minima. Classical PSO has many control parameters. The convergence of the algorithm depends on the value of the control parameters. Tuning a proper value for convergence of PSO algorithm is a tedious work. To avoid this problem a new PSO, which has only one control parameter and in which the movement of particles are inspired by the quantum mechanics is proposed [5].

3.2.1.2 Quantum Particle Swarm Optimization

Classical physics is dominated by two fundamental concepts. The first is the concept of a particle, a discrete entity with definite position and momentum which moves in accordance with Newton's laws of motion. The second is the concept of an electromagnetic wave, an extended physical entity with a presence at every point in space that is provided by electric and magnetic fields which change in accordance with Maxwell's laws of electromagnetism. The classical world picture is neat and tidy: the laws of particle motion account for the material world around us and the laws of electromagnetic fields account for the light waves which illuminate this world. This classical picture began to crumble in 1900 when Max Planck published a theory of black-body radiation; i.e. a theory of thermal radiation in equilibrium with a perfectly absorbing body [14]. Schrodinger's wave mechanics is now the backbone of our current conceptional understanding and our mathematical procedures for the study of quantum phenomena [12]. The role of the Schrodinger equation in quantum mechanics is analogous to that of Newton's Laws in classical mechanics. Schrodinger equation is a partial differential equation which describes how the wave function representing a quantum particle ebbs and flows.

In terms of classical mechanics, a particle is depicted by its position vector x_i and velocity vector v_i, which determine the trajectory of the particle. The particle moves along a determined trajectory in Newtonian mechanics, but this is not the case in quantum mechanics. In quantum world, the term trajectory is meaningless, because x_i and v_i of a particle can not be determined simultaneously according to uncertainty principle. Therefore, if individual particles in a PSO system have quantum behavior, the PSO algorithm is bound to work in a different fashion [13].

Trajectory analysis demonstrated that, to guarantee convergence of PSO algorithm, each particle must converge to its local attractor $p_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,n})$ of which the coordinates are defined as:

$$p_{i,j}(t) = (c_1 p_{i,j}(t) + c_2 p_{g,j}(t))/(c_1 + c_2) j = 1, 2, \ldots .n \tag{3.4}$$

where c_1 and c_2 are uniformly distributed random numbers in the interval [0,1]. It can be seen that the local attractor is a stochastic attractor of particle i that lies in a hyper-rectangle with P_i and P_g being two ends of its diagonal. In [20], the concepts of QPSO is defined as follows. Assume that each individual particle move in the search space with a δ potential on each dimension, of which the center is the point p_{ij}. For simplicity, we consider a particle in one-dimensional space, with point p the center of potential. Solving Schrodinger equation of one-dimensional δ potential well, we can get the probability density function Q and distribution function F as follows:

$$Q(X_{i,j}(t+1)) = (1/L_{i,j}(t))F(X_{i,j}(t+1)) \tag{3.5}$$

$$F(X_{i,j}(t+1)) = \exp^{-2|p_{i,j}(t)-X_{i,j}(t+1)|/L_{i,j}(t)} \tag{3.6}$$

where $L_{i,j}$ is standard deviation of the distribution, which determines search scope of each particle. Employing Monte Carlo method for, we can obtain the position of the particle using following equation [20]:

$$X_{i,j}(t+1) = p_{i,j} \pm (L_{i,j}/2)\ln(1/u) \tag{3.7}$$

where u is a random number uniformly distributed in (0,1). To evaluate $L_{i,j}(t)$, a global point called Mainstream Thought or mean best position of the population is introduced into classical PSO. The global point, denoted as m, is defined as the mean of the *pbest* positions of all particles. That is:

$$mbest = \frac{1}{M}\sum_{i=1}^{M} p_{i1}, \frac{1}{M}\sum_{i=1}^{M} p_{i2}, \ldots.. \frac{1}{M}\sum_{i=1}^{M} p_{id} \tag{3.8}$$

The values of $L_{i,j}(t)$ is determined by the following equation:

$$L_{i,j}(t) = 2\beta.|m_j(t) - X_{i,j}(t)| \tag{3.9}$$

and thus the position can be calculated by the following equation:

$$X_{i,j}(t+1) = p_{i,j}(t) \pm \beta.|m_j(t) - X_{i,j}(t)|.ln(1/u) \tag{3.10}$$

The parameter β is called contraction-expansion coefficient , which can be tuned to control the convergence speed of algorithms, β is the only parameter in QPSO algorithm. The main steps of QPSO is illustrated in Algorithm 1.1.

3.2.1.3 Weighted Quantum Particle Swarm Optimization

From Eq. (3.8), we can see that the *mbest* is simply the average on the personal best position of all particles, which means that each particle is considered equal and exert the same influence on the value of *mbest*. The equally weighted mean position, however, is something of paradox, compared with the evolution of social culture in real world. For one thing, although the whole social organism determines the Mainstream Thought, it is not proper to consider each member equal. In fact, the elitists play more important role in culture development. With this in mind a new control method for the QPSO is proposed in [20], where *mbest* in Eq. (3.8) is replaced by a weighted mean best position. The most important problem is how to evaluate

Algorithm 1.1 Quantum particle swarm optimization (QPSO) algorithm

```
Step 1: initialize population:(randomly assign positions for every particle using random uniform distribution)
for maximum number of iterations do
  Step 2: calculate mbest using Eq. 3.8
  Step 3: calculate fitness of each particle
  Step 4: update the pbest positions
  if particles fitness < pbest then
    set pbest = current value
    set pbest location = current location
  end if
  Step 4: update the gbest position
  for particle population do
    if particles fitness < gbest then
      set gbest = current value
      set gbest location = current location
    end if
  end for
  Step 5: update particle position according to Eq. 3.10
end for
```

Algorithm 1.2 Weighted quantum particle swarm optimization

```
Initialize population
(randomly assign positions for every particle using random uniform distribution)
for maximum number of iterations do
  Calculate mbest using Eq. (3.11)
  Calculate fitness of each particle
  Update the pbest positions
  if particles fitness < pbest then
    pbest = current value
    pbest location = current location
  end if
  Update the gbest position
  for Particle population do
    if particles fitness < gbest then
      gbest = current value
      gbest location = current location
    end if
  end for
  Update particle position according to Eq. (3.10)
end for
```

particle importance in calculate the value of *mbest*. Based on this we associate elitism with the particles' fitness value.

Describing it formally, we can rank the particle in descendent order according to their fitness value first, then assign each particle a weight coefficient α_i linearly decreasing with the particle's rank, that is, the nearer the best solution, the larger its weight coefficient. The mean best position *mbest* is calculated using Eq. (3.11):

$$mbest = \frac{1}{M}\sum_{i=1}^{M}\alpha_{ij} * p_{i,j} \tag{3.11}$$

where j = 1,2,…,n and α_{ij} is the weighted coefficient.

3.2.1.4 DWT-SVD Based Watermarking

Wavelet transform is obtained by calculating inner product of data to be transformed, with the translated and scaled mother wavelet function chosen for transform. Value of an inner product is express the resemblance of the mother wavelet in a certain translation and scaling state, with the data. Product result of each translation and scaling state is named as wavelet coefficient of certain translation and scaling state. DWT process is realized by filtering data that will be transformed, by various low pass and high pass filters and down scaling the results. Decomposition level is the unit (number) of the above process. Every decomposition level forms four band data called *LL*, *HL*, *LH* and *HH* sub-bands. This process can be continued up to reach desired level [23].

In past years, a singular value decomposition SVD-based watermarking technique and its variations have been proposed [23–28]. The core idea behind these approaches is to find the SVD of the cover image or each block of the cover image, and then modify the singular values to embed the watermark. There are two main properties to employ SVD method in digital watermarking scheme:

- The singular values of an image have very good stability, that is, when a small perturbation is added to an image, its singular values do not change significantly.
- Singular values represent intrinsic algebraic image properties.

In general, the watermark can be scaled by a scaling factor SF which is used to control the strength of the watermark. It is found that the scaling factor is set to be constant in some SVD-based studies. However, many argued that considering a single and constant scaling factor may not be applicable [25]. The larger the SF, the more the distortion of the quality of the host image (transparency) and the stronger the robustness. On the other hand, the smaller the SF, the better the image quality and the weaker the robustness [24]. SVD decomposes an $M \times N$ real matrix A into a product of three matrices $A = USV^T$, where U and V^T are $M \times N$ and $N \times N$ orthogonal matrices, respectively. S is an $N \times N$ diagonal matrix. The elements of S are only nonzero on the diagonal and are called the SVs of A.

3.2.1.5 Medical Image Watermarking

The healthcare professionals use the Internet for transmitting or receiving Electronic Patient Records (EPR) via e-mail. An EPR typically contains the health history of a patient, including X-ray images, CT-Scan images, physical examinations report, laboratory tests, treatment procedures, prescriptions, radiology examinations etc. An EPR can be represented in various forms such as diagnostic reports, images, vital signals, etc. An EPR transmitted through the Internet is very important since it contains the medical information of a person in digital format [29]. Therefore, for the medical image, the authenticity needs to ensure the image belongs to the correct patient, the integrity check to ensure the image has not been modified, and safe transfer are very big challenges. Also, when a digital medical image is opened for diagnosis, it is important that an automated framework exists to verify the authenticity and integrity of the image itself. Hospital Information System (HIS) and Picture Archiving and Communication System (PACS) have been established to provide security solutions to ensure confidentiality, integrity, and authentication [30]. Security of medical images, derived from strict ethics and legislative rules, gives rights to the patient and duties to the health professionals. In medical applications, it is very important to preserve the diagnostic value of images. For instance, artifacts in a patient's diagnostic image due to image watermarking may cause errors in diagnosis and treatment, which may lead to possible life-threatening consequences [29]. For this restrictions embedding watermarks and image compressions must not distort and degrade the quality of images. Therefore, minimum Peak Signal to Noise Ratio (PSNR) of 40–50 db is advised by previous works. More importantly, watermarks should survive the standard image processing like low pass filtering (LPF) which removes noise and improves visual quality; and High Pass Filtering (HPF) that enhances the information content [31].

3.3 The Proposed Medical Watermarking Approach

The design of an optimal watermarking for a given application always involves a trade-off between robustness and imperceptibility. Therefore, image watermarking can be considered as an optimization problem. This work aim to embed watermarking data using one of bio-inspired optimization techniques particle swarm optimization.

This section introduces the proposed adaptive watermarking approach in details including the embedding and extracting procedures. The proposed watermarking approach is designed in such a way that it achieve the imperceptibility and robustness requirements by invoking particle Swarm Optimization (PSO) technique with different improvement (QPSO, WQPSO). These methods are used in designing an adaptive quantization index modulation and singular value decomposition in conjunction with discrete wavelet transform (DWT) and discrete cosine transform (DCT).

3.3.1 Watermark Embedding Procedure

The watermark can be embedded into the host image through three consecutive phases, which are elaborated in this subsection. These phases are transformation, quantization and embedding phases. First, DWT is performed on the host image. Second, the low performed on DCT. Then a set of final quantization steps are modeled both the characteristics of the DCT domain human visual masking and particle swarm optimization of each block to ensure a high perceptual quality of watermarked image and a low bit error rate of the detected watermark. Finally, watermark is embedded into the singular values vector of each block by adaptive and optimized quantization steps. PSO helps search proper basic step of each block in order to optimize watermark embedding process. It is a difficult for determining the proper values of multiple basic quantization steps. In some cases, the choice of them may be based on some general assumption. Therefore, an efficient and optimal algorithm is required for achieving both invisibility and robustness. Here we use PSO with its variations (QPSO-WQPSO) to automatically determine these values without making any assumption.

3.3.1.1 Transformation Phase

The host image I_0 with $m \times n$ dimension is transformed into the wavelet domain using Discreet Wavelet Transform (DWT); three levels wavelet with filters of length 4 is used. We perform the Lth level discrete wavelet decomposition of the host image to produce a sequence of low frequency subband LL, and three high frequency subbands HL, LH, and HH, corresponding to the horizontal, vertical and diagonal details at each of the L resolution levels, and a gross approximation of the image at the coarsest resolution level. By taking the advantage of low frequency coefficients which have a higher energy value and robustness against various signal processing, we segment the *LL* subband into non-overlapping blocks A_i of Size $w \times w$. Then perform the Discreet Cosin Transform (DCT) on the low frequency coefficient *LL* for each block, followed by computation of the singular values vector of each frequency coefficient block A_i by SVD according to the Eq. 3.12.

$$DCT(A_i) = U_i S_i V_i^T \tag{3.12}$$

where $S_i = (\sigma_{i1}, \sigma_{i2}, ..., \sigma_{iw})$, denotes a vector formed by the singular values of the frequency coefficient block A_i. The N_i^s is then computed that represents each block by one value, and each block is quantized to a proper quantization step δ_i according two Eqs. 3.13 and 3.14.

$$N_i^s = \|s_i\| + 1 \tag{3.13}$$

$$N_i = \lfloor \frac{N_i^s}{\delta_i} \rfloor i = 1, 2, \ldots M \tag{3.14}$$

3.3.1.2 Quantization Phase

A set of final quantization steps are modelled both the characteristics of the DCT domain human visual system and weighted quantum particle swarm optimization of each block to ensure a high perceptual quality of watermarked image and a low bit error rate of the detected watermark. This final quantization step is determined as a combination of two quantization steps one come from Human Visual System (HVS) by using luminance mask M_i^L and texture mask M_i^T. The other quantization step determined by using PSO/QPSO/WQPSO training (Algorithm 1.1 and 1.2). When PSO/QPSO/WQPSO algorithm converge, it gives a basic quantization step δ_{i0} that used to formulate the final quantization step δ_i using Eq. 3.15.

$$\delta_i = \lfloor log2^{M_i^L \times M_i^T} \times 1000 \rfloor / 1000 + \delta_{i0} \tag{3.15}$$

Particle swarm optimization algorithms measure the fitness of each particle in each iteration. We have to select a fitness function that reflects both the imperceptibility and robustness. We adopt the same fitness function used in [32]:

$$f_i = \lfloor \frac{m}{\sum_{i=1}^{m} NC_i(w_i', w)} - NC(I_0', I_0) \rfloor^{-1} \tag{3.16}$$

where $NC(I_0', I_0)$ denotes 2-D normalized correlation between original and watermarked image, $NC_i(w_i', w)$ denotes 2-D normalized correlation between original and extracted watermark, and m represents number of attacking methods. Algorithm 1.2 shows the details steps of the quantization using weighted quantum particle swarm optimization.

Figure 3.1 shown in detailed how PSO-QPSO-WQPSO training algorithm used in formulating watermark image using optimization quantization step.

3.3.1.3 Watermark Embedding Phase

Before inserting watermark bits we apply a scrambling mechanism on original watermark according to scrambling technique mentioned in [8]. The modified watermark is embedded into the singular values vector of each block by adaptive and optimized quantization steps according to Eq. 3.17.

$$N_{iw} = \begin{cases} N_i + 1, & \text{if } (mod(N, 2), W_i) = (1, 1) \text{ } or \text{ } (0, 0) \\ N_i, & \text{otherwise} \end{cases} \tag{3.17}$$

Now we have to compute the modified singular value that will hold watermark information.

$$(\sigma_{i1}, \sigma_{i2}, ..., \sigma_{iw}) = (\sigma_{i1}, \sigma_{i2}, ..., \sigma_{iw}) \times (\frac{N_{iw}^s}{N_i^s}) \tag{3.18}$$

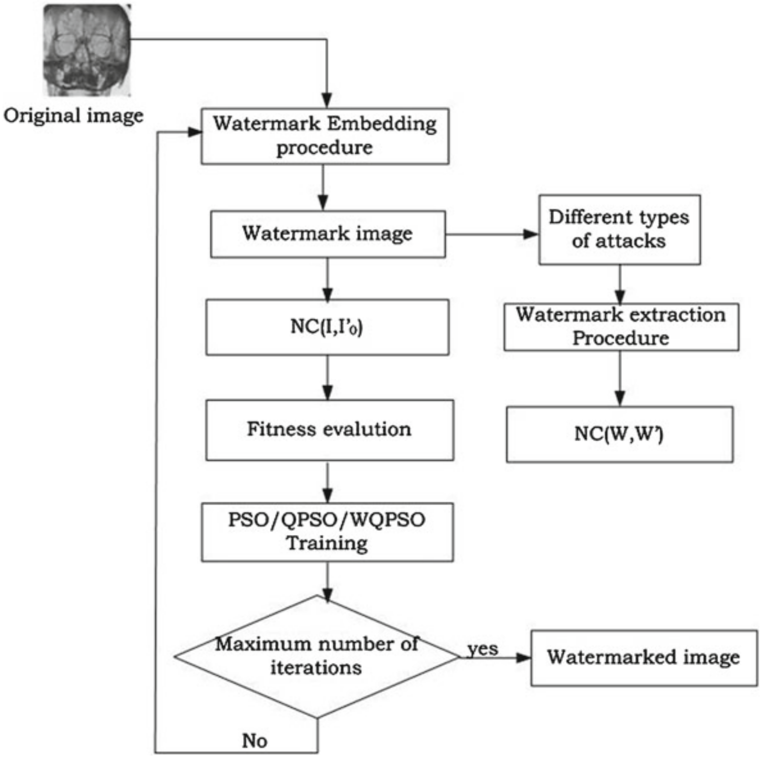

Original image
Watermark Embedding procedure
Watermark image
Different types of attacks
NC(I,I'₀)
Watermark extraction Procedure
Fitness evalution
NC(W,W')
PSO/QPSO/WQPSO Training
Maximum number of iterations
yes
Watermarked image
No

Fig. 3.1 Training of PSO-QPSO-WQPSO algorithms

where $N_{iw}^{s} = \sigma_i \times N_{iw} + 0.5$.

Finally, the watermarked block A_i' is computed with modified singular values. The watermarked low frequency sub-band LL is reshaped through A_i' performed on Inverse Discrete Cosine Transform (IDCT), then the watermarked image I_0' is obtained utilizing Inverse Discrete Wavelet Transform (IDWT).

3.3.2 Watermark Extraction Procedure

The watermark extracting procedure goes through the same steps of the embedding algorithm except that now we have an optimal final quantization steps which is derived during the embedding procedure. Watermark bits are extracted according to the following equation:

$$W_i' = \begin{cases} 1, & \text{if } mod(N_i', 2) = 0; \\ 0, & \text{else } mod(N_i', 2) = 1. \end{cases} \tag{3.19}$$

where

$$N_i = \lfloor \frac{N_i^s}{\delta_{final_i}} \rfloor \quad i = 1, 2, ..., M \tag{3.20}$$

3.4 Experimental Results and Analysis

In order to resist the normal signal processing and other different attacks, we wish the quantization step to be as high as possible. However, because the watermark directly affects the host image, it is obvious that the higher the quantization step, the lower the quality of the watermarked image will be. In other words, the robustness and the imperceptibility of the watermark are contradictory to each other.

In this section, some experimental results are demonstrated to show the effectiveness of the proposed adaptive digital watermarking approach in achieving the two contradictory requirements of watermarking system (robustness and the imperceptibility). We use the Peak Signal to Noise Ratio (PSNR) as an efficient measure of visual fidelity between the host image and the watermarked image, This is give us an indication of the imperceptibility factor. To investigate the robustness of watermark schemes, each watermarked image is attacked using different image processing attacks. To investigate the robustness of watermark schemes, each watermarked image is attacked using JPEG compression, Gaussian noise, Salt and Pepper noises, Gaussian filter, median filter, and geometrical attacks like image cropping and scaling. The watermarking scheme should be robust to signal processing attacks. Normalized Correlation (*NC*) is used for the evaluating the robustness of the watermarking scheme.

The results of various experimental analysis of proposed watermark approach using PSO, QPSO, WQPSO are compared with each others, all in terms of PSNR and NC values.

In all the experiments, different parameters are required to be initialized first in order to start training procedure of PSO, QPSO, and WQPSO. Table 3.1 shows such different parameters value.

It is very common and usual for physicians to participate in technical group communication existed between physicians and hospitals in order to diagnosing and spotting the patient's problem. There are several recipients and senders in this scenario. Transmitting Electronic Health Record (EHR) containing images and databases in these groups requires network bandwidth consumption management in order to ensure patient's privacy protection. In medical applications, it is very important to preserve the diagnostic value of images. Therefore, embedding watermarks must not distort and degrade the quality of images. Therefore, minimum peak signal to noise ratio (PSNR) of 40–50 dB is advised in previous works [31]. More importantly, watermarks should survive the standard image processing attacks like JPEG compression, Gaussian noise, Salt and Pepper noises, Gaussian filter, median filter, and geometrical attacks like image cropping and scaling.

Table 3.1 PSO, QPSO and WQPSO parameters setting

Method	Parameters	Value
PSO	Swarm population	20
	Number of generations	50
	Self-recognition component c_1	1.2
	Social component c_2	1.8
	Inertial weight w	Linearly decreasing from 0.9 to 0.4
QPSO	Swarm population	20
	Number of generations	50
	Contraction-expansion coefficient β	Linearly decreasing from 0.5 to 0.1
WQPSO	Swarm population	20
	Number of generations	50
	β coefficient	Linearly decreasing from 0.5 to 0.1
	α_i coefficient	Linearly decreasing from 1.5 to 0.5

In this simulation experiments, we assume a group of several medical professionals, which are x-ray images of size 512 × 512. The length of the watermark is 32 × 32 binary bits set.

3.4.1 Evaluation Analysis

The results of watermark embedding procedure on medical images are depicted in Fig. 3.2. Figures show the original and three watermarked images using PSO, QPSO, and WQPSO respectively. One can simply see that the watermarked images have good visual fidelity. It is hard to distinguish visually between original image and watermarked image.

The Peak Signal to Noise Ratio (PSNR) values used for quality comparison between the original and the watermarked images are utilized in Table 3.2. It can be seen that PSNR values in case of WQPSO is higher than PSNR for both QPSO and classical PSO which means the visual fidelity of WQPSO is superior to the two other methods.

Tables 3.3, 3.4, 3.5, 3.6, 3.7, 3.8 and 3.9 summarize the NC values resulted from applying the seven attacks described before—JPEG compression, gaussian noise, salt and pepper, gaussian filter, median filter, scaling, cropping—on set of medical images.

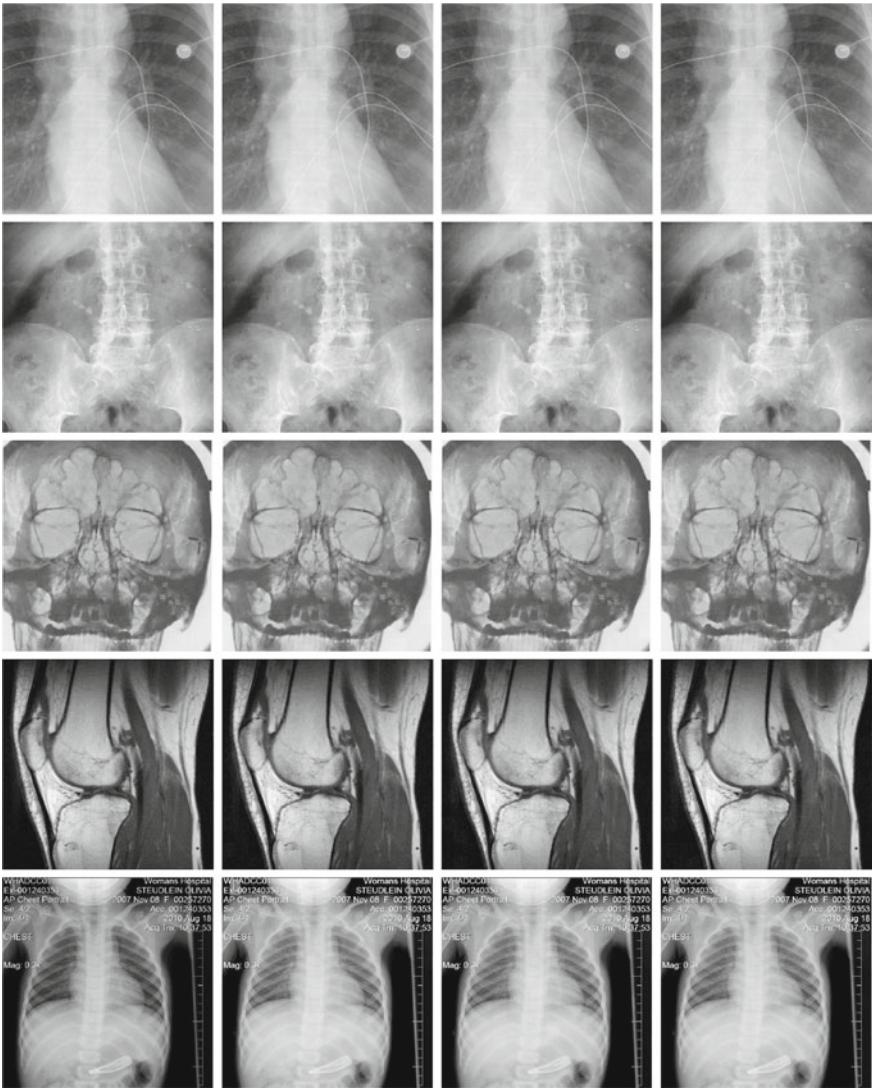


Fig. 3.2 From *left* to *right*: host image, watermarked image using classical PSO, QPSO, WQPSO respectively

Table 3.2 PSNR values for medical images with PSO, QPSO and WQPSO

Image	PSO	QPSO	WQPSO
Chest	54.42	54.32	54.82
Kidney	54.62	54.38	56.12
Skull	54.41	54.29	55.90
Knee	54.69	54.72	56.41
Stomach	54.61	55.32	54.72

Table 3.3 Robustness against JPEG compression for medical images

Method	QF	Chest	Kidney	Skull	Knee	Stomach
PSO	50	0.976	0.979	0.978	0.976	0.902
QPSO		0.982	0.982	0.982	0.982	0.917
WQPSO		0.982	0.982	0.982	0.982	0.936
PSO	40	0.971	0.977	0.977	0.976	0.895
QPSO		0.982	0.9827	0.982	0.982	0.917
WQPSO		0.982	0.9827	0.982	0.982	0.917
PSO	30	0.976	0.974	0.976	0.971	0.880
QPSO		0.964	0.982	0.982	0.982	0.917
WQPSO		0.964	0.982	0.982	0.982	0.907

Table 3.4 Robustness for Gaussian noise attacks for medical images

Var.	Alg.	Chest	Kidney	Skull	Knee	Stomach
0.001	PSO	0.971	0.976	0.971	0.939	0.870
0.005		0.857	0.879	0.877	0.723	0.757
0.001	QPSO	0.982	0.982	0.982	0.945	0.862
0.005		0.908	0.907	0.936	0.742	0.782
0.001	WQPSO	0.982	0.982	0.982	0.945	0.845
0.005		0.954	0.954	0.907	0.770	0.803

Table 3.5 Robustness for salt and pepper noise attacks for medical images

Density	Alg.	Chest	Kidney	Skull	Knee	Stomach
0.001	PSO	0.977	0.980	0.978	0.965	0.935
0.005		0.937	0.957	0.963	0.844	0.872
0.001	QPSO	0.982	0.982	0.982	0.973	0.926
0.005		0.963	0.963	0.982	0.888	0.898
0.001	WQPSO	0.982	0.982	0.982	0.963	0.935
0.005		0.964	0.982	0.982	0.870	0.899

Table 3.6 Robustness for Gaussian filter attacks for medical images

Size	Alg.	Chest	Kidney	Skull	Knee	Stomach
3 × 3	PSO	0.980	0.981	0.980	0.981	0.966
5 × 5		0.980	0.980	0.981	0.980	0.965
3 × 3	QPSO	0.982	0.982	0.982	0.982	0.973
5 × 5		0.982	0.982	0.982	0.982	0.963
3 × 3	WQPSO	0.982	0.982	0.982	0.982	0.973
5 × 5		0.982	0.982	0.982	0.982	0.973

Table 3.7 Robustness for median filter attacks for medical images

Size	Alg.	Chest	Kidney	Skull	Knee	Stomach
3 × 3	PSO	0.978	0.981	0.959	0.977	0.770
5 × 5		0.905	0.975	0.890	0.926	0.736
3 × 3	QPSO	0.982	0.982	0.982	0.982	0.779
5 × 5		0.936	0.982	0.889	0.936	0.723
3 × 3	WQPSO	0.982	0.982	0.982	0.982	0.824
5 × 5		0.917	0.982	0.917	0.945	0.721

Table 3.8 Robustness for scaling attacks for medical images

Ratio (%)	Alg.	Chest	Kidney	Skull	Knee	Stomach
75	PSO	0.980	0.981	0.980	0.981	0.837
50		0.937	0.980	0.980	0.979	0.824
75	QPSO	0.982	0.0982	0.982	0.982	0.842
50		0.982	0.0982	0.982	0.982	0.809
75	WQPSO	0.982	0.0982	0.982	0.982	0.879
50		0.982	0.0982	0.982	0.982	0.833

3.4.2 Convergence Comparison

From Tables 3.3–3.9, it is clear that the robustness performance of the watermarking approaches using QPSO and WQPSO as optimization procedure in estimating adaptive quantization step are superior. Comparing the performances of WQPSO and QPSO for different values of noise parameters, we shows that WQPSO provides higher *NC* values for most cases either in medical or ordinary image, indicating that watermarking based on WQPSO is more robust. This is due to the calculation method of mean best position with weight parameter introduced that can make the convergence speed of WQPSO higher, which may lead to good performance of the algorithm.

Figure 3.3 gives the comparison of convergence processes of PSO, QPSO and WQPSO averaged on 50 trial runs, these samples captured during assignment of different noise parameters shown in previous tables. It shows that WQPSO has the fastest convergence compared with PSO and QPSO. Since the parameter setting in

Table 3.9 Robustness for cropping attacks for medical images

Ratio (%)	Alg.	Chest	Kidney	Skull	Knee	Stomach
25	PSO	0.944	0.945	0.944	0.945	0.931
35		0.877	0.874	0.883	0.877	0.861
25	QPSO	0.946	0.946	0.944	0.946	0.937
35		0.880	0.880	0.863	0.900	0.862
35	WQPSO	0.946	0.946	0.946	0.946	0.937
35		0.880	0.900	0.871	0.910	0.854

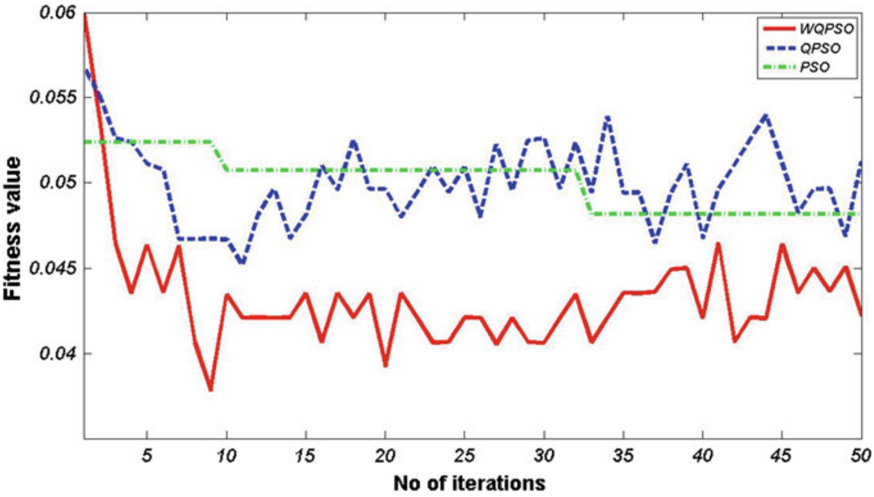


Fig. 3.3 Comparison of convergence process for skull images

WQPSO is the same as in QPSO, the fast convergence of WQPSO may be due to the weighted mean best position, which makes the particle converges to global best position more quickly.

3.5 Conclusion and Future Works

Watermarking images needs to satisfy two requirement, image fidelity and high robustness. In order to preserve both requirements, dozens of research papers are presented each year using the artificial intelligence techniques. In the presented watermarking approach, we proposed the use of weighted quantum particle swarm optimization algorithm (WQPSO) to determine the proper quantization step that was used to embed watermark bits into the singular values vector of each block of the host image. WQPSO introduced a linearly decreasing weight parameter to render the importance of particles in evolving populations. In general, the proposed approach performs much better than the classical optimization approaches, particle swarm optimization and quantum particle swarm optimization for almost all of the images and attacks. In our future work, we will be devoted to find out an adaptive method to update weight coefficient α_i instead of updating it using linearly decreasing method, and thus will enhance the performance of WQPSO further.

References

1. El Bakrawy, L.M., Ghali, N.I., Hassanien, A.E., Abraham, A.: An associative watermarking based image authentication scheme. In: The 10th International Conference on Intelligent Systems Design and Applications (ISDA2010), Cairo, Egypt, pp. 823–828 (2010)
2. Hassanien, A.E., Abraham, A., Grosan, C.: Spiking neural network and wavelets for hiding iris data in digital images. Soft Comput. **13**(4), 401–416 (2009)
3. Zhang, Q., Wang, Z., Zhang, D.: Memetic algorithm-based image watermarking scheme. In: Proceedings of the 5th International Symposium on Neural Networks: Advances in Neural Networks, pp. 845–853 (2008)
4. Findik, O., Babaoglu, I., Ulker, E.: A color image watermarking scheme based on hybrid classification method: particle swarm optimization and k-nearest neighbor algorithm. Opt. Commun. **283**(24), 4916–4922 (2010)
5. Sabat, S.L., Coelho, L.S., Abraham, A.: MESFET DC model parameter extraction using quantum particle swarm optimization. Microelectron. Reliab. **49**, 660–666 (2009)
6. Braudaway, G.: Protecting publicly-available images with an invisible image watermark. Proc. IEEE Int. Conf. Image Process. **1**, 524–527 (1997)
7. Hartung, F., Kutter, M.: Multimedia watermarking techniques. Proc. IEEE **87**(7), 1079–1107 (1999)
8. Hernandez, J., Gonzalez, F., Rodriguez, J., Nieto, G.: Performance analysis of a 2-d-multipulse amplitude modulation scheme for data hiding and watermarking of still images. IEEE J. Sel. Areas Commun. **16**(4), 510–524 (1998)
9. Arnold, M., Schmucker, M., Wolthusen, S.D.: Techniques and Applications of Digital Watermarking and Content Protection. Artech House, Norwood (2003)
10. Cox, I.J., Miller, M.L., Bloom, J.A.: Digital Watermarking. Morgan Kaufmann Publishers, San Francisco (2001)
11. Nikolaidis, A., Tsekeridou, S., Tefas, A., Solachidis, V.: A survey on watermarking application scenarios and related attacks. Proc. IEEE Int. Conf. Image Process. **3**, 991–994 (2001)
12. Fitts, D.D.: Principles of Quantum Mechanics as Applied to Chemistry and Chemical Physics. University of Cambridge, Cambridge (1999)
13. Coelho, L.S.: A quantum particle swarm optimizer with chaotic mutation operator. Chaos Solitons Fractals **37**, 1409–1418 (2008)
14. Phillips, A.C.: Introduction to Quantum Mechanics. British Library, London (2003)
15. Liu, J., Sun, J., Xu, W.: Quantum-behaved particle swarm optimization with immune operator. In: Foundations of Intelligent Systems, Lecture Notes in Computer Science, vol. 4203, pp. 77–83 (2006)
16. Kuk-Hyun, H., Jong-Hwan, K.: Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans. Evol. Comput. **6**(6), 580–593 (2002)
17. Jang, J.S., Han, K.H., Kim, J.H.: Face detection using quantum-inspired evolutionary algorithm. In: IEEE Congress Proceeding on Evolutionary Computation, pp. 2100–2106 (2004)
18. Yang, J., Li, B., Zhuang, Z.Q.: Multi-universe parallel quantum genetic algorithm and its application to blind-source separation. Proc. Int. Conf. Neural Netw. Sig. Process. **1**, 393–398 (2003)

19. Jianhua, X.: Improved quantum evolutionary algorithm combined with chaos and its application. Lect. Notes Comput. Sci. **5553**, 704–713 (2009)
20. Xi, M., Sun, J., Xu, W.: An improved quantum-behaved particle swarm optimization algorithm with weighted mean best position. Appl. Math. Comput. **205**(2), 751–759 (2008)
21. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. IEEE Proc. Neural Netw **IV**(1), 1942–1948 (1995)
22. Kennedy, J.: Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance. In: Proceedings of the 1999 Congress of Evolutionary Computation, vol. 3, pp. 1931–1938 (1999)
23. Aslantas, V., Dogan, A.L., Ozturk, S.: DWT-SVD based image watermarking using particle swarm optimizer. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 241–244 (2008)
24. Aslantas, V.: A singular-value decomposition-based image watermarking using genetic algorithm. Int. J. Electron. Commun. **62**, 386–393 (2007)
25. Lai, C.C., Huang, H.C., Tsai, C.C.: Image watermarking scheme using singular value decomposition and micro-genetic algorithm. In: Proceedings of International Conference on Intelligent Information Hiding and Multimedia, Signal Processing, pp. 469–472 (2008)
26. Qi, X., Bialkowski, S., Brimley, G.: An adaptive QIM-and SVD-based digital image watermarking scheme in the wavelet domain. In: Proceedings of IEEE International Conference on Image Processing, pp. 421–424 (2008)
27. Shaomin, Z., Liu, J.: A novel adaptive watermarking scheme based on human visual system and particle swarm optimization. In: Information Security Practice and Experience, Lecture Notes in Computer Science, vol. 5451, pp. 136–146 (2009)
28. Nikham, T., Amiri, B., Olamaei, J., Arefei, A.: An efficient hybrid evolutionary optimization algorithm based on PSO and SA for clustering. J. Zhejiang Univ. Sci. A **10**(4), 512–519 (2009)
29. Dharwadkar, N., Amberker, B., Panchannavar, P.: Reversible fragile medical image watermarking with zero distortion. In: International Conference on Computer and Communication Technology, pp. 248–254 (2010)
30. Kaur, R.: A medical image watermarking technique for embedding EPR and Its quality assessment using no-reference metrics. IJITCS 5(2), 73–79 (2013)
31. Fakhari, P., Vahedi, E., Lucas, C.: Protecting patient privacy from unauthorized release of medical images using a bio-inspired wavelet-based watermarking approach. Digit. Sig. Process. **21**(3), 433–446 (2011)
32. Soliman, M.M., Hassanien, A.E., Ghali, N.I., Onsi, H.M.: An adaptive watermarking approach for medical imaging using swarm intelligent. Int. J. Smart Home **6**, 37–51 (2012)

Chapter 4
Bio-inspiring Techniques in Watermarking Medical Images: A Review

Mona M. Soliman, Aboul Ella Hassanien and Hoda M. Onsi

Abstract Bio-inspiring (BI) is a well-established paradigm with current systems having many of the characteristics of biological computers and capable of performing a variety of tasks that are difficult to do using conventional techniques. BI is considered as one of the most important increasing fields, which attract a large number of researchers and scientists working in areas such as neuro-computing, global optimization, swarms and evolutionary computing. On the other hand, digital radiological modalities in modern hospitals have led to the producing a variety of a vast amount of digital medical files. Therefore, for the medical imaging, the authenticity needs to ensure the image belongs to the correct patient, the integrity check to ensure the image has not been modified, and safe transfer are very big challenges. The integrity of the images must be protected by using watermarking, which is called integrity watermark. At the same time the copyright and intellectual property of the medical images should be also protected, which is called copyright watermark. This chapter presents a brief overview of well known Bio-inspiring techniques including neural networks, genetic algorithm, swarms and evolutionary algorithms and show how BI techniques could be successfully employed to solve watermarking problem. Challenges to be addressed and future directions of research are also presented and an extensive bibliography is included.

M. M. Soliman (✉) · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Faculty of Computers and Information,
Cairo University, Cairo, Egypt
e-mail: monasolyman_it@yahoo.com
URL: http://www.egyptscience.net

A. E. Hassanien
e-mail: aboitcairo@gmail.com

H. M. Onsi
Faculty of Computers and Information, Cairo University, Cairo, Egypt

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_4, © Springer-Verlag Berlin Heidelberg 2014

4.1 Introduction

The development of multimedia and communication technologies has made medical images act important roles in the fields of telediagnosis and telesurgery. At the same time such advances provide new means to share, handle and process medical images, it also increases security issues in terms of: confidentiality, availability and integrity. Since any person with privilege can access to images which are contained in database and can modify them maliciously, the integrity of the images must be protected by using watermarking, which is called integrity watermark. Meanwhile, web-based image database system contains valuable medical image resources for not only research purpose but also commercial purpose. Therefore the copyright and intellectual property of the database should be also protected by a watermark, which is called copyright watermark. The basic principle of watermarking methods is to add copyright information into the original data, by embedding it into the original image. Then if the image is modified in any sense, it can be detected with the watermark. Initially, the watermark could be simply a unique number, such as the patient insurance code but as research moves into new paths, a new role has been given to the watermark: to include (apart from hospital digital signatures or copyright information), the electronic patient record, digital documents with diagnosis, blood test profiles or an electrocardiogram signal.

For the medical image, the authenticity needs to ensure the image belongs to the correct patient, the integrity check to ensure the image has not been modified, and safe transfer are very big challenges. Also, when a digital medical image is opened for diagnosis, it is important that an automated framework exists to verify the authenticity and integrity of the image itself. Hospital Information System (HIS) and Picture Archiving and Communication System (PACS) have been established to provide security solutions to ensure confidentiality, integrity and authentication [1]. Digital image watermarking provides copyright protection to digital image by hiding appropriate information in original image to declare rightful ownership [2]. The primary applications of watermarking are to protect copyrights and integrity verification. The main reason for protecting copyrights is to prevent image piracy when the transmitter sends it on the internet. For integrity verification, it is important to ensure that the medical image originated from a specific source and that it has not been changed, manipulated or falsified [3].

There are several types of algorithms for watermarking. Each type of algorithms has its own advantages and limitations. No method can provide fully perfect solution. Each type of solution has robustness to some type of attacks but is less resilient to some other types of attacks. In medical applications, because of their diagnostic value, it is very important to maintain the quality of images. For this matter, the development of a new algorithm that can satisfy both invisibility and robustness is needed. Improvements in performance of watermarking schemes can be obtained by several methods. One way is to make use of Intelligent computing techniques. The objective of this chapter is to present to the medical watermarking research communities some of the state-of-the-art in BI applications to medical watermarking problem and motivate research in new trend-setting directions. Hence, we review and discuss some representative methods to provide inspiring examples to illustrate how BI techniques can be applied to solve medical watermarking problems. We want to stress that the literature in this domain is of course huge and that therefore the work that we include here is only exemplatory and focussing on recent advances, while many other interesting research approaches had to be omitted due to space limitations [4].

The rest of the chapter is organized as follows. Section 4.2 introduces the fundamental aspects of the key components of modern BI techniques including: neural networks, evolutionary algorithms (EA), genetic algorithms (GA) and swarm intelligence. Of course there are many other methods of BI but, we only focusing on those BI methods used in medical image watermarking. Section 4.3 review some details of medical image watermarking categories and classification with brief illustration of the generic watermarking procedure. Section 4.4 reviews and discusses some successful approaches to illustrate how BI can be applied to medical image watermarking problems. Section 4.5 explore different watermarking assessment measures used in the evaluation of watermarking algorithms. Challenges and future trends are presented in Sect. 4.6 and an extensive bibliography is provided.

4.2 Bio-inspiring Computing

In this section, we explore an overview of modern BI techniques applied successfully on medical image watermarking problem including, neural networks, EAs, GAs and swarm intelligence.

4.2.1 Artificial Neural Networks

Artificial Neural Network (ANN) is a computational structure paradigm modeled on the biological process that is inspired by the way biological nervous systems, such as the brain, processes information. The key element of this paradigm is the novel structure of the information processing system [5]. Neural computing is an alternative to programmed computing which is a mathematical model inspired by biological models. This computing system is made up of a number of artificial neurons and a huge number of interconnections between them. There are six major components make up an artificial neuron [6]. These components are valid whether the neuron is used for input, output, or is in the hidden layers:

- *Weighting Factors*: A neuron usually receives many simultaneous inputs. Each input has its own relative weight, which gives the input the impact that it needs on the processing element's summation function. Some inputs are made more important than others to have a greater effect on the processing element as they combine to produce a neural response. Weights are adaptive coefficients that determine the intensity of the input signal as registered by the artificial neuron.

- *The summation function*: The input and weighting coefficients can be combined in many different ways before passing on to the transfer function. In addition to summing, the summation function can select the minimum, maximum, majority, product or several normalizing algorithms. The specific algorithm for combining neural inputs is determined by the chosen network architecture and paradigm. Some summation functions have an additional activation function applied to the result before it is passed on to the transfer function for the purpose of allowing the summation output to vary with respect to time.
- *Transfer Function*: The result of the summation function is transformed to a working output through an algorithmic process known as the transfer function. In the transfer function the summation can be compared with some threshold to determine the neural output. If the sum is greater than the threshold value, the processing element generates a signal and if it is less than the threshold, no signal (or some inhibitory signal) is generated.
- *Scaling and Limiting*: After the transfer function, the result can pass through additional processes, which scale and limit. This scaling simply multiplies a scale factor times the transfer value and then adds an offset. Limiting is the mechanism which insures that the scaled result does not exceed an upper, or lower bound. This limiting is in addition to the hard limits that the original transfer function may have performed.
- *Output Function "Competition"*: Each processing element is allowed one output signal, which it may give to hundreds of other neurons. Normally, the output is directly equivalent to the transfer function's result. Some network topologies modify the transfer result to incorporate competition among neighboring processing elements.

In most cases ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance with a chosen learning algorithm. Learning algorithms can be supervised, unsupervised, or reinforced [7]. The reader may refer to [8–10] for an extensive overview of the artificial neural networks.

4.2.2 Evolutionary Algorithms

Evolution in nature is responsible for the design of all living beings on earth, and for the strategies they use to interact with each other. Evolutionary algorithms employ this powerful design philosophy to find solutions to hard problems. The idea of applying the biological principle of natural evolution to artificial systems, introduced more than four decades ago, has seen impressive growth in the past few years. Known as evolutionary algorithms or evolutionary computation, these techniques are common nowadays, having been successfully applied to numerous problems from different domains, including optimization, automatic programming, circuit design, machine learning, economics, ecology and population genetics [11].

Over many generations, natural populations evolve according to the principles of natural selection and "survival of the fittest". By mimicking this process, evolutionary algorithms are able to "evolve" solutions to real world problems, if they have been suitably encoded [12].

Evolutionary computation makes use of a metaphor of natural evolution [13]. According to this metaphor, a problem plays the role of an environment wherein lives a population of individuals, each representing a possible solution to the problem. The degree of adaptation of each individual (i.e. candidate solution) to its environment is expressed by an adequacy measure known as the fitness function. The phenotype of each individual, i.e. the candidate solution itself, is generally encoded in some manner into its genome (genotype). Like evolution in nature, evolutionary algorithms potentially produce progressively better solutions to the problem. This is possible thanks to the constant introduction of new genetic material into the population, by applying so-called genetic operators that are the computational equivalents of natural evolutionary mechanisms.

There are several types of evolutionary algorithms, among which the best known are genetic algorithms [14], genetic programming [15], evolution strategies [16], evolutionary programming [17] and learning classifier systems [18]; though different in the specifics they are all based on the same general principles. All share a common conceptual base of simulating the evolution of individual structures via processes of selection, mutation and reproduction.

These processes depend on the perceived performance of the individual structures as defined by the environment. EAs deal with parameters of finite length, which are coded using a finite alphabet, rather than directly manipulating the parameters themselves. This means that the search is unconstrained neither by the continuity of the function under investigation, nor the existence of a derivative function.

The application of an evolutionary algorithm involves a number of important considerations. The first decision to take when applying such an algorithm is how to encode candidate solutions within the genome. The representation must allow for the encoding of all possible solutions while being sufficiently simple to be searched in a reasonable amount of time. Next, an appropriate fitness function must be defined for evaluating the individuals. The (usually scalar) fitness must reflect the criteria to be optimized and their relative importance. Representation and fitness are thus clearly problem-dependent, in contrast to selection, crossover and mutation, which seem prima facie more problem-independent. Practice has shown, however, that while standard genetic operators can be used, one often needs to tailor these to the problem as well.

4.2.3 Genetics Algorithms

Genetic algorithms (GAs), introduced by Hollard in his seminal work, are commonly used as adaptive approaches that provide a randomized, parallel and global search method based on the mechanics of natural selection and genetics in order to find solutions. In biology, the gene is the basic unit of genetic storage [19]. Within cells, genes are strung together to form chromosomes. The simplest possible sexual reproduction is between single-cell organisms. The two cells fuse to produce a cell with two sets of chromosomes, called a diploid cell. The diploid cell immediately undergoes meiosis. In meiosis, each of the chromosomes in the diploid cell makes an exact copy of itself. Then the chromosome groups (original and copy) undergo crossover with the corresponding groups, mixing the genes somewhat. Finally the chromosomes separate twice, giving four haploid cells. Mutation can occur at any stage, and any mutation in the chromosomes will be inheritable. Mutation is essential for evolution. There are three types relevant to genetic algorithms: point mutations where a single gene is changed, chromosomal mutations where some number of genes are lost completely, and inversion where a segment of the chromosome becomes flipped.

The power of GAs as mentioned in [20] comes from the fact that the technique is robust and can deal successfully with a wide range of problem areas including those which are difficult for other methods to solve. GAs are not guaranteed to find the global optimum solution to a problem but they are generally good at finding acceptably good solutions to problems.

In GA, a candidate solution for a specific problem is called an individual or a chromosome and consists of a linear list of genes. Each individual represents a point in the search space, and hence a possible solution to the problem. A population consists of a finite number of individuals. Each individual is decided by an evaluating mechanism to obtain its fitness value. Based on this fitness value and undergoing genetic operators, a new population is generated iteratively with each successive population referred to as a generation. The GAs use three basic operators (reproduction, crossover and mutation) to manipulate the genetic composition of a population [21]. A population is created with a group of randomly individuals. The individuals in the population are then evaluated by fitness function. Two individuals (off-spring) are selected for the next generation based on their fitness. Crossover is a process yielding recombination of bit strings via an exchange of segments between pairs of chromosomes to create the new individuals. Finally, mutation has the effect of ensuring that all possible chromosomes are reachable or a certain number of generations have passed.. The reader may refer to [22–26] for an extensive overview of the GAs.

4.2.4 Swarm Intelligence

The expression **swarm intelligence** was introduced by Beni and Wang in 1989, in the context of cellular robotic systems, SI systems are typically made up of a population of simple agents interacting locally with one another and with their environment. Although there is normally no centralized control structure dictating how individual agents should behave, local interactions between such agents often lead to the emergence of global behavior. Examples of systems like this can be found in nature, including ant colonies, bird flocking, animal herding, bacteria molding and fish schooling [27]. Optimization techniques inspired by swarm intelligence have become increasingly popular during the last decade [28]. They are characterized by a decentralized way of working that mimics the behavior of swarms. The advantage of these approaches over traditional techniques is their robustness and flexibility. These properties make swarm intelligence a successful design paradigm for algorithms that deal with increasingly complex problems. Two successful examples of optimization techniques inspired by swarm intelligence are : ant colony optimization and particle swarm optimization [29]. Ant Colony Optimization (ACO) has been successfully applied to solve various engineering optimization problems. ACO algorithms can also be used for clustering datasets and to optimize rule bases.

Dorigo and Blumb in [30] introduced the first ACO algorithms in the early 1990s. The development of these algorithms was inspired by the observation of ant colonies. Ants are social insects. They live in colonies and their behavior is governed by the goal of colony survival rather than being focused on the survival of individuals. The behavior that provided the inspiration for ACO is the ants foraging behavior, and in particular, how ants can find shortest paths between food sources and their nest. While moving, ants leave a chemical pheromone trail on the ground [31]. Ants can smell pheromone. When choosing their way, they tend to choose, in probability, paths marked by strong pheromone concentrations. As soon as an ant finds a food source, it evaluates the quantity and the quality of the food and carries some of it back to the nest. During the return trip, the quantity of pheromone that an ant leaves on the ground may depend on the quantity and quality of the food. The pheromone trails will guide other ants to the food source. The concept of particle swarms, although initially introduced for simulating human social behaviors, has become very popular these days as an efficient search and optimization technique. PSO has gained increasing popularity among researchers and practitioners as a robust and efficient technique for solving difficult optimization problems. In PSO, individual particles of a swarm represent potential solutions, which move through the problem search space seeking an optimal, or good enough, solution. The particles broadcast their current positions to neighboring particles. The position of each particle is adjusted according to its velocity (i.e. rate of change) and the difference between its current position, respectively the best position found by its neighbors, and the best position it has found so far. As the model is iterated, the swarm focuses more and more on an area of the search space containing high-quality solutions [32]. We have to note that PSO is mainly used for continuous optimization while ACO is mainly used for combinatorial optimization. The reader may refer to [33–38].

4.3 Medical Image Watermarking: Classification and Generic Model

Digital radiological modalities in modern hospitals have led to the producing a variety of a vast amount of digital medical files. It is very common and usual for physicians to participate in technical group communication existed between physicians and hospitals in order to diagnosing and spotting the patient's problem. Hospital Information System (HIS) comprising radiology information system (RIS) and picture archiving and communication system (PACS) based on Digital Imaging and Communication in Medicine standard (DICOM) has facilitated offering various e-Health services. These e-Health services are introducing new practices for the profession as well as for the patients by enabling remote access, transmission and interpretation of the medical images for diagnosis purposes [39]. The health-care professionals use the Internet for transmitting or receiving Electronic Patient Records (EPR) via e-mail. An EPR typically contains the health history of a patient, including X-ray images, CT-Scan images, physical examinations report, laboratory tests, treatment procedures, prescriptions, radiology examinations etc. An EPR can be represented in various forms such as diagnostic reports, images, vital signals, etc. An EPR transmitted through the Internet is very important since it contains the medical information of a person in digital format [40].

4.3.1 Medical Watermarking Classification

As ordinary watermarking techniques, Medical image watermarking techniques can be categorized in different ways. They can be classified according to the type of watermark being used, i.e. the watermark may be visible or invisible. Visible watermarking is the process of embedding the watermark into some visible parts of the digital media [41]. Examples of visible watermarking can be seen as TV logos and company emblems. As for invisible watermarking, it is the process of embedding the watermark confidentially into the parts of the digital media known only by the owner [42].

For different purposes, digital watermarking has been branched into two classifications: robust watermarking technique and fragile watermarking technique [43]. Robust watermarks are designed to resist intentional or unintentional image modifications for instance filtering, geometric transformations, noise addition, etc. In contrast, fragile watermarking is used to determine the modifications on the digital media that is, to ensure the integrity of the digital media [41]. The design goal of robust watermarking is to make the embedded watermarks remain detectable after being attacked. In contrast, the requirements of fragile watermarking are to detect the slightest unauthorized modifications and locate the changed regions.

Another classification is based on domain which the watermark is applied i.e. the spatial domain or the frequency domain. The earlier watermarking techniques were almost in spatial domain. Spatial-domain schemes embed the watermark by directly modifying the pixel values of the original image [3]. Spatial domain methods are less complex and not robust against various attacks as no transform is used in them. The frequency domain techniques based on transforming the original media to frequency coefficient by using some transformations such as discrete Fourier transformation, discrete cosine transformation and discrete wavelet transformation. Then, the watermark is embedded by modifying coefficients [44]. Frequency domain techniques are robust as compared to spatial domain methods [45].

There are three possible categories for medical image watermarking have been identified in the literature survey [46]. The first category is based on region of non-interest. A medical image in case of clinical outcome can be divided in two parts, the region of interest (ROI) where the diagnosis focuses, or it can be defined as an area including important information and must be stored without any distortion, while the region of non-interest (RONI), is considering the remaining area. The definition of the ROI space depends on the existence of a clinical finding and its features [47]. The second category is based on reversibility. Here the watermarking scheme should able retrieve the original image from the watermarked image without any loss of information at the extraction process. The reversible watermarking not only provides authentication and tamper proofing but also can recover the original image from the suspected image. After the verification process if the transmitted image is deemed to be authentic the doctor reconstitutes the original image (without any degradation) and uses it in its diagnosis avoiding all risk of modification [48]. The third category is based on non-reversible. Here, tolerable information loss is accepted as in lossy compression. Any watermarking scheme can be classified into either reversible or irreversible. It can use region of non-interest for embedding. However, the selection of region of images for embedding is application specific.

4.3.2 Generic Medical Image Watermarking System

"Watermarking" is the process of hiding digital information in a carrier signal; the hidden information should, but does not need to contain a relation to the carrier signal. The information to be embedded in a signal is called a digital watermark. The signal where the watermark is to be embedded is called the host signal. All watermarking systems as shown in Fig. 4.1 including medical image watermarking system are usually divided into three distinct steps, embedding, attack and detection.

In embedding, an algorithm accepts the host and the data to be embedded, and produces a watermarked signal. For a blind watermark, the goal is that the digital data appears to be the same as before the embedding process, and the casual user is unaware of the watermark's presence. Then, the watermarked digital signal is transmitted or stored, usually transmitted to another person. If this person makes a modification, this is called an attack. While the modification may not be malicious, the term attack arises from copyright protection application, where third parties may attempt to remove the digital watermark through modification. We can classify attacks into the Simple attack, Detection-disabling or Geometric attacks, Ambiguity attacks and Removal attacks.

Detection (often called extraction) is an algorithm which is applied to the attacked signal to attempt to extract the watermark from it. If the signal was unmodified during transmission, then the watermark still is present and it may be extracted. In robust digital watermarking applications, the extraction algorithm should be able to produce the watermark correctly, even if the modifications were strong. In fragile digital watermarking, the extraction algorithm should fail if any change is made to the signal. The watermark detector decides whether a watermark is present or not (refer to [49]).

According to [50] watermark algorithms can be classified into generations based on the watermark embedding domain. The first generation include those algorithms

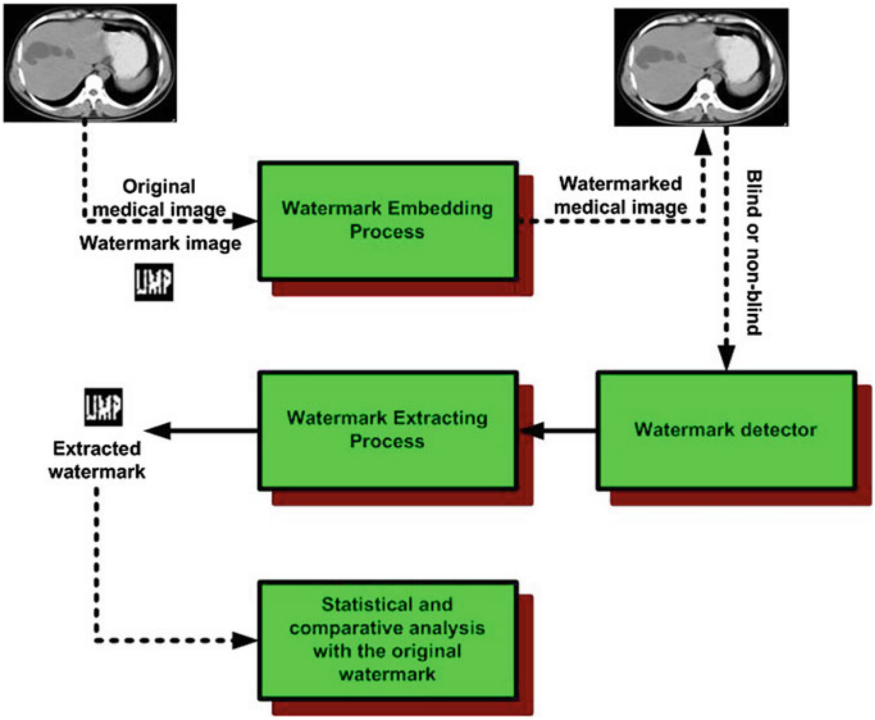

Original
medical image
Watermark image
Watermark Embedding Process
Watermarked
medical image
Blind or non-blind
Watermark Extracting Process
Watermark detector
Extracted
watermark
Statistical and comparative analysis with the original watermark

Fig. 4.1 General medical watermarking system

that embed watermark signal into spatial domain. There have been several enhancements since the first generation to improve the performance in terms of capacity, invisibility, and robustness of the watermark. The second generation algorithms use various transformations to insert the watermark in the transform domain coefficients to enhance robustness. Third generation techniques build on existing first and second generation algorithms and also include hybrid domains to allow information fusion from different domains. The proposed third generation algorithms explore the use of computational intelligence techniques to insert a high capacity watermark in both the spatial and transform domains.

4.4 BI in Medical Image Watermarking

4.4.1 Artificial Neural Networks in Medical Image Watermarking

Only few works have been reported to use neural network to embed watermark into medical images [51] watermark mammograms images. These images are watermarked in order to proof its integrity; not modified by unauthorized person, and to ascertain the authenticity; ensuring that the image belong to the correct patient and source. Mammograms contain diagnostic information which can be used for early detection of breast cancer diseases and breast abnormality.

Olanr in [51] have been exploited using the Complex version of ANN (CVNN), trained by Complex backpropagation (CBP) algorithm. This technique was used to embed and detect forge watermark in Fast Fourier Transform FFT domain. The performance of the algorithm has been evaluated using mammogram images. The imperceptibility and detection accuracy was appraised with objective performance measure. Results indicate that watermarked mammogram were perceptually indistinguishable from the host mammogram, hence the application of the developed CVNN-based watermarking technique in medical images can improve correct diagnoses. Ability of the algorithm to localize modification undergone makes it a unique and efficient algorithm for authentication and tamper detection as well as blind detection applications. CVNN is use to process complex valued data (CVD) such as in image with real and imaginary component. CVNN is made up of Complex-Valued Feedforward (CVFF) and Complex Back-Propagation (CBP) algorithm. CVNN has been studied and developed by authors in solving various problems. The CVNN consists of an interconnection of Complex-Valued (CV) neurons and complex valued synaptic weights. It processes information using a connectionist approach to computation in complex domain.

Oues in [52], attempts to define an adaptive watermarking scheme based on Multi-Layer Feed forward (MLF) neural networks. Neural network applied to digital watermark embedded process simulates human visual characteristic to determine the maximum watermark embedded intensity endured by middle frequency coefficients in every one of (8*8) image block DCT coefficients. The watermarking scheme in this work has been tested on the medical images with size of 512 × 512 pixels. The experimental results demonstrated that the proposed method significantly improve the perceptual quality of a watermarked image while preserving the robustness against various attacks such as of compression, cropping and filtering. The trade-off between the imperceptibility and robustness is one of most serious challenges in digital watermarking system, in particular the medical imaging.

4.4.2 Genetic Algorithms in Medical Image Watermarking

The learning capabilities of GA make an effective trade off between the watermark payload and imperceptibility, through effective selection of threshold. This trade off has much more importance in case o sensitive imagery such as medical imagery. In [21] author developed an intelligent reversible watermarking approach for medical images. Companding technique is effectively used to achieve higher PSNR value for the images and is controlled using threshold. We have observed that the threshold value has a pronounce effect on the actual payload available for watermark embedding. The higher the threshold, the lower is the companding, and the corresponding companding error, and the higher is the effective payload. Thus, with change in the threshold, the effective payload and PSNR values are also changed but in reciprocating manner. The characteristics of each wavelet subband are used to evolve a threshold matrix. An optimum threshold value is computed for each of the wavelet subband.

Reference [49] presents a robust technique embedding the watermark of signature information or textual data around the ROI of a medical image based on genetic algorithms. A fragile watermark is adopted to detect any unauthorized modification. The embedding of watermark in the frequency domain is more difficult to be pirated than in spatial domain. The embedded watermark in the coefficients of the transformed image will be somewhat disturbed in the process of transforming the image from its frequency domain to spatial domain because of deviations in converting real numbers into integers. This work developed a technique to correct the errors by using genetic algorithm. In order to protect the copyright of medical images, the watermark is embedded surrounding their ROI parts. A high compression ratio is preferable for reducing transmission time. But, it is difficult to attain the same compression ratio for an image with low PSNR, because it might degrade the original image, making it difficult for clinical reading. Hence it is necessary to find an optimal trade-off between the image quality and compression ratio. In addition, it is also essential to ensure that the compressed image has sufficient bandwidth to accommodate the watermark payload. Meanwhile, the embedded watermark is pre-processed by SPIHT (set partitioning in hierarchical trees) Near-Lossless Compression, whose main requirement is that of ensuring that the maximum error between the original and the compressed image does not exceed a fixed threshold. In the same line, the concept of near-lossless watermarking has been introduced recently to satisfy the strict requirements for medical image watermarking. Moreover, these techniques do not adaptively arrive at an optimal compression ratio. A single compression technique might not be suitable for all medical images because of their differing noise characteristics. Reference [53] attempts to investigate, for the first time, the application of GA in achieving an optimal compression ratio for dual watermarking in wavelet domain without degrading the image.

Reference [54] presents a lossless data hiding method using integer wavelet transform and Genetic Programming (GP) based intelligent coefficient selection scheme. By exploiting information about the amplitude of the wavelet coefficient and the type of the sub band, GP is used to evolve a mathematical function in view of the payload size and imperceptibility of the marked image. The evolved mathematical function acts like a compact but robust coefficient map for the reversible watermarking approach. Information is embedded into the least significant bit-plane of those high frequency wavelet coefficients that are intelligently selected by the Genetic Programming module.

4.4.3 Swarms Intelligent in Medical Image Watermarking

Image watermarking can be viewed as an optimization problem. Reference [55] proposed an adaptive and optimal watermark method for brain magnetic resonance images. First it have used segmentation to extract region of interests (ROI). Patient information and hospital logo has been used as a watermark and embeded in the non-region of interest part so that ROI remain same after embedding watermark. Watermark has been embedded in the Discrete Wavelet Transform (DWT) domain. Particle swarm optimization (PSO) has been used to optimize the strength of watermark intelligently. This work proposed an adaptive watermark method for brain magnetic resonance images that can be used to secure medical information. First segmentation to extract region of interests (ROI) is performed. Patient information and hospital logo has been used as a watermark and embedded in the non-region of interest part so that ROI remain same after embedding watermark. Watermark has been embedded in the Discrete Wavelet Transform (DWT) domain. Particle swarm optimization (PSO) has been used to optimize the strength of watermark intelligently.

Soliman et al. in [56–59] discuss the subject of the optimization of medical image watermarking and provide different solutions. Figure 4.2 shows the model designed by Soliman et al. [56] to use PSO in building robust system of medical image watermark.

Soliman et al. in [56] introduce an adaptive watermarking scheme in medical imaging based on swarm intelligence. The watermark bits are embedded on singular value vector of each embedding block within low frequency sub-band in the hybrid DWT-DCT domain. In adaptive watermark scheme the quantization step that used in determining the embedding watermark is changed for every image. For a single image the quantization step is not fixed but it change its value through PSO training procedure till reaching the best quantization step and hence best locations for watermark bits to be embedded. The embedding strength is more or less proportional to the perceptual sensitivity to distortions for using adaptive quantization step size. In order to resist the normal signal processing and other different attacks, the quantization step has to be as high as possible. However, because the watermark directly affects the host image, it is obvious that the higher the quantization step, the lower the quality of the watermarked image will be. In other words, the robustness and the imperceptibility of the watermark are contradictory to each other. The experiments is performed on different MRI medical images, and the proposed approach performance compared to other scheme with scheme built on ordinary methods and its shown its superior.

Soliman et al. in [57] present a novel application of Quantum Particle Swarm Optimization (QPSO) in the field of medical image watermarking for copyright protection and authentication. The global convergence of PSO cannot always be guaranteed because the diversity of population is decreased with evolution developed. To deal with this problem, concept of a global convergence guaranteed method called as Quantum behaved Particle Swarm Optimization (QPSO) was developed [60]. It provides a good scheme for improving the convergence performance of PSO because it is a theoretically global convergence algorithm. It utilizing human visual system (HVS) characteristics and QPSO algorithm in adaptive quantization index modulation and singular value decomposition in conjunction with discrete wavelet transform (DWT) and discrete cosine transform (DCT). This work provides an enhanced version of PSO using quantum theory. Quantum computing is a new class of computing based on the concepts and principles of quantum theory, such as superposition of

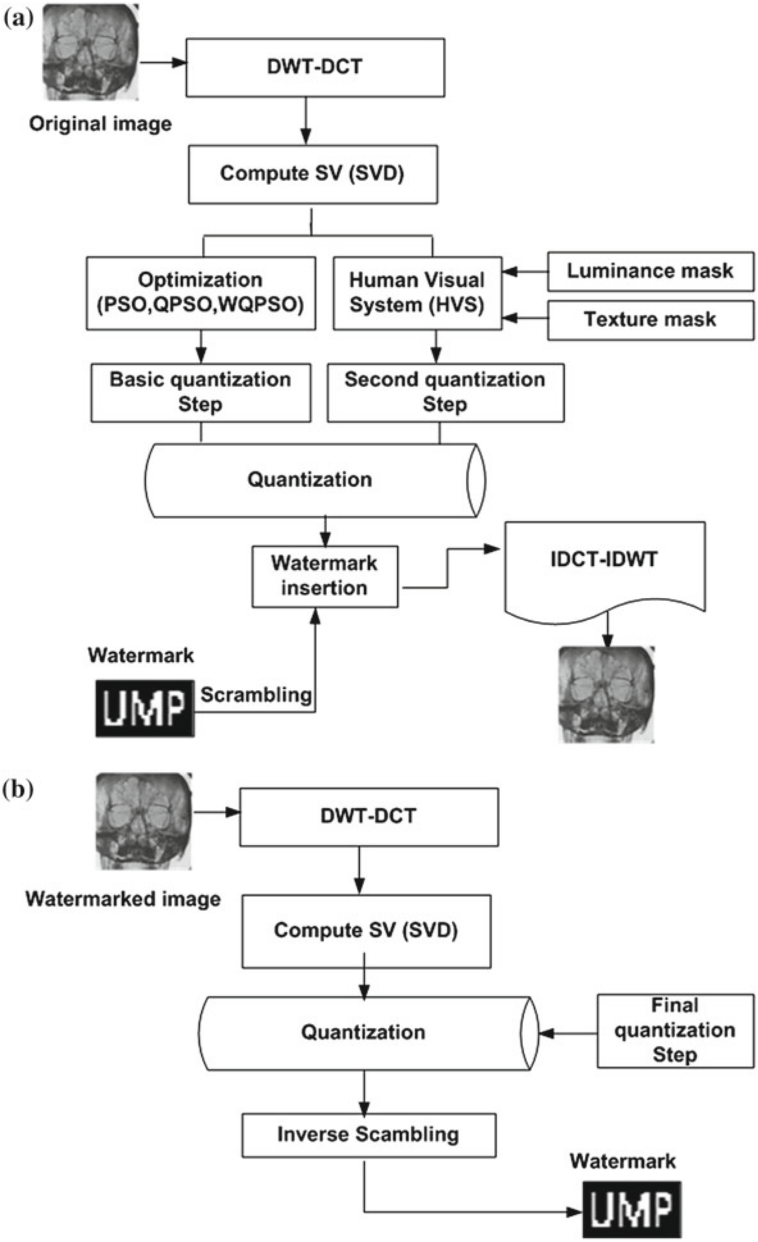


Fig. 4.2 Proposed model of medical image watermark

quantum states, entanglement and intervention. This research area was first proposed by Benioff in the early 1980s. Since Benioff in [61] introduced it, quantum computing has developed rapidly, and it also has been turning out that quantum computing has significant potential to be applied to various difficult problems including optimization. In terms of classical mechanics, a particle is depicted by its position vector x_i and velocity vector v_i, which determine the trajectory of the particle. The particle moves along a determined trajectory in Newtonian mechanics, but this is not the case in quantum mechanics. Experimental results prove the effectiveness of the proposed algorithm that yields a medical watermarked image with good visual fidelity, at the same time watermark able to withstand a variety of attacks including JPEG compression, Gaussian noise, Salt and Pepper noises, Gaussian filter, median filter, image cropping and image scaling.

Another modification had been added to the work proposed in [56], where, a novel watermarking approach based on weighted quantum particle warm optimization (WQPSO) is presented [58] with the aim to balance the global and local searching abilities, and focusing on adaptive determination of the quantization parameters for singular value decomposition. WQPSO introduce a new control parameter in QPSO algorithm, where in QPSO the best position of the particle swarm is computed and the current position of each particle is updated. In WQPSO the best particle position is replaced by a weighted mean best position. The most important problem is how to evaluate particle importance in calculate the value of weighted best position. the particles were ranked in descendent order according to their fitness value first, then assign each particle a weight coefficient linearly decreasing with the particle's rank, that is, the nearer the best solution, and the larger its weight coefficient.

4.4.4 Hybrid Bio-inspiring Systems in Medical Watermarking

Fakhari et al. in [1] combine recently developed stochastic and bio-inspired optimization algorithms called Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) to improve the performance of data hiding In designing a digital image watermarking system, it encounter two conflicting objectives which are visual quality and robustness. This work, propose an image watermarking method based on the discrete wavelet transform for the application of tracing colluders in clinical environment. In their work, 24 particles are used for PSO optimization. As mentioned before, watermark keys is embedded in all 4 frequency parts and 4 times in each, so the dimension of our problem is 16. Each particle represents 16 parameters to be searched. At this stage, the proposed algorithm educate watermarking algorithm in a way that it can resist different kinds of attacks in addition to having a high level of perceptual quality. Similarly and by considering the important factors explained in the previous part, GA is an efficient approach to search for optimal places, resulting in optimum performance of our proposed digital watermarking scheme. In this work, 24 chromosomes are used for GA optimization. Each chromosome represents 16 places to be searched with the resolution of 20 bits per variable, resulting in a total length of 320 bits. Training the watermarking algorithm to be resistant against different kinds of attacks and having a high level of perceptual quality are equivalent to converging in acceptable time and number of iterations. To achieve this goal, the mentioned 16 factors should be set properly and this is carried out by getting benefit from a GA. The GA block uses information from the fitness function to optimize the watermarking algorithm and find the best possible places like PSO.

Inspired by PSO and GA, Soliman et al. in [59] introduce a hybrid d Genetic Particle Swarm Optimization (GPSO) technique by combining the advantages of both PSO and GA. The algorithm starts by applying PSO procedure in the search space and allow particles to adjust their velocity and position according to PSO equations, then in the next step we select a certain number of particles according to GA selection methods. The particles are matched into couples. Each couple reproduces two children by crossover. Then some children are adjusted by applying mutation process. These children are used to replace their parents of the previous particles to keep the number of particles unchanged. By combination of PSO and GA, evolution process is accelerated by flying behavior and the population diversity is enhanced by genetic mechanism. The proposed GPSO is modeled to solve such optimization problem of medical image watermarking.

4.5 Watermarking Assessment Measures

Although digital watermarking provides better data security and authentication, however, the major drawback is distortions/visual artifacts introduced during data embedding which makes it difficult in detecting forged watermarks introduced by attackers. The accuracy of diagnosis and treatment of patients greatly depends on the received (watermarked) image by clinician. So, it is very important to preserve the diagnostic value of images. For instance, artifacts in a patient's diagnostic image due to image watermarking may cause errors in diagnosis and treatment, which may lead to possible life-threatening consequences [40]. For this restrictions embedding watermarks and image compressions must not distort and degrade the quality of images. Therefore, minimum Peak Signal to Noise Ratio (PSNR) of 40–50 db is advised by previous works. More importantly, watermarks should survive the standard image processing like low pass filtering (LPF) which removes noise and improves visual quality; and High Pass Filtering (HPF) that enhances the information content [1].

Watermarking systems can be characterized by a number of defining properties. The relative importance of each property is dependent on the requirements of the application and the role the watermark will play. In fact,even the interpretation of a watermark property can vary with the application. These watermarking properties can be used as performance criteria to evaluate watermarking schemes and provide some favorable guidance for the design of watermarking schemes with certain application background. In this Section, some watermarking performances are introduced [62].

According to [63] there have not a complete evaluation system about digital watermark. Because the uniform description of the performance, the test methods, the method of attack, the standard test procedure have not been established. How to judge the performance of a watermarking system is very important. The evaluation of a watermarking system is associated to the watermark robustness and imperceptibility. So, watermark should with stand any alteration of the watermarked image, whether it is unintentional or malicious. This is a key requirement. Researchers currently do not appear to have found a watermarking method that is 100 % robust. If the original watermark(s) is (are) removed the image quality must degrade so the image will be unusable according to there requirements of the application.

4.5.1 Imperceptibility or Transparency

Imperceptibility of digital watermark is also known as transparency. The watermark signal should be imperceptible to the end user who is listening to or viewing the host signal. This means that the perceived quality of the host signal should not be distorted by the presence of the watermark [64]. Ideally, a typical user should not be able to differentiate between watermarked and un-watermarked signals.

Transparent assessment is divided into the subjective evaluation and objective evaluation [63]:

- *Subjective evaluation*: refers to the human visual effect as the evaluation criterion. That is given by the observer to judge the image quality. The mathematical models can not be used quantitatively to describe the image quality, and it is too time consuming. The application of subjective evaluation is very limited, so we want to use the objective, stable mathematical model to express the image quality.
- *Objective evaluation*: Image objective evaluation method used mathematical model and computed similarity between the image distortion and the original image (or distortion) and quantized evaluation scores. MSE (mean squared error) and peak signal to noise ratio (PSNR) is widely used as an evaluation criteria of watermark imperceptibility [65]. The mean-square error between any signals S and S^* is defined as:

$$MSE(S, S^*) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} ||S_{i,j} - S^*_{i,j}||^2 \tag{4.1}$$

when S and S^* are identical, then MSE (S, S) = 0. A related distortion measure is the peak signal-to-noise ratio (PSNR), measured in decibels (dB). The PSNR is defined as follows:

$$PSNR = 10log_{10}(\frac{MAX_i^2}{MSE}) = 20log_{10}(\frac{MAXi}{\sqrt{MSE}}) \tag{4.2}$$

where $MAX_i = \max(S^*_{i,j}, 1 \leq \mathrm{i}, \mathrm{j} \leq \mathrm{m})$. The higher the PSNR ($S$, S^*), the less distortion between S and S^*. If the signals are identical, then PSNR (S, S^*) $= \infty$.

Because the mean square error and peak signal to noise ratio reflect the difference of the original image and restore the image in whole, not reflect in the local of images, they can not reflect the human visual characteristics. There are other variety of evaluation methods based on the above principle, such as: assessing the quality of enhanced images based on human visual perception, image quality assessing model by using neural network and support vector machine, gradient information based image quality assessment, image quality assessment method based on contrast sensitivity, and so on [63].

4.5.2 Robustness

The watermarking system should be robust enough to detect and extract the watermark similar to the original one. Different types of alteration (distortions) which are known as attacks can be performed to degrade the image quality [66]. The distortions are limited to those factors which do not produce excessive degradations in the image otherwise the transformed object would be unusable. These distortions also introduce degradation on the performance of the watermark extraction algorithm. To test for the robustness of the methods or a combination of the methods, an attack is performed intentionally on a watermarked document in order to destroy or degrade the quality of the hidden watermark. According to [67] we can classify attacks into the following groups:

- Simple attack: Simple attack aim to modify the entire cover image without extraction of the watermark. Examples of such attacks include compression, addition of noise and editing.
- Detection-disabling or Geometric attacks: The objective of these attacks is not removal of watermark but to damage the watermark. Watermark still exist but not detectable. Normally, they make some geometric distortions such as zooming, rotating the object, cropping or pixel permutation, shift in spatial/temporal direction and removal/insertion etc.
- Ambiguity attacks: These attacks try to deceive the detection process through false watermarked data. In this attack many additional watermarks to discredit the original owner so that it is not clear that which watermark is the original watermark.
- Removal attacks: The objective of these attacks is to detect and then remove the embedded watermark without harming the cover media. Examples include collusion attack, denoising, use of the conceptual cryptographic weakness of the watermarking system, quantization, averaging, filtering, printing and scanning.

Robustness metrics include the correlation measure and bit error rate. The correlation measure usually use NC (normalized correlation) coefficient as the similarity measure of extracted watermark and the original watermark. NC equation is defined as following:

$$Corr = \frac{\sum_{n=1}^{N-1} (w_n' - w^{-'})(w_n - w^{-})}{\sqrt{\sum_{n=1}^{N-1} (w_n' - w^{-'})^2 \sum_{n=1}^{N-1} (w_n - w^{-})^2}} \tag{4.3}$$

where $w^{-'}$ and w^{-} indicate respectively the averages of the watermark bit sequence w_n' and w_n. This correlation value measures the similarity between two strings and varies between 1 (orthogonal sequence) and $+1$ (the same sequence).

The trade-off between the imperceptibility and robustness is one of most serious challenges in digital watermarking system, in particular the medical imaging.

4.6 Conclusions and Future Directions

Bio-inspiring has increasingly gained attention in watermarking research. The main purpose of this chapter was to present the computational intelligence and medical watermarking research communities some of the state-of-the-art and recent advances in BI applications to medical image watermarking, and to inspire further research and development on new applications and new concepts in new trend-setting directions. In medical applications, because of their diagnostic value, it is very important to maintain the quality of images. There is always a trade-off between the imperceptibility and robustness. It is one of most serious challenges in digital watermarking system, in particular the medical imaging. For instance, artifacts in a patient's diagnostic image due to image watermarking may cause errors in diagnosis and treatment, which may lead to possible life-threatening consequences . For this matter, the development of a new algorithm that can satisfy both imperceptibility and robustness is needed. Improvements in performance of watermarking schemes can be obtained by several methods. One way is to make use of Bio-inspiring techniques. The learning capability of all BI techniques is used to make an optimal trade-off between imperceptibility and robustness through effective selection of threshold. This trade-off has much more importance in case of sensitive imagery such as medical imagery. BI techniques can be used also to find which are watermarks embedding places. These parameters are optimally varied to achieve the most suitable places for watermark embedding that achieve best values for both imperceptibility and robustness.

Different BI techniques such as restricted Boltzmann machine, deep belief network, rough sets, swarm intelligence, artificial immune systems and support vector machines, could be successfully employed to tackle various problems in watermark embedding and extraction procedure of medical image watermarking. Also a hybrid system of different BI techniques can be built to combine the advantages of participating techniques in the hybrid system.

References

1. Fakhari, P., Vahedi, E., Lucas, C.: Protecting patient privacy from unauthorized release of medical images using a bio-inspired wavelet-based watermarking approach. Digit. Signal Process. **21**, 433–446 (2011)
2. Gunjal, B.L., Mali, S.N.: ROI based embedded watermarking of medical images for secured communication in telemedicine. Int. J. Comput. Commun. Eng. **68**, 293–298 (2012)
3. Kallel, M., Lapayre, J.C., Bouhlel, M.S.: A multiple watermarking scheme for medical image in the spatial domain. GVIP J. **7**(1), 37–42 (2007)
4. Horng, J.-T., Wu, L.-C., Liu, B.-J., Kuo, J.-L., Kuo, W.-H., Zhang, J.-J.: An expert system to classify microarray gene expression data using gene selection by decision tree. Expert Syst. Appl. **36**(5), 9072–9081 (2009)
5. Olanrewaju, R.F., Khalifa, O., Abdul Latif, K.N.: Computational intelligence: its application in digital watermarking. Middle-East J. Sci. Res. (Math. Appl. Eng.). **13**, 25–30 (2013)
6. Beckenkamp, F.: A component architecture for artificial neural network systems. A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in University of Constance, Computer and Information Science (2002)
7. Moore, A.W.: Reinforcement learning: a survey. J. Artif. Intell. Res. **4**, 237–285 (1996)
8. Haykin, S.: Neural Networks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River (1999)
9. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995)
10. Motsinger, Alison A., Dudek, Scott M., Hahn, Lance W., Ritchie, Marylyn D.: Comparison of neural network optimization approaches for studies of human genetics. EvoWorkshops **2006**, 103–114 (2006)
11. Ghosh, A., Dehuri, S.: Evolutionary algorithms for multi-criterion optimization: a survey. Int. J. Comput. Inf. Sci. **2**(1), 38–57 (2004)
12. Fogel, D.B.: Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, 2nd edn. IEEE Press, Piscataway (1999)
13. Andres, C., Reyes, P., Sipper, M.: Evolutionary computation in medicine: an overview. Artif. Intell. Med. **19**, 1–23 (1999)
14. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
15. Koza, J.R.: Genetic Programming. The MIT Press, Cambridge (1992)
16. Back, T.: Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms. Oxford University Press, New York (1996)
17. Fogel, L.J., Owens, A.J., Walsh, M.J.: Artificial Intelligence Through Simulated Evolution. Wiley, New York (1967)
18. Smolinski, T.G., Milanova, M.G., Hassanien, A.E.: Applications of Computational Intelligence in Biology: Current Trends and Open Problems, Studies in Computational Intelligence, vol. 122. Springer, Berlin (2008)
19. Purves, W., Orians, G., Heller, C.: Life, the Science of Biology, Sinauer, Sunderland (1995)
20. Beasley, D., Bully, D.R., Martinz, R.R.: An overview of genetic algorithms: part 1, fundamentals. Univ. Comput. **15**(2), 58–69 (1993)
21. Juang, C.F.: A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. IEEE Trans. Syst. Man Cypern. Part B: Cybern. **34**(2), 997–1006 (2004)
22. Imade, H., Morishita, R., Ono, I., Ono, N., Okamoto, M.: A framework of grid-oriented genetic algorithms for large-scale optimization in bioinformatics. In: The Proceeding of the 2003 Congress on Evolutionary Computation, vol. 1, pp. 623–630 (2003)
23. Jiao, C.Y., Li, D.G.: Microarray image Converted database—genetic algorithm application in bioinformatics. Int. Conf. BioMed. Eng. Inform. **1**, 302–305 (2008)
24. Glen, R.C., Payne, A.W.R.: A genetic algorithm for the automated generation of molecule within constraints. J. Comput.-Aided Mol. Des. **9**, 181–202 (1995)

25. Venkatasubramanian, V., Chan, K., Caruthers, J.M.: Evolutionary design of molecules with desired properties using the genetic algorithm. J. Chem. Inf. Comp. Sci. **35**, 188–195 (1995)
26. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Corporation Inc., Reading (1989)
27. Fleischer, M.: Foundations of Swarm Intelligence: From Principles to Practice, CSHCN; TR 2003–5 (2005)
28. Blum, Ch., Merkle, D.: Swarm Intelligence Introduction and Applications, Natural Computing Series. Springer, Berlin (2008)
29. Felix, T.S., Kumar, M.T.: Swarm Intelligence Focus on Ant and Particle Swarm Optimization. I-Tech Education and Publishing, Vienna (2007)
30. Dorigoa, M., Blumb, Ch.: Ant colony optimization theory: a survey. Theor. Comput. Sci. **344**(2005), 243–278 (2005)
31. Dorigoa, M., Stutzle, T.: Ant Colony Optimization. MIT Press, Cambridge (2004)
32. Pant, M., Thangaraj, R.: Particle swarm optimization: performance tuning and empirical analysis. Stud. Comput. Intell. **203**, 101–128 (2007)
33. Blum, Christian: Ant colony optimization: introduction and recent trends. Phys. Life Rev. **2**(4), 353–373 (2005)
34. Parpinelli, R.S., Lopes, H.S.: New inspirations in swarm intelligence: a survey. Int. J. Bio-Inspired Comput. **3**(1), 1–16 (2011)
35. Kentzoglanakis, K., Poole, M.: A swarm intelligence framework for reconstructing gene networks: searching for biologically plausible architectures. IEEE/ACM Trans. Comput. Biol. Bioinf. **29**, 358–371 (2011)
36. Das, S. et al.: Swarm intelligence algorithms in bioinformatics. Stud. Comput. Intell. **94**, 113–147 (2008)
37. Kennedy, J., Eberhart, R., Shi, Y.: Swarm Intelligence, pp. 1931–1938. Morgan Kaufmann Academic Press, San Francisco (2001)
38. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
39. Nyeem, H., Boles, W., Boyd, C.: Review of medical image watermarking requirements for teleradiology. Digit. Imaging **26**(2), 326–343 (2013)
40. Dharwadkar, N.V., Amberker, B.B., Panchannavar, P.B.: Reversible Fragile Medical Image Watermarking with Zero Distortion. In: Computer and Communication Technology (ICCCT), pp. 248–254. Allahabad, Uttar Pradesh (2010)
41. Umaamaheshvari, A., Thanushkodi, K.: Performance analysis of watermarking medical images. Life Sci. J. **10**(1), 2653–2660 (2013)
42. Fndk, O., Babaoglu, I., Ulker, E.: A color image watermarking scheme based on artificial immune recognition system. Expert Syst. Appl. **38**, 1942–1946 (2011)
43. Aslantas, V.: A singular-value decomposition-based image watermarking using genetic algorithm. Int. J. Electron. Commun. **62**(5), 386–394 (2008)
44. Pal, K., Ghosh, G., Bhattacharya, M.: Biomedical image watermarking in wavelet domain for data integrity using bit majority algorithm and multiple copies of hidden information. Am. J. Biomed. Eng. **2**(2), 29–37 (2012)
45. Bhatnagar, G., Raman, B.: A new robust reference watermarking scheme based on DWT-SVD. Comput. Stand. Interfaces **31**(5), 1002–1013 (2009)
46. Poonkuntran, S., Rajesh, R.S., Eswaran, P.: Analysis of difference expanding method for medical image watermarking. In: International Symposium on Computing, Communication, and Control (ISCCC 2009), vol. 1, pp. 31–34 (2011)
47. Fotopoulos, V., Stavrinou, M.L., Skodras, A.N.: Medical Image Authentication and Self-Correction through an Adaptive Reversible Watermarking Technique. In: 8th IEEE International Conference on Bio Informatics and Bio Engineering, Athens, pp. 1–5 (2008)
48. Kallel, I.M., Bouhlel, M.S., Lapayre, J.C.: Improved tian method for medical image reversible watermarking. GVIP J. **7**(2), 2–5 (2007)
49. Frank, Y.S., Wu, Y.T.: Robust watermarking and compression for medical images based on genetic algorithms. Inf. Sci. **175**, 200216 (2005)

50. Motwani, M.C.: Third generation 3D watermarking: applied computational intelligence techniques. A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering. University of Nevada, Reno (2011)
51. Olanrewaju, R.F., Khalifa, O.O., Hashim, A.H., Zeki, A.M., Aburas, A.A.: Forgery detection in medical images using complex valued neural network (CVNN). Aust. J. Basic Appl. Sci. **5**(7), 1251–1264 (2011)
52. Oueslati, S., Cherif, A., Solaimane, B.: Adaptive image watermarking scheme based on neural network. Int. J. Eng. Sci. Technol. **3**(1), 748–756 (2011)
53. Ramya, M.M., Murugesan, R.: Joint image-adaptive compression and watermarking by GA-based wavelet localization: optimal trade-off between transmission time and security. Int. J. Image Process. **6**(6), 478–487 (2012)
54. Usman, I., Khan, A., Ali, A., Choi, T.S.: Reversible watermarking based on intelligent coefficient selection and integer wavelet transform. J. Innov. Comput. Inf. Control **5**(12), 46754682 (2009)
55. Aleisa, E.A.: A secure transmission of medical images over wireless networks using intelligent watermarking. Life Sci. J. **10**(2), 2438–2444 (2013)
56. Soliman, M.M., Hassanien, A.H., Ghali, N., Onsi, H.: An adaptive watermarking approach for medical imaging using swarm intelligent. Int. J. Smart Home **6**(1), 37–50 (2012)
57. Soliman, M.M., Hassanien, A.H., Onsi, H.: An adaptive medical images watermarking using quantum particle swarm optimization. In: 35th International Conference on Telecommunications and Signal Processing (TSP), pp. 735–739, Prague, Czech Republic (2012)
58. Soliman, M.M., Hassanien, A.H., Onsi, H.: An adaptive watermarking approach based on weighted quantum particle swarm optimization. Submitted to Neural Computing and Applications (2013)
59. Soliman, M.M., Hassanien, A.H., Onsi, H.: The way of improving PSO performance: medical imaging watermarking case study. In: A Joint Conference of the 8th International Conference on Rough Sets and Current Trends in Computing (RSCTC 2012), pp. 237–242, Chengdu (2012)
60. Sun, J., Feng, B., Xu, W.B.: Particle swarm optimization with particles having quantum behaviour. In: IEEE Proceedings of Congress on Evolutionary Computation, pp. 325–331 (2004)
61. Benioff, P.: The computer as a physical system: a microscopic quantum mechanical hamiltonian model of computers as represented by turing machines. J. Stat. Phys. **22**(5), 563–591 (1980)
62. Zhou, X., Wang, S., Xiong, Sh: Attack model and performance evaluation of text digital watermarking. J. Comput. **5**(12), 1933–1941 (2010)
63. Zhang, X., Zhang, F., Xu, Y.: Quality evaluation of digital image watermarking, ISNN 2011, Part II, LNCS 6676, pp. 241250 (2011)
64. Gordy, J.D.: Performance evaluation of digital watermarking algorithms. A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering, faculty of graduate studies. University of Calgary, Alberta (2000)
65. Abdallah E.A.: Robust digital watermarking techniques for multimedia protection. A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science, Concordia University, Montreal, Quebec, Canada (2009)
66. Pal, K., Ghosh, G., Bhattacharya, M.: Biomedical image watermarking in wavelet domain for data integrity using bit majority algorithm and multiple copies of hidden information. Am. J. Biomed. Eng. **2**(2), 29–37 (2012)
67. Jan, Z.: Intelligent image watermarking using genetic programming. A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science, National University of Computer and Emerging Sciences Islamabad

Chapter 5
Efficient Image Authentication and Tamper Localization Algorithm Using Active Watermarking

Sajjad Dadkhah, Azizah Abd Manaf and Somayeh Sadeghi

Abstract Due to the increasing number of forged images and possibilities of effortless digital manipulations, certain organizations have started to focus on approaches that help preserve their digital data integrity. Image authentication systems try to accurately verify the integrity of digital images, of which digital watermarking is known to be one of the most precise techniques in authenticating the originality of digital images. This chapter has presented a novel image authentication system with accurate tamper localization ability. In the proposed algorithm a 16-bit watermark key has been created from each block of pixels in a host image. The generated 16-bit watermark key will be embedded into the host image by utilizing a fragile watermarking algorithm. The security of the watermarking algorithm will be ensured by using the proposed block cipher algorithm, which encrypts the user key and watermarking algorithm. The proposed tamper detection algorithm conducts two comprehensive comparisons, to ensure the accuracy of the results. The high quality watermarks and powerful tamper detection approaches, along with less computational complexity are the main advantages of the proposed image authentication system, which makes it suitable for real-time application. Several tampering experiments have been conducted to examine the proposed algorithm. The experiment

S. Dadkhah
Faculty of Computing, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia
e-mail: dsajjad2@live.utm.my

A. A. Manaf (✉)
Advanced Informatics School, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia
e-mail: azizah07@ic.utm.my

S. Sadeghi
Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
e-mail: ssomayeh@siswa.um.edu.my

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_5, © Springer-Verlag Berlin Heidelberg 2014

results have clearly proved that, the proposed method is not only efficient, but also very accurate in detecting different types of digital image manipulations, such as, small region tampering, cropping tampering and bit tampering.

5.1 Introduction

The fast development of digital technology and image editing tools have led to a wide range of professional digital images manipulations, which first precisely appear as original images, but are later proven to be forged. At present, image authentication and tampering detection systems are struggling to confirm the valid originality of the digital images that are presented in certain situations. In significant situations, like evidence in court of law, a small amount of ambiguity can change the path of judgments. In some cases, such as, tampered images on the news or in other forms of social media could cause some serious problems, which could also jeopardies a person's life [1, 2].

As a result, the technological world is in a desperate need for some reliable and efficient image authentication schemes. In order to have an impeccable digital image authentication system, certain aspects have to be considered to preserve the authenticity and integrity of digital images. Verifying the originality of the digital image alone is not enough, locating the tampered areas makes the authentication procedure more complete, by separating the parts of the image that are reliable from the parts that are forged. In special cases like military maps and medical imagery, this aspect is considered to be very crucial [3, 4].

The best approach to validate the integrity and authenticity of the digital image is by using the digital watermarking. Digital watermarking is a procedure of embedding significant information, such as, binary values into digital images, videos or audio waves. The digital watermark can appear on the surface of the images, which is known as visible watermarking, and could be used as owner copy rights. However the visible watermarking can be forged or omitted with a simple editing technique, which makes it inappropriate for the purpose of tamper detection [5, 6]. On the other hand, invisible watermarking embeds meta-data into a digital media in a way that it becomes imperceptible by naked eyes, and therefore is used for several purposes, such as, authentication and tamper localization [7, 8].

It has been declared by established practices that an efficient watermarking system for authenticating purpose should satisfy the following features, and has to be carefully designed to make a trade-off between properties [9, 10]. In imperceptibility, the watermarked information and the original data should be perceptually indistinguishable; however, in this concept, fidelity and quality are important measurements of perceptiveness. Fidelity is a similarity between signals before embedding and after embedding the meta-data. A high-fidelity means the host is very similar to the original after the watermarking process. On the other hand, a low-fidelity means that, the features are dissimilar, or distinguishable from the original. At the same time, a high-quality image simply looks good and contains a high Peak Signal-to-Noise

Ratio (PSNR) value [11]. In this chapter the quality of the watermarked image has been measured by calculating the PSNR.

Invisibility is the next factor for a perfect tamper detection, which uses watermarking technique; the embedded information should not damage the value of the image and should be imperceptible. Security is the third factor and it is mainly an important issue for authentication a watermarking system [12, 13]. The aspect of security is covered in our proposed algorithm by adding an encryption key to the watermarking procedure. This action makes it impossible for attackers to forge original images, even though the forgers might be aware of the functionality of the algorithm.

The invisible watermarking is categorized into three different classifications: the fragile watermarking, the semi-fragile, and the robust watermarking [14, 15]. The key feature of the robust watermarking is its resistance against more planned attacks, which means that, this watermarking method can resist specific signal processing operations. This ability makes the robust watermarking suitable for copyright protection applications [16, 17]. The semi-fragile watermarking is de-signed to detect any unauthorized modification, while allowing some image processing operations. Both, the semi-fragile watermarking and fragile watermarking are developed to be used for image authentication, but the semi-fragile leans more towards tamper localization, and discovering the extra modifications carried out by attackers. As against the fragile watermarking, the semi-fragile watermarking has been considered weak against complicated attacks, but the compression attacks like JPEG are detectable by such watermarking technique [18, 19].

The fragile watermarking systems are intended to sense any potential manipulations that influence the watermarked pixel values. The fragile watermarking has been considered weak and sensitive against any kind of attack, and this characteristic makes this kind of watermarking method suitable for authentication and tamper localization applications [15]. The aim of the fragile watermarking is to produce a verification message, which is made from specific image characteristic and embed into the host image. The message created by the watermarking procedure is known as watermarked, and has be to extracted from watermarked image to be compared to the new message created by tamper detection or the authentication procedure. Certain watermarking procedures, other than the authentication and the tamper localization have the ability of recovering the tampered area.

Generally, Least Significant Bit (LSB) watermarking is used to prevent the image from having a noticeable change, and to preserve the good quality of the host image. The primary element of the authentication procedure is dividing the original image into blocks of pixels. The sizes of the blocks are changeable, depending on the size of the block selected by the tamper detection procedure and its accuracy and changes in time consumption [20]. In this chapter a novel block-based watermarking method for image authentication and tamper localization is proposed. In order to evaluate the image authentication method, various tampering experiments can be conducted. Tamper detection methods with the ability of detecting collage attacks can achieve robustness. Collage attacks include deleting some portion of the image or copying and pasting some object/s inside the host image.

The remaining of the chapter is organized as follows: Sect. 5.2 briefly reviews some existing methods on digital image authentication and image tamper detection. Section 5.3 introduces the structure of the proposed authentication system with a detailed discussion on the proposed watermarking and embedding technique; the Sect. 5.3 discusses the novel two-level image tamper detection proposed by this chapter. Section 5.4 outlines several experiments that were conducted in order to examine, the tamper detection rate, watermarked image perceptibility, and tamper localization, which has been conducted by determining the false positive and negative of the proposed tamper localization algorithm; furthermore, the results of comparative analysis between the proposed authentication system and several current digital image authentication approaches have also been presented. Finally, Sect. 5.5 concludes by highlighting the contribution of the proposed algorithm.

5.2 Related Work

Over the recent years, the topic of image forgery and tamper detection has drawn the attention of many organizations to overcome these issues. The importance of saving the integrity and confidentiality of digital data has made a huge impact on several researches as well. In this section we have reviewed some of the existing fragile and non-fragile watermarking algorithms for tamper detection.

5.2.1 Digital Image Authentication Methods

Wong [21] has proposed an authentication algorithm using the LSB embedding technique. This is the fragile method, where the image is divided into several non-overlapping blocks of the same sized (8 × 16) pixels, after which seven of the Most Significant Bit (MSB) of each pixel inside the blocks are extracted and converted by the hash function. The XOR function is then applied between the blocks of message and the host image; and the result is encrypted by public-key and embedded into the LSB of each pixel inside blocks, within the host image. Fridrich [22] has adopted Wongs technique, and proposed a tamper detection method using the fragile watermarking, with the ability of tamper localization. The highlight of his method is detecting the Vector Quantization Attack, where he was able to locate the tamper zones with an accuracy of 83 %.

Zhang and Wang [23] have proposed a tamper detection method using 3 Least Significant Bit (3LSB) embedding technique. In this method the five most significant bits (5MSBs) of pixels are converted to thirty one validation bits; and any changes on the first five significant bits would affect their authentication bits. The embedding procedure of this fragile watermarking technique takes place in the last three significant bits of each pixel. The content of the watermark in this method is generated by dividing the authentication bits into a certain number of subsets using his proposed formula. This method is able to locate the tamper regions, as long as the tamper areas are not too wide. The PSNR of the watermarked image reaches the satisfactory amount of 37 dB.

Ohkita et al. [24] have proposed a fragile watermarking method by adopting Zhang and Wangs method. Their proposed LSB watermarking technique has tamper localization ability, but is limited in the sense of tamper ratio. Lin et al. [25] have proposed a fragile hierarchical digital watermarking for image authentication, which took advantage of counting the number of bits along with finding the intensity of the relationship between the blocks. Chang et al. [26] have verified that, Lins method was not able to fully detect all kinds of tamper, such as, deletion by examining a few of the scanned attacks. Zhang and Wang [27] have proposed a hierarchical image tamper detection method, which uses both, the block and pixel data to construct the watermarking information. Block features are used to authenticate the watermarked image, and pixel features to locate the tamper regions. This method was vulnerable to specific tampering attacks like collage attacks, etc.

Later Lee and Shinfeng [28] have proposed a fragile watermarking method, which divides the image into two non-overlapping blocks, and embeds the watermarking information into the same pixel on both divided blocks. The advantage of this method is that, it is possible to recover the watermark, in case it is destroyed. The dual watermarking method proposed by Lee was weak against any bit tampering attack to the first five significant bits of pixels. This flaw was emphasized by Chaluvadi and Prasad [29] a year later, where they proposed a method, which created the watermark information by calculating the average intensity of the pixels inside each block and embedding the three least significant bits of each pixel.

Lin et al. [30] proposed a digital image authentication algorithm which divides the original image to blocks of size 16 × 16 pixels. Their proposed algorithm performed Discrete Cosine Transform (DCT) to each divided blocks, then the coefficient of the converted pixels are embedded with the generated watermark information's. Their proposed algorithm achieved 90 % tamper detection rate for digital images with slight compressions. A semi-fragile tamper detection algorithm proposed by Ho et al. [31] utilized Pinned Sine Transform (PST) domain. In their proposed algorithm, the digital image is divided into blocks of size 8 × 8 pixels, the pinned and boundary field of all the divided blocks are generated by utilizing PST. The embedding procedure for their proposed algorithm, locates the high frequency coefficients of each divided block in the pinned field.

Moreover, Zhao et al. [32] proposed two tamper detection algorithm by using active and passive techniques, the active watermarking proposed by them performed the embedding operation in Slant Transform (SLT) domain. Their proposed algorithm divided the original image into blocks of size 8 × 8 pixels and extracts the 7 most significant bits of the pixel to generate the watermark information. The semi-fragile watermarking algorithm proposed by Zhao et al. [32] achieved 98 % tamper detection rate for copy-move tampering attacks.

To the best of our knowledge, the recent tamper detection algorithm which utilizes watermarking techniques, are not designed to be efficient in localizing small regions. However, Rosales et al. [33] proposed two tamper detection algorithm for digital image recovery, their proposed algorithm utilized Integer Wavelet Transform (IWT). Their proposed algorithm embeds the generated watermarked by utilizing quantization; however their proposed algorithm could perform accurately if the proper threshold is selected. Tong et al. [34] proposed tamper detection algorithm with capability of self-recovery, their proposed algorithm is superior when the tampered area is relatively large.

Dadkhah et al. [35] have proposed a two level image tampering detection schemes by using the 3 Least Significant Bit (3LSB) watermarking technique. Their proposed technique improved Chaluvadis tamper detection rate by 40 %. The digital image authentication technique proposed by them has been very accurate in localizing the tamper region and meeting the satisfactory PSNR value of 38.2 dB. Sheng and Tu [36] have proposed a fragile watermarking method for tamper detection, based on Lees [28] method; where they used block features, such as, the smooth and non-smooth blocks for image authentication. Yeo and Lee [37] have proposed the block based image authentication system using a reversible watermarking technique, where they achieved 97 % as the image authentication rate, but their method could not detect small region tampers. Furthermore, the proposed tamper detection algorithm in this research improves the tamper detection rate and watermark quality of our previous work [35].

5.3 The Proposed Algorithm

In this section of the research, we have reviewed the main procedure of the proposed image authentication and tamper detection algorithm. The main procedure of the proposed watermarking algorithm starts with a secret key created by user to preserve the security of the proposed algorithm as illustrated in Fig. 5.1. The concept of adding a secret key to the algorithm is for eliminating any possible threat of an attack on the algorithm itself. Thus, a 16-bit secret key is generated by the user and embedded into the two random blocks of size 4×4 pixels, inside the host image.

As illustrated by the Fig. 5.1 the watermarking procedure starts by dividing the original image into non-overlapping blocks of 4×4 pixels in size. As described in the previous section, the fragile watermarking is very sensitive to sense any small or large amount of manipulations. The primary structure of the proposed tamper detection technique in this chapter, utilize the same outline but completely distinctive methodology as our previous tamper detection technique [35], in which the watermarking procedure took advantage of the parity check, which is counted the number of bits inside the image pixels and the embedding procedure took place in the 3 least significant bits (3LSB) of each pixel value, which affected the perceptibility of the watermarked image. The length of the watermark key for each block was 12 bit with size of each block was 2×2 pixels. In this chapter a new watermarking procedure has been proposed, which still takes the advantages of the parity check concept. The length of the watermarked key for each block has been extended to 16 bit per block,

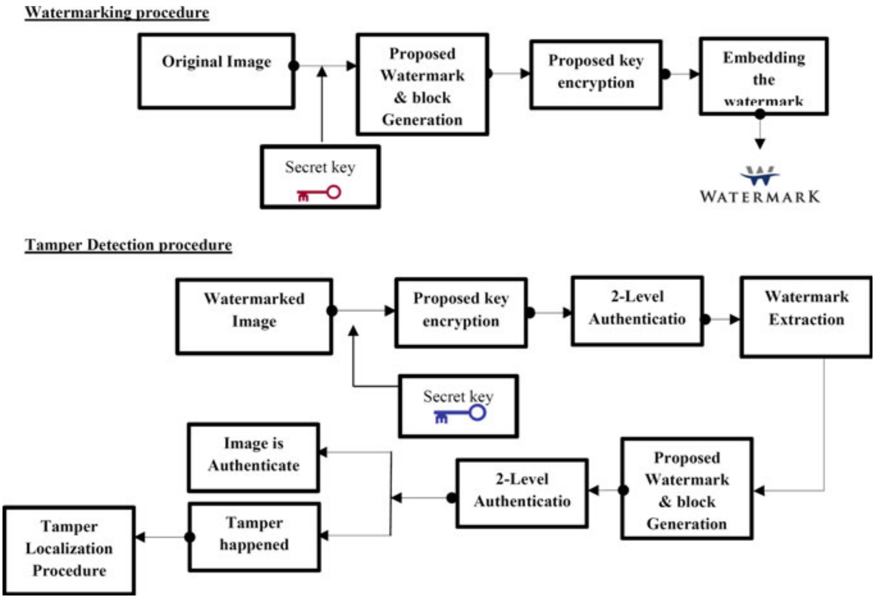


Fig. 5.1 General procedure of proposed watermarking and authentication

and the embedding procedure takes place in the LSBs of the pixels inside the divided blocks.

The proposed tamper detection procedure as shown in Fig. 5.1 includes a two-level comparison, where the first stage of the authentication procedure starts with checking the secret key inserted by the user. The secret key is embedded in the watermarking procedure, so if the new key matches with the embedded key, the authentication procedure can continue scanning the whole image. The tamper detection procedure includes generating the watermark in the same way as the watermarked key was extracted before, and is then compared to the new generated watermark. If both the 16-bit watermark keys were equal, then the watermarked image is authenticated, or else the image is regarded as forged, and as illustrated, the tamper regions will be located. The proposed authentication technique has been designed to be compatible to gray-scale and RGB images.

5.3.1 The Proposed Watermarking Scheme

In this section of the chapter the procedure of the proposed watermarking method is described step by step. As illustrated in Fig. 5.2, the original image is divided into non overlapping blocks of sizes 4 × 4 pixels, and each block comprises 16 distinct pixels. The first step of the watermarking procedure is receiving a 16-bit length

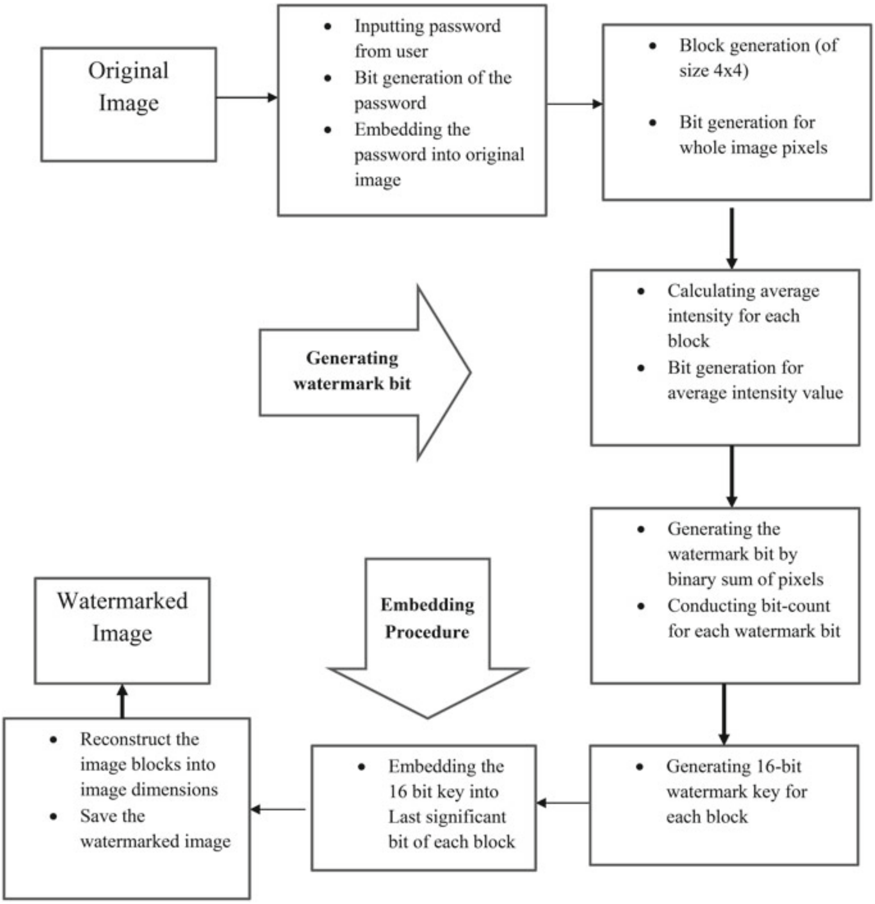

Original Image
Inputting password from user
Bit generation of the password
Embedding the password into original image
Block generation (of size 4x4)
Bit generation for whole image pixels
Generating watermark bit
Calculating average intensity for each block
Bit generation for average intensity value
Generating the watermark bit by binary sum of pixels
Conducting bit-count for each watermark bit
Embedding Procedure
Watermarked Image
Reconstruct the image blocks into image dimensions
Save the watermarked image
Embedding the 16 bit key into Last significant bit of each block
Generating 16-bit watermark key for each block

Fig. 5.2 General flowchart of proposed watermarking algorithm

password from a user as a secret key; and this "key" is for encrypting the algorithm to stop attackers, who have complete knowledge about the proposed watermarking algorithm. The secret key has been embedded into the least significant bit of the two random blocks. In this phase, procedure of the proposed encryption key is presented. The flowchart illustrated in Fig. 5.3 is the detailed explanation of the proposed encryption technique.

However, as illustrated in Fig. 5.2 after completion of the proposed embedding procedure, all the pixels with new values revert to its decimal format. All the pixel blocks are restored to their original locations, to form the digital image equipped with proposed watermark. The process of producing watermark and embedding procedure will be described in following sub chapters.

The Proposed Key Encryption Algorithm

In order to secure the algorithm from attackers, who knows the structure of the proposed watermarking algorithm, an encryption algorithm has been proposed to encrypt the key entered by user and by the watermark itself; the proposed encryption algorithm takes advantage of simple block ciphers, which create high confusion for attacker. The proposed key encryption steps are as follows:

1. Convert the password to 64 bit binary, as illustrated in Fig. 5.3.
2. Shift all the rows by 6 places.
3. Swap rows 8 and 6.
4. Swap rows 4 and 2.
5. Apply XOR function between new shifted matrix and first pass matrix.

However, as outlined in Fig. 5.3 there is limitation of the key length. The maximum length of the key depends on the number of watermark bit specified for the key embedding process. In the proposed key encryption algorithm, the number of key embedding procedure is specified as 4, as mentioned before each watermark key has the length of 16 bit, thus the total bit length specified for the embedding procedure is 64 bit. As Fig. 5.2 is illustrated the first phase of encryption is to convert the entered key into 64-bit binary stream.

Moreover, after the conversion of the watermark keys, as Fig. 5.3 illustrated, a combination of different rows and columns swings will be conducted to create an illusion. The proposed movement of the rows and columns is inspired by block-based encryption techniques. The main goal of the block-based encryption technique is to rearrange normal pattern of the numbers in a way that the new numbers is totally different from the first value. As described before, greater number of swings will create harder situation for the attacker to crack the password.

In addition, as Fig. 5.3 is illustrated the next step after rows swapping stage, a “XOR” function will be performed between the new shifted values and the first watermarked key. Different encryption algorithms use different mathematical operation such as “XOR” and “OR” operation. However, the key concept of using logical operations such as “XOR” is to produce a completely new value. As Fig. 5.4 is illustrated, in this phase of encryption operation, the best logical operation for producing different value is“XOR” because both matrices are containing binary values with different amount of “0” and “1”. Furthermore, any other logical operation such as “OR" and “AND” would create less ambiguity for the attacker.

Furthermore, the next phase of proposed encryption technique is to shift all the columns of the new converted matrix by 6 places up, and then apply the same swap movement but this time on the column of the shifted matrix. The next step is to convert the result of the last operation into an array of size 4 × 4 to make the key encryption matrices ready for watermark embedding procedure. As mentioned in previous section, by adding the key encryption algorithm, the proposed tamper detection algorithm will be robust to different type of collage attacks. The importance of key encryption is highlighted when the attacker know the structure of the victims algorithm. The proposed encryption algorithm takes advantage of simple block ciphers, which create high confusion for attacker.

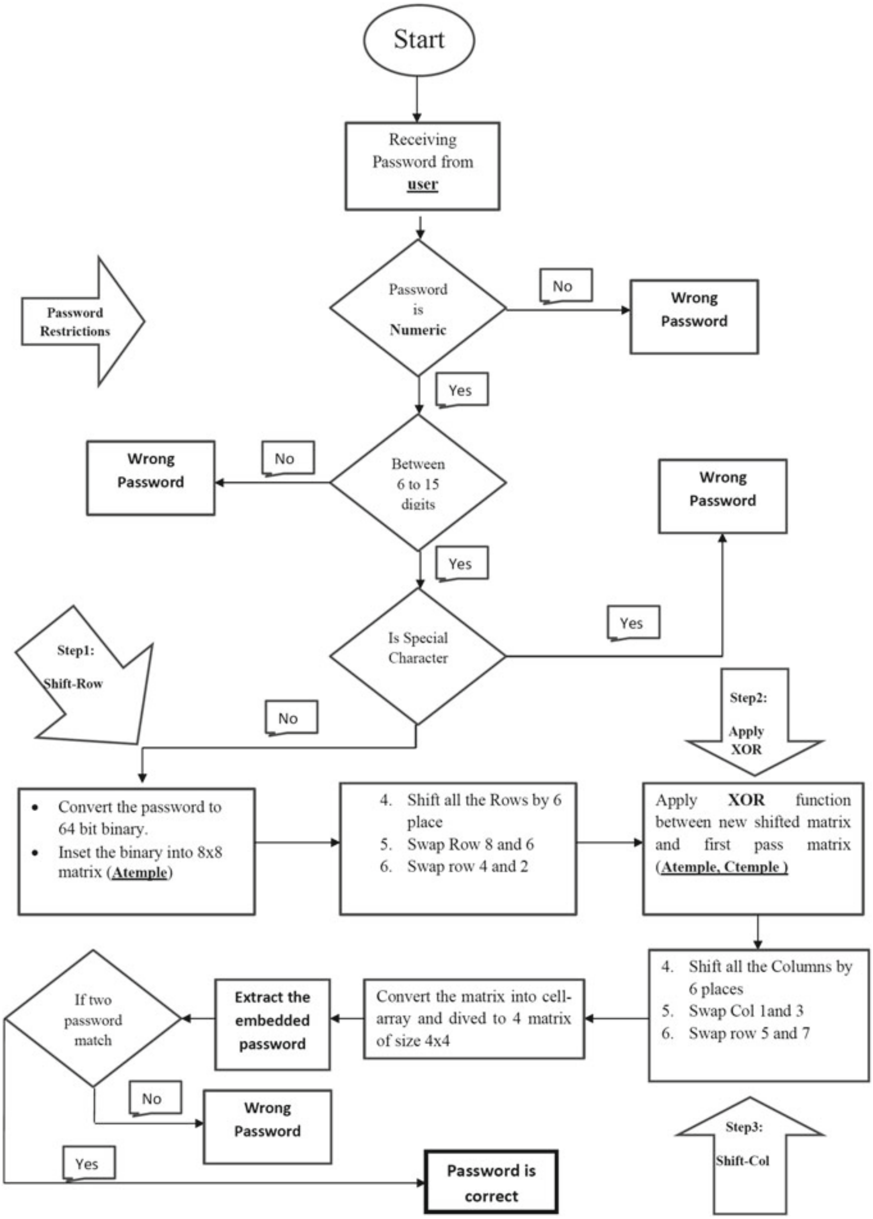
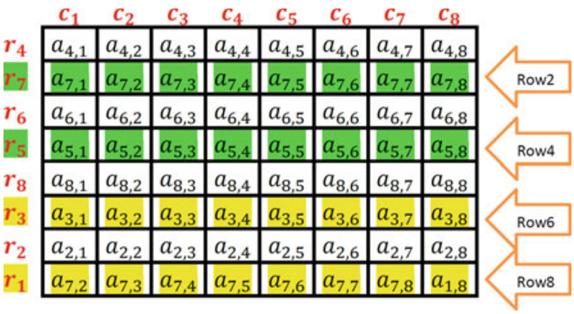

Start
Receiving Password from user
Password Restrictions
Password is Numeric
No
Wrong Password
Yes
Between 6 to 15 digits
No
Wrong Password
Yes
Is Special Character
Yes
Wrong Password
Step1: Shift-Row
No
Step2: Apply XOR
Convert the password to 64 bit binary.
Inset the binary into 8x8 matrix (Atemple)
4. Shift all the Rows by 6 place
5. Swap Row 8 and 6
6. Swap row 4 and 2
Apply XOR function between new shifted matrix and first pass matrix (Atemple, Ctemple)
4. Shift all the Columns by 6 places
5. Swap Col 1and 3
6. Swap row 5 and 7
Convert the matrix into cell-array and dived to 4 matrix of size 4x4
Extract the embedded password
If two password match
No
Wrong Password
Yes
Password is correct
Step3: Shift-Col

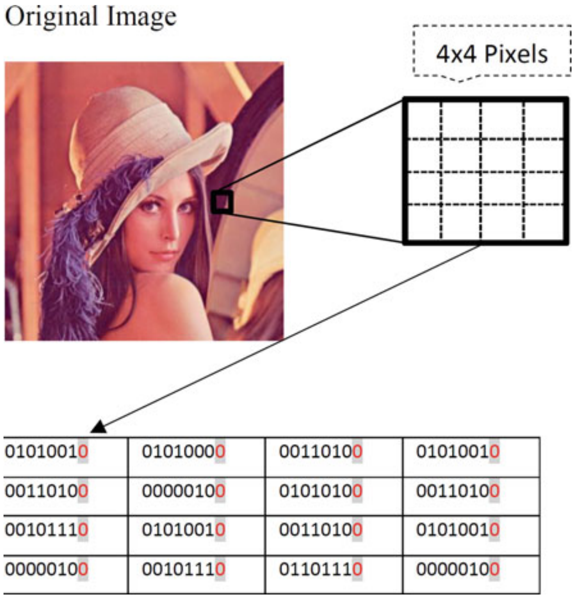
Fig. 5.3 Proposed key encryption procedure



	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
r_4	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$	$a_{4,5}$	$a_{4,6}$	$a_{4,7}$	$a_{4,8}$
r_7	$a_{7,1}$	$a_{7,2}$	$a_{7,3}$	$a_{7,4}$	$a_{7,5}$	$a_{7,6}$	$a_{7,7}$	$a_{7,8}$
r_6	$a_{6,1}$	$a_{6,2}$	$a_{6,3}$	$a_{6,4}$	$a_{6,5}$	$a_{6,6}$	$a_{6,7}$	$a_{6,8}$
r_5	$a_{5,1}$	$a_{5,2}$	$a_{5,3}$	$a_{5,4}$	$a_{5,5}$	$a_{5,6}$	$a_{5,7}$	$a_{5,8}$
r_8	$a_{8,1}$	$a_{8,2}$	$a_{8,3}$	$a_{8,4}$	$a_{8,5}$	$a_{8,6}$	$a_{8,7}$	$a_{8,8}$
r_3	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	$a_{3,5}$	$a_{3,6}$	$a_{3,7}$	$a_{3,8}$
r_2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	$a_{2,5}$	$a_{2,6}$	$a_{2,7}$	$a_{2,8}$
r_1	$a_{7,2}$	$a_{7,3}$	$a_{7,4}$	$a_{7,5}$	$a_{7,6}$	$a_{7,7}$	$a_{7,8}$	$a_{1,8}$

Row2
Row4
Row6
Row8

Fig. 5.4 Key encryption matrix



01010010	01010000	00110100	01010010
00110100	00000100	01010100	00110100
00101110	01010010	00110100	01010010
00000100	00101110	01101110	00000100

Fig. 5.5 Original image division

Proposed Watermark Generation Procedure

The first step of proposed watermark generation divides the image into non overlapping blocks of size 4 × 4. As illustrated in Fig. 5.5, after dividing the original image into blocks of 4 × 4 pixels, the least significant bit of each pixel is padded to zero, before starting any operation. The reason for such action is to construct the standard template, which might be helpful in the tamper detection procedure. Thus, to create the same situation in both, watermarking and tamper detection, the least Significant Bit of each pixel is padded to zero.

Furthermore, as the following MATLAB code illustrated, if the proposed algorithm recognize the digital image as a RGB image, all three channel (red, green and blue) of the digital image will be separately divided into non-overlapping blocks.

As following computer program illustrated, all three channel of the digital image is divided into non-overlapping blocks of size 4 × 4 pixel. This operation is implemented into MATLAB programing language by converting the channel of the digital image into a cell-array; in this case each element of the cell-array contains arrays with size of 4 × 4 pixels.

Example of a MATLAB Computer Program

```
redChannel = IR(:, :, 1);
greenChannel = IR(:, :, 2);
blueChannel = IR(:, :, 3);

\%Dividing the image to blocks of size 4x4
C = cell(1,no_block_cell);
C1=cell(1,no_block_cell);
C2=cell(1,no_block_cell);

for i=1:no_of_block_height
 for j=1:no_of_block_width
     C{(i-1)*no_of_block_width+j}=redChannel
       (4*i-3:4*i,4*j-3:4*j);
     C1{(i-1)*no_of_block_width+j}=greenChannel
       (4*i-3:4*i,4*j-3:4*j);
     C2{(i-1)*no_of_block_width+j}=blueChannel
       (4*i-3:4*i,4*j-3:4*j);
  end
end
```

(Implementation of Dividing the Image into 4 × 4 block)

The next step towards creating the 16-bit watermark calculates the average intensity with the help of Eq. (5.1), and the average intensity here is the average of the pixel value inside each block. As there is a limitation for the watermark length, the most sensitive operation related to all the 16 pixels inside each block is selected, in order to sense any small change inside of pixel value. The average length should be numbers between 0 and 16 bits, but in the proposed algorithm only an 8 bit of the watermark specifies the average intensity inside the blocks and it is given as.

$$A = \frac{\sum_{n=1}^{16} (a_n)}{B} \tag{5.1}$$

where “a” is the value of the pixel inside each block and “B” is the number of pixels inside each block. The number of bits dedicated to the average intensity is 8 bits. The first 8 bits of the watermark is the value of A(x). The next 8 bits are constructed by using the bit summation and parity check. To find the most fragile algorithm within each block of the original image, several operations, such as, the average intensity is examined. The result of such expressions in tamper detection is usually fair, but not optimal. As illustrated by the following MATLAB programming code, for the next 8 bits of watermark, parity check is calculated for the results of the addition of rows inside each block, which is defined as Eq. (5.2).

$$f = \begin{cases} 0 & \text{if count is even} \\ 1 & \text{Otherwise} \end{cases} \tag{5.2}$$

Example of a MATLAB code for paritty check

```
\%Proposed Paritty check Rows

for e=1:A_size
  for eee=1:4
   if mod (sum(dec2bin(A2{1,e}(1,eee))==’1’), 2)== 0

         A2{1,e}(1,eee) = 0;
    else
         A2{1,e}(1,eee) = 1;

    end
  end
end

\%Proposed Paritty check coloumns

for e=1:A_size
  for eee=1:4
   if mod (sum(dec2bin(A3{1,e}(eee, 1))==’1’), 2)== 0

          A3{1,e}(eee,1) = 0;
     else
          A3{1,e}(eee,1) = 1;
   end
  end
end
```

(Implementation of the Proposed Pritty-check)

Two distinct additions have to be calculated. The first procedure adds the pixel value in all the four columns of each block, and adds them all together as demonstrated by Fig. 5.6. The same action is repeated for all the rows of each block. The results of the addition of A and B are converted to the binary form. The parity check for each value is then calculated. Parity check is the process of counting the number of 1s in each binary value.

At this stage, there are two numbers with the length of 8 bits. The combination of both numbers must not exceed 8 bits. In order to optimally combine these two

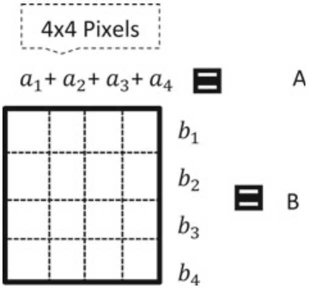


Fig. 5.6 Row and column addition

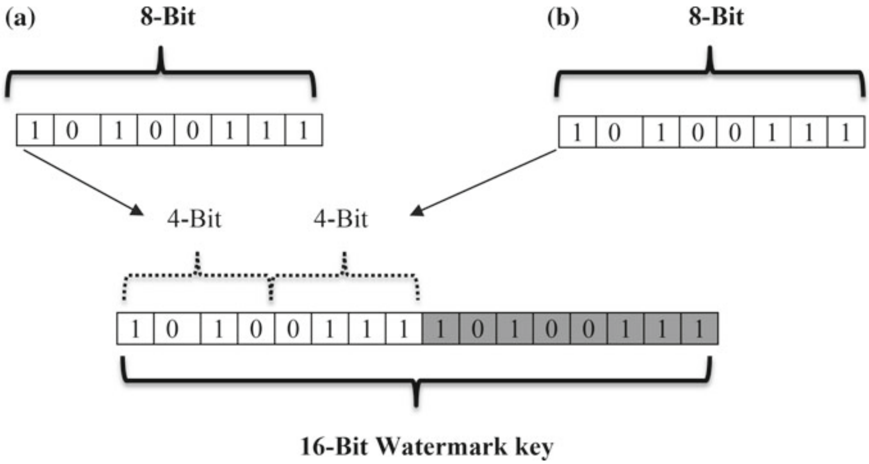


Fig. 5.7 Combination of the first 8 bit

numbers, the first 8 bit value is divided into two (4 bits) as illustrated by previous MATLAB code for paritty-check, then the parity is calculated by Eq. (5.2). In this way the results of the parity check for both, A and B is 4 bit long, which together creates the second 8 bit of the watermark. As shown in Fig. 5.7, the first 4 bit of the watermark is constructed by Eq. (5.1), which is the average intensity of the pixels inside each block. Any major change in pixel values will affect the first 8 bit of the proposed watermark. The second 8 bit combination will take care of the minor tampering; the perfect mixture of row and column summation parity checks creates an efficient 16 bit of watermark for the tamper detection procedure, in this case.

Moreover, as Fig. 5.8 demonstrated, the combination of the average intensity of the whole block is where the parity checks of the row summation and column summation, creates the 16 bit of the watermark. Selecting the sum operation for the second part is the key to success for this proposed algorithm. After a comprehensive research on fragile watermarking, and examining the different operational sensitivity,

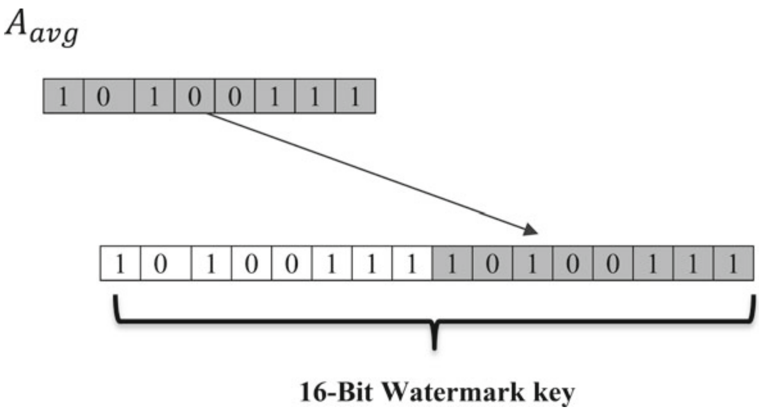


Fig. 5.8 Watermark bit generation second 8-bit

the bit addition turns out to be the most delicate operation that can be used for tamper detection and localization.

Proposed Embedding Algorithm

In addition, in the proposed watermarking procedure as shown in previous section, the first 7 bit of the watermark is constructed by Eq. (5.1) which is the average intensity of the pixels inside each block and the 8th watermarked bit is created by conducting parity check on the first seven watermarked bits. Any major change in pixel values will affect the first 8 bit of the proposed watermark. The second 8 bit combination will take care of the minor tampering; the perfect mixture of row and column summation parity checks creates an efficient 16 bit of watermark for the tamper detection procedure in this case.

The 16 bit watermark key for each block is then generated. The embedding procedure takes place in the least significant bit (LSB) of each pixel value inside each block as shown in Fig. 5.2 in last section. Utilizing the least significant bit for the embedding procedure has several advantages such as less time consumption, less change in the pixel value which results in a better image perceptibility and more sensitivity relations to any tampering and manipulation. Figure 5.2 illustrates, after embedding the 16 bit of the watermark key, the image is reconstructed from all the watermarked blocks to the original dimensions and the watermarked image is then produced.

5.3.2 The Proposed Tamper Detection Scheme

The proposed image authentication flowchart is presented in Fig. 5.9. The first level of the proposed image tamper detection starts with checking the secret pass key, as entered by a user. After inserting the secret key, the extraction progress is invoked and the two keys are compared. As illustrated in Fig. 5.9, in case the two values do not match, the corresponding watermarked image will be marked as not original and there will be no further progress. However, the detailed explanation of the proposed key-decryption and key comparison will be presented in the following sub section.

After the confirmation of the secrete keys, the tamper detection procedure continues by dividing the image into non-overlapping blocks of size 4 × 4 pixels. As mentioned in the previous section, to create the same atmosphere just as it was before the watermarking procedure, all the least Significant Bits should be converted to zero. Before padding and converting the LSBs to zero, the watermark extraction procedure should be completed.

Once all the least Significant Bits are shifted to zero, the procedure of creating the watermark key for each block is started, by producing the first 8 bits. The first 8 bits of the watermark key is created from the average intensity of the pixel values inside the blocks. The average intensity is created by Eq. (5.1). The first 7 bits is the binary form of the average intensity, and the last bit is the parity check of the seven bits, which are calculated by Eq. (5.2).

After generating the first 8 bits of the watermark for the tamper detection procedure, the sum of rows and columns for each block is calculated. As mentioned in the watermarking procedure, there is a limitation for the embedding procedure, which is 8 bits. Thus, these two values have to be combined in a way that, they would still be able to sense any minor manipulations in the pixel bit values. In order to do this combination, the first total sum of the results is divided into two equal 4 bits, and the parity check is calculated for the first half. The same procedure is done for the total of the column results. In this way the second 8 bit of watermark is generated.

As illustrated by Fig. 5.9, after generating the 16 bit watermark, the first level of the proposed tamper detection continues by comparing between the new generated watermark values "A" and the extracted watermark values "A". This comparison is conducted for every single block inside the watermarked image. If the average intensity value of the particular block is equal to the extracted value, the specific block is marked as authenticated, and in any other case it is marked as not original. The authenticated blocks are sent to the second level of the tamper detection procedure, which is the comparison of the bit in the second half of the generated watermark "B", and the 8 bits extracted from the watermarked image "B". If both values are equal the image is marked as original, otherwise the particular block is marked as tampered.

The second level of the proposed tamper detection algorithm starts with performing the bit-count function for each pixel stored in the cell-array called "A3". As mentioned in the past sub-chapter, the result of the bit-count function is based on

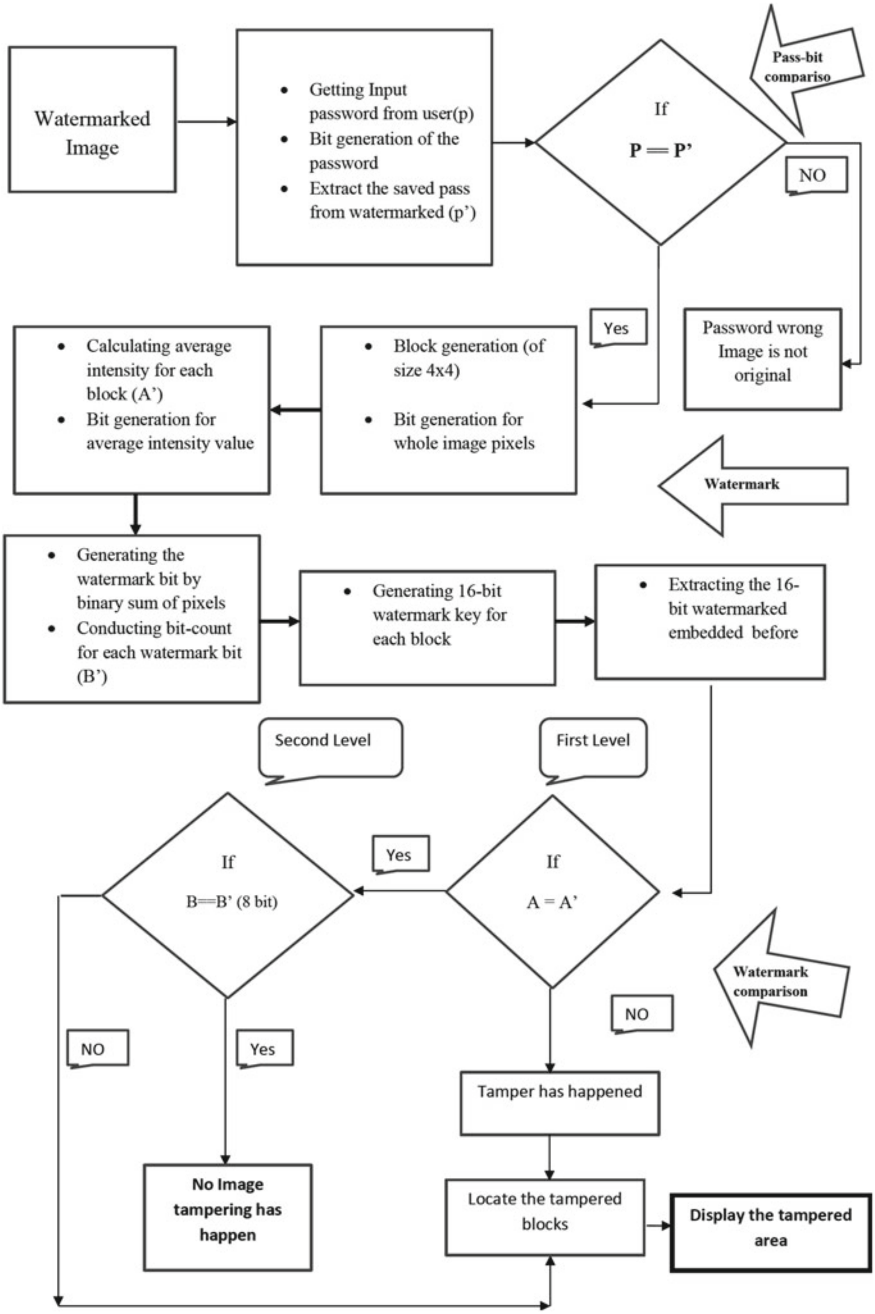

Watermarked Image
Getting Input password from user(p)
Bit generation of the password
Extract the saved pass from watermarked (p')
If
P == P'
Pass-bit compariso
NO
Yes
Password wrong Image is not original
Calculating average intensity for each block (A')
Bit generation for average intensity value
Block generation (of size 4x4)
Bit generation for whole image pixels
Watermark
Generating the watermark bit by binary sum of pixels
Conducting bit-count for each watermark bit (B')
Generating 16-bit watermark key for each block
Extracting the 16-bit watermarked embedded before
Second Level
First Level
If
B==B' (8 bit)
Yes
If
A = A'
Watermark comparison
NO
Yes
NO
Tamper has happened
No Image tampering has happen
Locate the tampered blocks
Display the tampered area

Fig. 5.9 Proposed authentication flowchart

the number of ones among the binary form of the pixel. However, after storing the result of the proposed bit-count function inside the cell-array “A3”, the second level of tamper detection will be conducted as illustrated in following programming code. Mainly, the second level of the proposed tamper detection algorithm, consist of a compressive bit-comparison, between the extracted bits from the primary cell-array “C” and the results of the proposed bit-count function which is stored in cell-array “A3”. In case there is any tamper detected the location of the tampered block along with all the pixels stored in the block will be stored in the cell-array Tampered as demonstrated in following MATLAB code.

Example of a MATLAB code for Tamper Detection

```
\%Second Level of Tamper detection and localization

for e=1:A_size
  for eee=1:4
   if mod (sum(dec2bin(A3{1,e}(eee, 1))==’1’), 2)== 0

               A3{1,e}(eee,1) = 0;
        else
               A3{1,e}(eee,1) = 1;
    end
  end
end

for e=1:A_size
  for eee=1:4
     if bitget (C{1,e}(2,eee),1) ~= A3{1,e}(eee,1)
       Tampered{1,e}= C{1,e};
       Tampered1{1,e}= C1{1,e};
       Tampered2{1,e}= C2{1,e};
     end
  end
end
```

(Implementation of Second-Level Tamper Detection)

All the blocks, which are marked as tampered will be presented as tampered regions inside the forged image. The tamper detection process skips the two blocks, in which the secret keys are stored. The two-level inspection in the proposed algorithm makes sure that, no manipulation is invisible. To evaluate the tamper detection rate, several different experiments have been conducted and are explained in the following section.

5.4 Experimental Results

In this section, we have discussed all the experiments conducted. In order to evaluate efficiency of the proposed method, some visibility examinations, along with various tampering attacks are examined, and the results have been compared with some current, related image authentication methods. To show the quality of watermark image, several screened result shots have been discussed in the following sub sections. To evaluate the perceptibility of the output, image peak signal-to-noise ratio (PSNR) has been calculated by Eq. (5.3). PSNR is used to evaluate the quality difference between the source image and the modified image.

$$PSNR = 10.\mathrm{Log}_{10}(\frac{1}{mn}\sum_{j=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2) \tag{5.3}$$

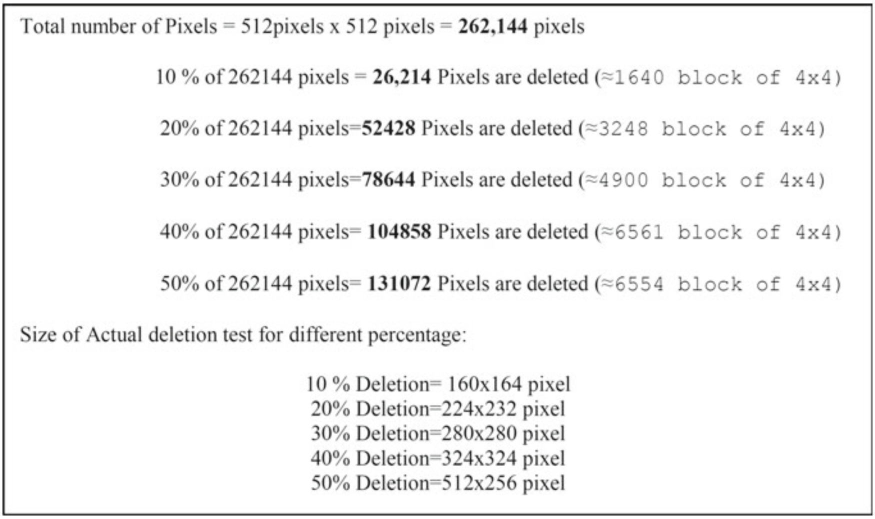
where, m × n are the dimensions of the monochrome image of I and K (I denotes original version of the original image and K denotes the gray-scale of the watermark image). To evaluate the efficiency of the proposed authentication algorithm, different types of collage attack, such as, the deletion attack, copy and paste attack, and drawing tampering have been conducted. To evaluate the tamper detection rate, the same method as [35] has been used in this chapter. In this method the bit tampering attack is used, where certain amounts of bits inside the blocks are manipulated and the tamper detection rate is calculated by the following equation:

$$D_t = (\frac{Number\ of\ detected\ blocks}{Number\ of\ tampered\ blocks}) \times 100 \tag{5.4}$$

To evaluate the tamper detection rate, two types of bit tampering have been conducted. Both tampering attacks are based on the number of blocks detected by this method. In the first attack, the number of tampered bits is unknown, because this attack changes the number of bits in the first five most significant bits into 1. The value of bits that are already 1 cannot be considered as been tampered. The second bit tampering, (which is the same as the tamper attack tested in [29, 35]), is a decent way for the tamper detection rate measurements. The details of the bit tampering attacks have been described in following sections.To evaluate the overall performance of the proposed tamper detection and tamper localization, false positive (FP) and false negative (FN), and average of detection rate (AV) have been measured by the following formula:

$$FP = (\frac{Number\ of\ Pixels\ in\ Untampared\ Part,\ Detected\ as\ Tampered}{Total\ Number\ of\ Pixels\ in\ Untampered\ Region}) \tag{5.5}$$

$$FN = (\frac{Number\ of\ Pixels\ in\ Tampered\ Part,\ Detected\ as\ Untampered}{Total\ Number\ of\ Pixels\ in\ Tampered\ Region}) \tag{5.6}$$



Total number of Pixels = 512pixels x 512 pixels = **262,144** pixels

10 % of 262144 pixels = **26,214** Pixels are deleted (≈1640 block of 4x4)

20% of 262144 pixels=**52428** Pixels are deleted (≈3248 block of 4x4)

30% of 262144 pixels=**78644** Pixels are deleted (≈4900 block of 4x4)

40% of 262144 pixels= **104858** Pixels are deleted (≈6561 block of 4x4)

50% of 262144 pixels= **131072** Pixels are deleted (≈6554 block of 4x4)

Size of Actual deletion test for different percentage:

10 % Deletion= 160x164 pixel
20% Deletion=224x232 pixel
30% Deletion=280x280 pixel
40% Deletion=324x324 pixel
50% Deletion=512x256 pixel

Fig. 5.10 Size of the tampering attacks

$$FN = (1 - (\frac{FN + FP}{1 + R})) \times 100 \tag{5.7}$$

"R" is number of regions, which are tampered. To have fair evaluation, for tamper detection rate, and measuring the false positive and false negative, a number of standard 512512 gray-scale images have been selected. However, for the rest of the experiments, standard images with different dimensions have been chosen to illustrate the flexibility of the proposed algorithm. However, for the rest of the experiments, standard images with different dimensions are selected to illustrate the flexibility of the proposed algorithm. Furthermore, the other factor that is considered in evaluating the proposed tamper detection algorithm is the size of the tampering attacks. Thus, different percentages of the tampering attacks such as 10, 20, 30, 40 and 50 % are selected. Figure 5.10 is illustrating the method that is utilized for the calculation of tampering attacks size.

5.4.1 Watermark Quality Experiments Result

Throughout this section the image quality check has been completed by both, checking the image attributes, and measuring the PSNR of the output image. As illustrated in Fig. 5.11, the proposed outcome of the watermark quality is fairly high. The images illustrated in Fig. 5.11 are of different types, colors, and dimensions. The results of the proposed watermarking algorithm clearly prove that, the difference between the original image and watermarked image is unrecognizable by the naked eye.

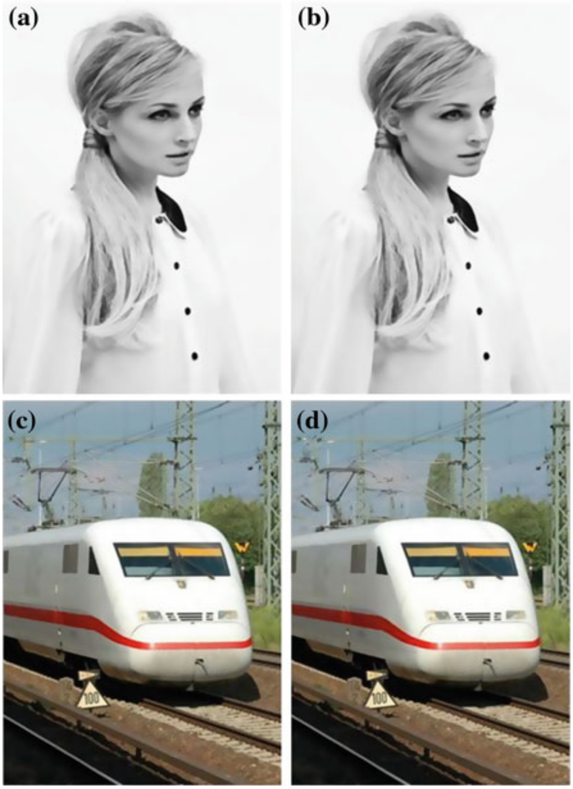


Fig. 5.11 **a** Lady original, **b** Lady watermarked, **c** Train original, **d** Train watermarked

To truly evaluate the quality of the watermark image, the PSNR of the image has to be calculated. In this chapter the PSNR has been calculated by Eq. (5.3). Any amount of PSNR that is higher than 30 dB is considered satisfactory; the average value of PSNR for this proposed watermarking scheme is 51.14 dB, which can be considered as being very high among the image authentication methods, which use the watermarking technique. However, if a tamper detection or image authentication method produces a good quality watermark along with a decent tamper detection rate, it can be used for several real-time image processing applications. Table 5.1 illustrates the PSNR value of 15 grayscale image with different sizes.

However, as Fig. 5.12 illustrated, the PSNR value variations of the proposed watermarked images is between 50.60 and 51.15 db. Moreover, the size of the watermarked image, sometimes have a minor effect on the overall value of PSNR. The PSNR value test is conducted for several images with different type, different size and different dimensions. As Table 5.1 illustrated, the PSNR values demonstrate a high quality of the proposed watermarking output.

Table 5.1 PSNR value of watermark

Image	PSNR peak signal-to-noise ratio (db) Watermarking algorithm 1
Lena512×512.jpg	50.9388
Cameraman512×512.jpg	51.1344
Cameraman256×256.jpg	51.1760
Image4_240×360.jpg	51.1099
Image12_416×536.jpg	51.1332
Figureprint.jpg	51.1231
Deert512×512.jpg	51.1344
JetPlane512×512.jpg	50.6014
Partot512×512.jpg	51.1522
Peppers512×512.jpg	51.1432
Pirate512×512.jpg	50.6525
Tank512×512.jpg	50.6145
Womn_blonde512×512.jpg	51.1522
Livingroom512×512.jpg	51.1296
Lake256×256.jpg	50.8681
Average	51.0042

5.4.2 Tamper Detection Experiments Results

Several tampering attacks have been conducted to evaluate the proposed image authentication and the tamper localization scheme. The first digital image database selected by this research which introduced the famous digital images such as Lena to the world of image processing is a database produced by Signal and Image Processing Institute in University of Southern California [38]. The digital images selected for comparison, are selected from the same source which has been used by the other publications for evaluation their results. However, more than 2,000 digital image is collected from trusted academic and research sources [38–41]. Digital images collected from the mentioned sources have different category, type and dimensions. Following are the completed information about the collected images:

- Dimension: 512 × 512 Pixels, 256 × 256 Pixels, 300 × 256 Pixels, etc.
- Format: *.BMP,*.TIF,*.JPEG,*.PNG,*.PPM, *.GIF
- Pixel depth: 8 bits/pixel …

5.4.3 Deletion and Copy-Move Tampering Attacks

As described in past sub-chapter, several digital tampers with different size and different target locations will be examined to determine the efficiency of the

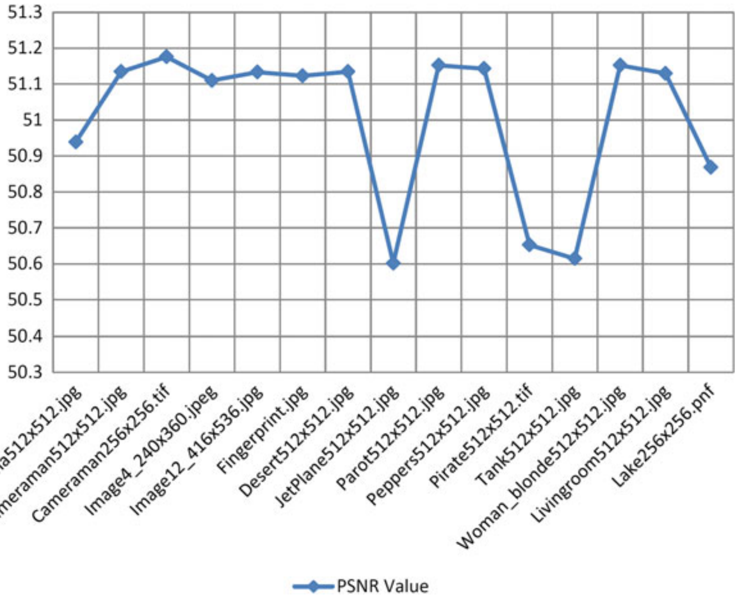


Fig. 5.12 PSNR value frequency chart

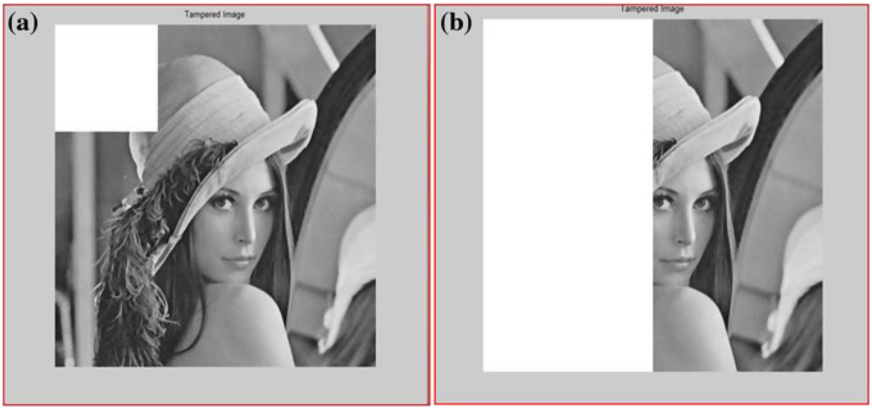


Fig. 5.13 **a** 10 % deletion on the edges, **b** 50 % deletion on the edges

proposed algorithm. However, every tamper detection algorithm produce different tamper detection rate against tampers that occurs on the edges of the digital images and tampers which occur inside the digital image. Thus, this part of the proposed evaluation is divided into, tampers which targeted edges of the digital image and tampers that aim anywhere but corners of the digital image. The Fig. 5.13 is illustrating different percentages of Deletion attack on the edge of the digital image.

Table 5.2 False positive for deletion attack inside image

Image	Different percentage of deletion attack inside the digital image				
	10 %	20 %	30 %	40 %	50 %
Lena	0.0050	0.0043	0.0061	0.0166	0
Cameraman	0.0056	0.0088	0.0061	0.0082	0.0156
Desert	0.0056	0.0043	0.0061	0.0082	0.0156
JetPlane	0.0050	0.0043	0.0061	0.0082	0.0156
Parrot	0.0056	0.0088	0.0061	0.166	0
Peppers	0.0056	0.0043	0.0061	0.0082	0.0156
Pirate	0.0056	0.0088	0.0061	0.0082	0.0156
Tank	0.0027	0.0043	0.0061	0.0166	0.0156
Woman_blonde	0.0056	0.0043	0.0436	0.0165	0
Livingroom	0.0056	0.0088	0.0061	0.0184	0.0156
Lake	0.0056	0.0088	0.0061	0.0166	0.0156
Average	0.005227	0.006345	0.009509	0.012936	0.011345

Deletion attacks in this section have been performed to evaluate the accuracy of tamper localization of the proposed method. To truly evaluate the power of tamper localization, different sizes and portions of an image have been tested. For substitution tampering, different portions of the test image are copied and pasted. In some cases an outside object is copied and pasted into the test image. For detecting single pixel deletion tampering, square portion deletion and non-straight deletions have been examined. Furthermore in order to cover all the angles for substitution tampering attacks, the same color substitution, drawing tampering, and different color substitutions have been tested.

Several deletions tampering located on the edges are experimented in this section. The proposed tamper detection algorithm achieved 100 % average tamper detection results, for deletion tampering on the edges. Thus, there is no need to illustrate the result separately inside a table. However, tampering attack which is located inside the digital image has different result which is illustrated in Table 5.2. The false positive results are calculated by Eq. (5.5).

Table 5.3 illustrates the average detection rate for five distinctive tamper sizes, which are calculated according to formula demonstrated in (5.7). The result established inside the Table 5.3; clearly verify the efficiency of the proposed algorithm in detection of deletion attack with different dimensions. However, later on this sub-chapter, the experiment result of the proposed algorithm, will be compared to different image authentication algorithms.

As Fig. 5.14 illustrated, with increasing of the tamper size, the average tamper detection rate will be decreased. Nevertheless, the decreasing value is not noticeable and the range of tamper detection frequency is higher than 99.20–100 %. However, as illustrated in the following figure, the range of change for tamper detection rate for 50 % tamper is abnormal.

The average tamper detection rate which is measured by Eq. (5.7) is the base evaluation factor in this section. As illustrated in Table 5.4., the performance of the

Table 5.3 Average tamper detection rate of deletion attacks inside the image

Image	Different percentage of deletion attack tamper detection algorithm 1				
	10 %	20 %	30 %	40 %	50 %
Lena	99.75	99.78	99.69	99.17	100
Cameraman	99.72	99.56	99.69	99.59	99.22
Desert	99.72	99.78	99.69	99.59	99.22
JetPlane	99.75	99.78	99.69	99.59	99.22
Parrot	99.72	99.56	99.69	99.17	100
Peppers	99.72	99.78	99.69	99.59	99.22
Pirate	99.72	99.56	99.69	99.59	99.22
Tank	99.86	99.78	99.69	99.59	99.22
Woman_blonde	99.72	99.78	99.62	99.17	100
Livingroom	99.72	99.56	99.69	99.08	99.22
Lake	99.72	99.56	99.69	99.17	99.22
Average (%)	99.74	99.69	99.68	99.37	99.45

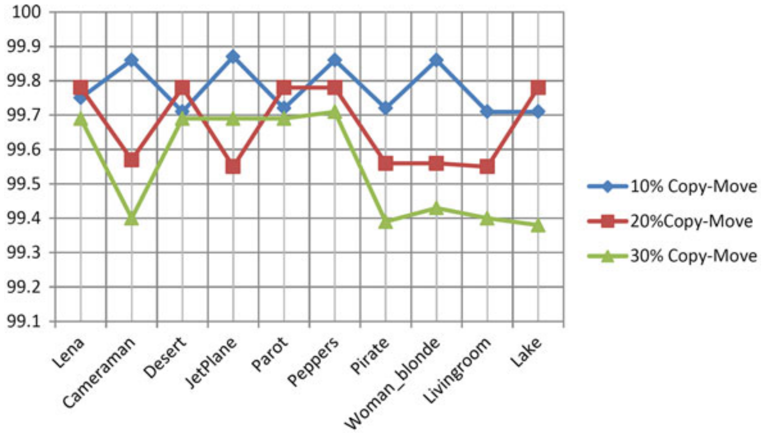


Fig. 5.14 Frequency of tamper detection rate for deletion attack

proposed algorithm is compared with DCT (a fragile image authentication using Discrete Cosine transform) [30], PST (Pinned Sine transform based for image tamper detection) [31] and SLT (semi-fragile Slant Transform for image authentication) [32]. To have fair comparison, six gray-scaled image with size of 512 × 512 pixel are selected, the same digital images that is used by [30, 32]. As illustrated in the Table 5.4, three different percentage of Copy-Move attacked is conducted. As described before, the copy-move tampering is conducted for more than 100 random places inside the image.

However, as Table 5.4 and Fig. 5.15 demonstrated, the overall performance of this research proposed algorithm for 10 % tampering rate is 99.7 % which is more efficient comparing to the other three algorithm. To evaluate the accuracy of proposed tamper localization, an edge deletion tamper has been performed as depicted in Fig. 5.16

Table 5.4 Comparison between the proposed algorithm performance and three other method in terms of tamper detection rate with false positive and false negative

Image	Different percentage of copy-move attack algorithm											
	10 %				20 %				30 %			
	SLT	PST	DCT	Our	SLT	PST	DCT	Our	SLT	PST	DC	Our
Lena	96.0	97.6	95.5	99.7	97.9	98.7	97.1	99.78	98.3	99.0	97.3	99.6
Baboon	96.7	97.3	96.3	99.8	98.1	98.8	96.5	99.57	98.5	99.0	97.0	99.4
Bridge	96.3	97.5	95.4	99.7	97.9	98.6	97.0	99.78	98.2	98.8	96.9	99.6
Trucks	95.7	97.6	95.1	99.8	97.6	98.7	96.7	99.55	98.3	98.8	96.9	99.6
Ship	96.1	97.6	95.2	99.7	97.7	98.8	96.2	99.78	98.3	98.8	96.7	99.6
San Die-go	96.6	97.6	95.4	99.8	98.1	98.8	96.6	99.78	98.6	99.0	97.5	99.7
Average	96.2	97.5	95.5	99.7	97.9	98.7	96.7	99.6	98.4	98.9	97.0	99.5

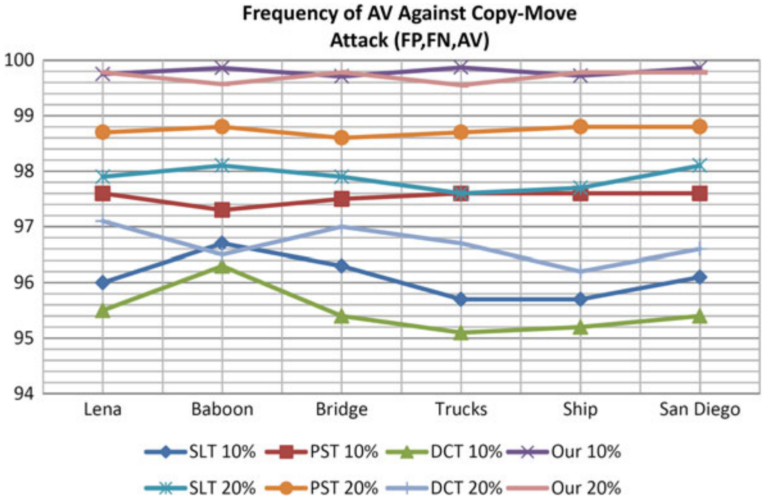


Fig. 5.15 Comparing the AV's of proposed algorithm to three other algorithm

Parts A–D. As illustrated in Part C of Fig. 5.16, the edge of Lenas hat is deleted, and the algorithm used here has accurately detected the deleted regions in Part D. A series of different deletion tampering has also been illustrated in Parts E–H. As shown in Fig. 5.16, various deletions tampering, such as, square deletion, small pixel deletion, rectangle deletion, and finally the deletion of the flag have been conducted. The proposed tamper detection method has successfully located all the tampers, and presented the tampered areas in Part H of Fig. 5.17.

Figure 5.17 illustrates the different kinds of tampering attacks conducted. As illustrated, several substitution tampering attacks have been performed. Part A–D in Fig. 5.17 demonstrates a copy-and-paste modification. It means that, some portion of the watermarked image is copied and pasted into the same image, to hide the man

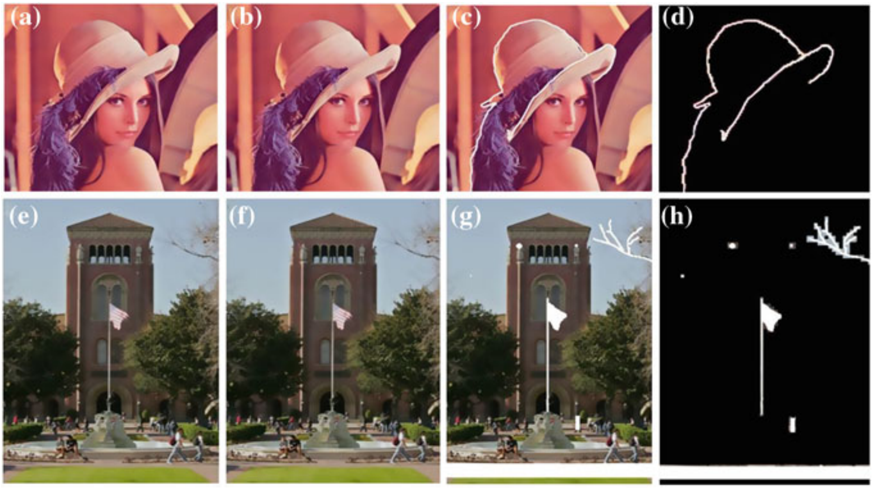


Fig. 5.16 **a** Original Lena, **b** watermarked Lena, **c** edge deletion tamper Lena, **d** edge deletion detected, **e** original square, **f** watermarked square, **g** deletion tamper, **h** deletion tampered detected

lying under the bushes. As illustrated in Part C of Fig. 5.17, the same man and the house on the right hand corner are completely hidden by local objects.

The proposed algorithm has the ability to demonstrate all the numbers of the tampered blocks. However, another type of substitution tampering has also been performed and shown in Parts E–H of Fig. 5.17, and is known as drawing tampering. The date on the right top corner of the image is written by the attacker on the watermarked image, along with a copy of an outside object (the tank)inside the image.

Finally the last part of the Part I-L of Fig. 5.17 illustrates the importance of digital data integrity. This part of the image represents a crime scene, which is forged in part K, and the proposed image authentication method has accurately detected the forged regions. Finally, the experiment results illustrated in Figs. 5.16 and 5.17 clearly proves the capabilities of this proposed algorithm in authenticating the original images along with detecting all types of collage tampering attacks.

As mentioned before, false negative value for tamper detection procedure is the number of pixels in an actual tampered part of the digital image which is detected as original. However, in some special scenarios, the proposed tamper detection algorithm, produce small amount of false negative.However, the proposed tamper detection algorithm perform slightly different against tampering attacks which contains mostly smooth color, such as deletion tampering. Furthermore, as Table 5.5 demonstrated the amount of false positive and false negative generated by proposed tamper detection algorithm is insignificant. As described before, the false positive value for tamper detection procedure is the number of pixel in the un-tampered part of digital image which is detected as a tampered pixel. The false positive values and false negative values, in Table 5.5 are calculated by Eqs. (5.5) and (5.6).

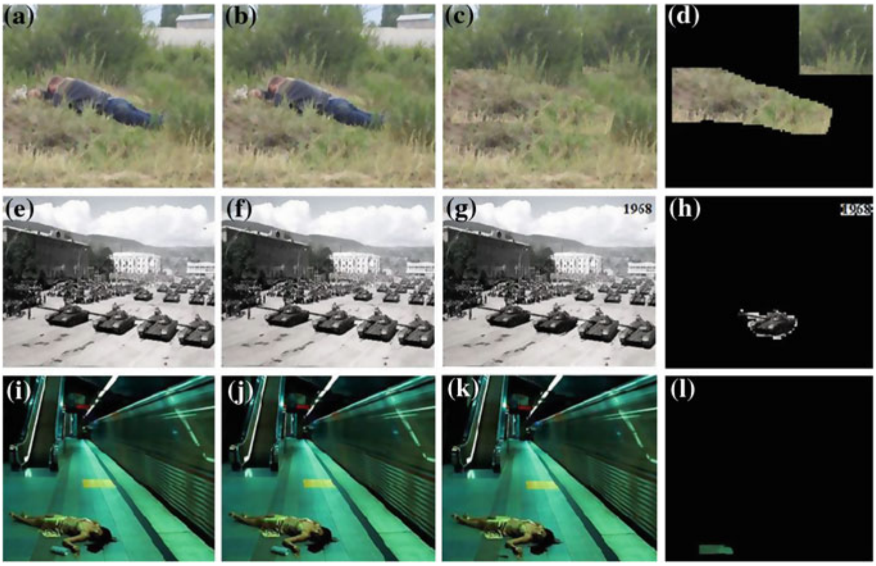


Fig. 5.17 **a** Original lying man, **b** watermarked lying man, **c** copy-paste attack, **d** copy-paste tampering detected, **e** originals tanks, **f** watermarked tanks, **g** same substitution tampering, **h** tamper detected, **i** original crime scene, **j** watermarked crime scene, **k** copy-paste tamper, **l** copy-paste tamper detected

As illustrated in Table 5.5, the proposed tamper detection and tamper localization algorithm are pretty sensitive and very accurate in detecting the tamper regions. This evaluation test has been conducted with several images, and the results were almost identical. As mentioned in previous sections, the aim of the proposed watermarking algorithm is to achieve a good level of efficiency. This experiment result which has been conducted by using Eq. (5.7), illustrates a good level of efficiency for the proposed tamper detection and tamper localization algorithm.

Table 5.5 illustrates the average tamper detection rate against the copy-move tampering attack, which targeted the middle section of the watermarked image. As described earlier in this chapter, the average tamper detection rate is composed of false positive, false negative and number of the regions, which are tampered. The formula for average tamper detection is Eq. (5.7). However, the copy-move tampering attack for this phase of the research will be conducted on number of standard 512 × 512 gray-scale images.

5.4.4 Bit Tampering Experiments Results

In order to evaluate tamper detection rate for the proposed Algorithm 2, a distinct bit tampering experiment, same as [29] and [35], has been performed, and the results of

Table 5.5 Average tamper detection for copy-move tampering located inside

Image	Different percentage of copy-move attack algorithm (inside image)		
	10 %	20 %	30 %
Lena	99.75	99.78	99.69
Cameraman	99.86	99.57	99.40
Desert	99.71	99.78	99.69
JetPlane	99.87	99.55	99.69
Parot	99.72	99.78	99.69
Peppers	99.86	99.78	99.71
Pirate	99.72	99.56	99.39
Woman_Blonde	99.86	99.56	99.43
Livingroom	99.71	99.55	99.40
Lake	99.71	99.78	99.38
Average (%)	99.77	99.66	99.54

experiment have been calculated by Eq. (5.4). To have a fair comparison, tampering attacks have been conducted on the same image, which is the gray-scaled Lena with a 256 × 256 pixel size.

This bit tampering attack targets 10 % of all the blocks inside the watermarked image. The proposed method created a total number of 4,096 blocks with sizes of 256 × 256 pixels. All the tamper blocks have been selected in way that, every two tampering blocks are separated by a distance of 10 blocks. The aim of the Attack-1 is to change the value of the fourth least significant bit of every block to value 1. It means, if the value of the fourth LSB is already 1, no change will take place. On the other hand, Attack-2 will change the value of the fourth LSB to 0, if the value is 1, and also it changes the value to 1, if the fourth LSB value is already 0. Bit tampering attacks are not visible to the naked eye.

Figure 5.18 illustrates the performance of Attack-2 on the gray-scaled Lena image with a 256 × 256 pixel size. As described before, 10 % of all the blocks are selected for this tampering attack. The total number of blocks for the proposed watermarking method is 4,096; it means that, exactly 409 blocks are targeted by this bit tampering attack. The proposed method has been able to detect all the 409 tampered blocks.

Moreover, tamper detection rates in Table 5.6 are calculated based on Attack-1 and Attack-2 by Eq. (5.4). Table 5.6 illustrates the comparison between the tamper detection rate of this research, and three other methods. As described earlier, the tamper detection rate has been calculated by Eq. (5.4), which is the number of the detected tampered blocks, divided by the actual number of tampered blocks. Table 5.7 illustrated the result of proposed tamper detection algorithm against the Bit-tampering Attack-1 and Attack-2.

The reason the differences in number of detected blocks in Table 5.6 is because, the block size proposed by the other three tampering detection methods is 2 × 2 pixels. For instance, the total number of tampered blocks in Lee and Shinfengs [28] method was 16,384 blocks while the number of actual tampered blocks for Attack-2 is 1,636

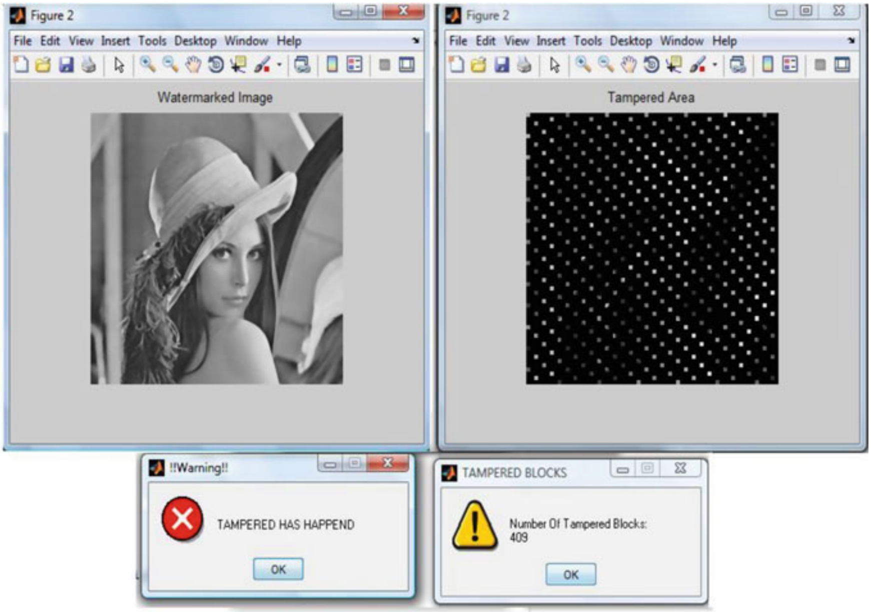


Figure 2
File Edit View Insert Tools Desktop Window Help
Watermarked Image
Figure 2
File Edit View Insert Tools Desktop Window Help
Tampered Area
!!Warning!!
OK
TAMPERED BLOCKS
OK

TAMPERED HAS HAPPEND

Number Of Tampered Blocks:
409

Fig. 5.18 Bit tampering attack-2 completely detected

Table 5.6 Tamper detection rate

Methods	Attack-1	Attack-2
10 % of blocks 2 × 2 Pixels	1636	1636
10 % of blocks 4 × 4 Pixels	409	409
Lee and Shinfeng [28]	Nil (0 %)	Nil (0 %)
Chaluvadi et al. [29]	649 (47.59 %)	849 (51.79 %)
Dadkhah et al. [35]	1187 (81.41 %)	1636 (100 %)
Our method	396 (94.3 %)	409 (100 %)

blocks, which is approximately 10 % of the total blocks. Furthermore, as described in Sect. 5.1, Lees method cannot detect any manipulation in the five most significant bits (MSB). This approach shows 40 % improved tamper detection rate of Chaluvadi et al. [29] as in our last publication Dadkhah et al. [35]; however, because the 3 least significant bits were used in the embedding procedure, the watermark perceptibility has been reduced. As illustrated in Table 5.6 our new proposed method carries a better tampering detection rate, compared to the other three methods. Moreover this proposed authentication scheme creates a watermark with high PSNR value along with the accurate tamper localization and high tamper detection rate.

As Table 5.7 illustrated, the tamper detection rate for Bit-tampering Attack-1 and Attack-2 are different depend on the number of tampered blocks. However, as described before, number of actual tampered blocks is unknown at first. Thus, in order to calculate tamper detection rate for the proposed tamper detection algorithm

Table 5.7 Bit-tampering attack results for attack-1 and attack-2

Image	Bit tampering attact results for algorithm 1		
	Attact 1 (Detected blocks)	Number of actual tampered blocks	Attack 2 (Detected blocks)
Lena	1610 (99.87 %)	1612	1637 (99.99 %)
Cameraman	1402 (99.71 %)	1406	1635 (99.81 %)
Desert	1610 (99.93 %)	1611	1635 (99.81 %)
JetPlane	1429 (99.93 %)	1430	1636 (99.87 %)
Parot	1344 (99.92 %)	1345	1634 (99.75 %)
Peppers	1401 (100 %)	1401	1630 (99.51 %)
Pirate	1603 (100 %)	1603	1635 (99.81 %)
Woman_blonde	1631 (100 %)	1631	1627 (99.32 %)
Livingroom	1636 (100 %)	1636	1623 (99.08 %)
Lake	1574	1574	1632 (99.63 %)
Average (%)	99.87		99.85

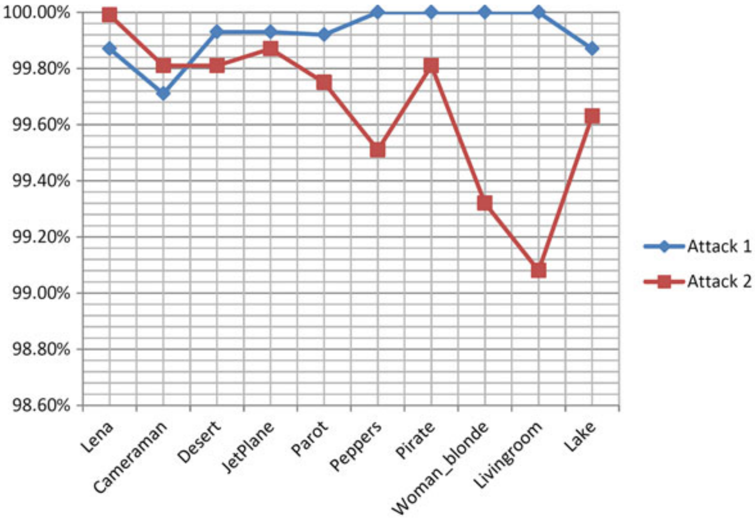


Fig. 5.19 Attack-1 and attack-2 experiment result's frequency

against the Bit-tampering Attack-1, utilizing the Eq. (5.4) is not enough. As illustrated in Table 5.7, to present the true value of tamper detection rate, number of actual tampered pixels has to be calculated. To determine the number of actual tampered block, a proposed function in MATLAB programming language is utilized. The proposed program, distinguish the least significant bits which are altered to value of one and the least significant bits that already contain the value of one. The Fig. 5.19 presents the frequency of the proposed algorithms tamper detection rate against Attack-1 and Attack-2.

5.5 Conclusion

In this chapter, we had proposed a digital image authentication method with an accurate tamper localization capability. The proposed watermarking algorithm generate 16-bit watermark by utilizing proposed binary operations. Various types of tampering attacks were performed in order to evaluate the proposed tamper detection method. The proposed tamper localization algorithm is evaluated by calculating False positive (FP) and False Negative (FN), and the results are compared to several current image authentication techniques, and the comparison results clearly proved the high level of our tamper detection rate. The proposed tamper detection algorithm in this chapter improved our digital image authentication algorithm [35] in terms of tamper detection rate and PSNR value .Having a high tamper detection rate together with a high quality watermark at same time is an added advantage of this proposed method. Finally, all the advantages that have been mentioned make this authentication algorithm very suitable for real-time developments, such as, forensic applications, identity card, e-passport and biometrics. For future investigations, the adding of tamper region recovery should be considered, and in order to add this feature the proposed embedding procedure has to be further explored.

Acknowledgments The authors would like to thank Universiti Teknologi Malaysia for their educational and financial support.

References

1. Vrusias, B., et al.: Forensic Photography. University of Surrey, Computing Department (2001). Technical Report
2. Qi, X., Xin, X.: A quantization-based semi-fragile watermarking scheme for image content authentication. J. Vis. Commun. Image Represent. **22**(2), 187–200 (2011)
3. Ho, A.T.: Semi-fragile watermarking and authentication for law enforcement applications. In: Innovative Computing, Information and Control, p. 286 (2007)
4. Yeung, M.M., Mintzer, F.: An invisible watermarking technique for image verification. In: IEEE image processing, (1997)
5. Cox, I., et al.: Digital Watermarking and Steganography, 2nd edn. Morgan Kaufmann, USA (2007)
6. Lin, W., et al.: Multimedia Analysis, Processing and Communications, vol. 346, p. 139183 (2011)
7. Chang, C.-C., Hu, Y.-S., Lu, T.-C.: A watermarking-based image ownership and tampering authentication scheme. Pattern Recogn. Lett. **27**(5), 439–446 (2006)
8. Lin, T.-C., Lin, C.-M.: Wavelet-based copyright-protection scheme for digital images based on local features. Inf. Sci. **179**(19), 3349–3358 (2009)
9. Guo, H., Georganas, N.D.: Jointly verifying ownership of an image using digital watermarking. Multimedia Tools Appl. **27**(3), 323–349 (2005)
10. Zhu, X., Ho, A.T., Marziliano, P.: A new semi-fragile image watermarking with robust tampering restoration using irregular sampling. Signal Process.: Image Commun. **22**(5), 515–528 (2007)
11. Tsolis, D., et al.: Applying robust multibit watermarks to digital images. J. Comput. Appl. Math. **227**(1), 213–220 (2009)

12. Huang, C.-H., Wu, J.-L.: Fidelity-guaranteed robustness enhancement of blind-detection watermarking schemes. Inf. Sci. **179**(6), 791–808 (2009)
13. Wang, S.-S., Tsai, S.-L.: Automatic image authentication and recovery using fractal code embedding and image inpainting. Pattern Recogn. **41**(2), 701–712 (2008)
14. Kundur, D., Hatzinakos, D.: Digital watermarking for telltale tamper proofing and authentication. Proc. IEEE **87**(7), 1167–1180 (1999)
15. Yu, G.-J., Lu, C.-S., Liao, H.-Y.M.: Mean-quantization-based fragile watermarking for image authentication. Opt. Eng. **40**(7), 1396–1408 (2001)
16. Li, C.-T., Si, H.: Wavelet-based fragile watermarking scheme for image authentication. J. Electron. Imaging **16**(1), 013009–013009-9 (2007)
17. Yen, E., Tsai, K.-S.: HDWT-based grayscale watermark for copyright protection. Expert Syst. Appl. **35**(1), 301–306 (2008)
18. Lin, C.-Y., Chang, S.-F.: Semifragile watermarking for authenticating JPEG visual content. In: Electronic Imaging. International Society for Optics and Photonics, (2000)
19. Lin, H.-Y.S., et al.: Fragile watermarking for authenticating 3-D polygonal meshes. IEEE Trans. Multimedia **7**(6), 997–1006 (2005)
20. Zhang, X., et al.: Reversible fragile watermarking for locating tampered blocks in JPEG images. Signal Process. **90**(12), 3026–3036 (2010)
21. Wong, P.W.: A public key watermark for image verification and authentication. In: Image Processing, ICIP 98. pp. 455–459, (1998)
22. Fridrich, J.: Security of fragile authentication watermarks with localization. In: Proceeding of SPIE Security and Watermarking of Multimedia Contents, pp. 691–700 (2002)
23. Zhang, X., Wang, S.: Statistical fragile watermarking capable of locating individual tampered pixels. IEEE Signal Process. Lett. **14**(10), 727–730 (2007)
24. Ohkita, K., et al.: Improving capability of locating tampered pixels of statistical fragile watermarking, In: Digital Watermarking, pp. 279–293. Springer, (2009)
25. Lin, S.D., Kuo, Y.-C., Huang, Y.-H.: An image watermarking scheme with tamper detection and recovery. In: IEEE Innovative Computing, Information and Control, (2006)
26. Chang, C.-C., Fan, Y.-H., Tai, W.-L.: Four-scanning attack on hierarchical digital watermarking method for image tamper detection and recovery. Pattern Recogn. **41**(2), 654–661 (2008)
27. Zhang, X., Wang, S.: Fragile watermarking scheme using a hierarchical mechanism. Signal Process. **89**(4), 675–679 (2009)
28. Lee, T.-Y., Lin, S.D.: Dual watermark for image tamper detection and recovery. Pattern Recogn. **41**(11), 3497–3506 (2008)
29. Chaluvadi, S.B., Prasad, M.V.: Efficient image tamper detection and recovery technique using dual watermark. In: Nature & Biologically Inspired Computing, (2009)
30. Lin, E.T., Podilchuk, C.I., Delp, E.J.: III Detection of image alterations using semifragile watermarks. In: Electronic Imaging, (2000)
31. Ho, A.T., Zhu, X., Woon, W.: A semi-fragile pinned sine transform watermarking system for content authentication of satellite images. In: IEEE Geoscience and Remote Sensing Symposium, (2005)
32. Zhao, X., Bateman, P., Ho, A.T.: Image Authentication Using Active Watermarking and Passive Forensics Techniques, in Multimedia Analysis, Processing and Communications, pp. 139–183. Springer, (2011)
33. Rosales-Roldan, L., et al.: Watermarking-based image authentication with recovery capability using halftoning technique. Signal Process. Image Commun. **28**(1), 69–83 (2013)
34. Tong, X., et al.: A novel chaos-based fragile watermarking for image tampering detection and self-recovery. Signal Process. Image Commun. **28**(3), 301–308 (2013)
35. Dadkhah, S., Manaf, A.A., Sadeghi, S.: Efficient digital image authentication and tamper localization technique using 3Lsb watermarking. Int. J. Comput. Sci. Issues (IJCSI), **9**(1), (2012)
36. Hsu, C.-S., Tu, S.-F.: Image tamper detection and recovery using differential embedding strategy. In: IEEE Communications, Computers and Signal Processing (PacRim), (2011)

37. Yeo, D.-G., Lee, H.-Y.: Block-based image authentication algorithm using reversible watermarking, In: Computer Science and Convergence, pp. 703–711. Springer (2012)
38. Weber, A.G.: The USC-SIPI image database version 5. USC-SIPI Rep. **315**, 1–24 (1997)
39. Pullen, M. Crime Scene. High Fashion Crime Scenes, (2010) http://www.melaniepullen.com/
40. Christlein, V., et al.: An evaluation of popular copy-move forgery detection approaches. IEEE Trans. Inf. Forensics Secur. **7**(6), 1841–1854 (2012)
41. Perona, M.F.A.P.: Computational vision at Caltech. [Image database], (2010) http://www.vision.caltech.edu/archive.html

Part II
Mobile Ad Hoc Networks and Key Managements

Chapter 6
TARA: Trusted Ant Colony Multi Agent Based Routing Algorithm for Mobile Ad-Hoc Networks

Ayman M. Bahaa-Eldin

Abstract In this chapter, a model for using Multi Intelligent Agents and Swarm Intelligence for trusted routing in mobile ad-hoc networks (MANETs) is presented. The new algorithm called TARA is proved more efficient in different ways than the existing protocols. TARA uses a local trust evaluation method to indicate an objective trust value for each node within the network depending on its forwarding behavior; therefore, a multi agent system is hosted on each node and monitors its forwarding behavior. Both the number and total size of packets being forwarded are considered. The trust value of each node is directly applied to the route as it is being built, so there is no need to propagate the trust values like other trusted protocols. Ant Colony Optimization (ACO) is used to find the best route for delivery and the processes of strengthening and evaporation of the route pheromone value are applied. Simulation using a random-way-point mobility model with different parameters had been carried out and comparison with other protocols complexity and performance was presented.

6.1 Introduction

Mobile ad-hoc networks (MANETs) are a collection of mobile nodes communicating with each other via multi-hope wireless links. Each node in MANETs acts as a host and a router in the same time. A routing protocol is required to establish and maintain such data links taking into consideration all the features and characteristics of mobile nodes, malicious behavior and wireless communication.

MANETs routing protocols are classified into two categories, table-driven (proactive) and on-demand (reactive) [1] protocols. On-demand routing protocols perform better with significantly lower overheads than table- driven routing protocols in many situations. Several ad-hoc on demand routing protocols have been proposed, for

A. M. Bahaa-Eldin (✉)
Department of Computer and Systems Engineering, Ain Shams University, Cairo, Egypt
e-mail: ayman.bahaa@eng.asu.edu.eg

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_6, © Springer-Verlag Berlin Heidelberg 2014

example, ad-hoc on demand distance vector AODV [2], dynamic source routing DSR [3], and temporally ordered routing algorithm TORA [4].

In general, both types of routing protocols for MANETs are designed based on the assumption that all participating nodes are fully cooperative. Due to Magnets' characteristics such as openness, mobility, dynamic topology and protocol weaknesses, these may be targeted by attackers in a number of ways [5]. Several "secure" routing protocols have been proposed for MANETs [6–8]. Most of them assume centralized units or trusted third parties, which actually destroy the self-organization nature of MANETs. These protocols are effective to fight against external attacks, but are not able to prevent selfishness like misbehaviors. For example, a node may refuse to forward data packets for other nodes to save its battery. Therefore, a comprehensive approach is necessary for MANETs to prevent both attacks and misbehaviors. This is achieved by developing mechanisms for measuring the trustworthiness of the network nodes. The measure of the trustworthiness of such nodes is done through a term called trust level, which results in what is called trusted routing protocols.

Many trusted routing protocols have been suggested as an effective security mechanism in MANETs [9–11]. In these protocols, measuring the node's trust level is the challenging issue due to the characteristics of MANETs. These protocols classified the trust relation as direct and indirect relation. Each node has a direct trust relation with the nodes located inside its communication range (neighbors); the direct trust relation is computed by monitoring the behavior of the neighbors in the routing process. On the other hand, the indirect trust relation is concerned with the other nodes located outside the node's communication range (non-neighbors); a useful method to compute the indirect trust relation is flooding the network with request messages and waiting replies. Evaluating the direct and indirect trust relation consumes both bandwidth and energy, delays the route discovery process and complicates the routing process due to the additional computational overhead.

Intelligent and bio-inspired routing protocols had also been introduced [12, 13]. Ant Colony Routing Algorithm for MANETs (ARA) [14] is a good example for such a protocol that performs comparably to DSR with much less overhead.

Recently, Objective Local Trust model had been presented. This model rely on using a multi agent system to calculate an objective trust model for each node in the MANET and avoids misbehaving routes. This protocol contributed by the author is called ATDSR [15].

In this chapter, a novel trusted ant colony based dynamic source routing protocol (TARA) is proposed for MANETs. TARA is based on the ATDSR and ARA algorithms. The main objective of this protocol is to manage trust and reputation with minimal overhead in terms of extra messages and time delay. Swarm Intelligence is used to build and maintain the routes, while self-monitoring of each node to find out its trust value is used. An objective model for measuring the trust value is presented. This object is achieved through installing a multi-agent system (MAS) in each participated node in the network. TARA in comparison with other protocols provides better security with significantly less overhead in terms of extra messages and time delay in finding trusted end-to-end routes.

6.2 Background

In this section, a brief background about routing, trust management and Swarm Intelligence is provided.

6.2.1 Mobile Ad-Hoc Networks

Unlike traditional networks, an ad-hoc network as its name dictates is a temporary network built and operated for a special purpose, “ad-hoc” in Latin means “for this”, and then terminates. One main feature of such networks is that the lack of infrastructure, that is there is no dedicated transmission devices like routers, switches and access points. There are many types of ad-hoc networks like mobile adhoc networks MANETs and wireless sensor networks WSNs and others.

One of the most general types of ad-hoc networks is the MANET type. MANETs represent complex distributed systems that comprise wireless mobile nodes that can freely and dynamically self-organize themselves into arbitrary and temporary, “ad-hoc” network topologies, allowing people and devices to seamlessly internetwork in areas with no pre-existing infrastructure. An ad-hoc device is by definition a source of information, a drain for information and a router for information flows from other devices in the network [16]. MANET allows unrestricted mobility of the mobile devices, as long as at least one device is within transmission range. Direct neighbors are used to route information to nodes outside the transmission range of a wireless node. The transmission range of a node is restricted by the limited battery power of the device, thus the transmission range is relatively small in comparison to the potential overall size of MANET. Mobile routers could also act as gateways to other wireless MANETs.

As any other wireless network, MANETs face some issues related to wireless communication like:

1. Collision of Frames.
2. Delay in Frame Transmission.
3. Interference on Wireless Links.
4. Violation of Security Goals.

In addition to those challenges, MANETs have the following characteristics.

1. Communication carried out via wireless means.
2. Nodes perform the roles of both hosts and routers.
3. Neither centralized controller nor infrastructure can be set up anywhere.
4. Dynamic network topology with frequent routing updates.
5. Autonomous, no infrastructure is needed.
6. Energy constraints applies to both the computational power of nodes and their ability to participate in communication.

7. Bandwidth constraints applies to all nodes.
8. Limited physical security.

Generally, the communication terminals have a mobility nature, which makes the topology of the distributed networks time varying. The dynamical nature of the network topology increases the challenges of the design of MANETs. Each radio terminal is usually powered by energy limited power source (as rechargeable batteries). The power consumption of each radio terminal could be divided generally into three parts, power consumption for data processing, power consumption to transmit its own information to the destination and finally power consumption when the radio terminal is used as a router, i.e. forwarding the information to another terminal in the network. The energy consumption is a critical issue in the design of MANETs.

The mobile devices usually have limited storage and low computational capabilities. They heavily depend on other hosts and resources for data access and information processing. A reliable network topology must be assured through efficient and secured routing protocols for MANETs.

MANETs are useful in areas that have no fixed infrastructure and hence need alternative ways to deliver services. Ad-hoc Networks work by connecting mobile devices to each other in the transmission range through automatic configuration, i.e., setting up an ad-hoc network that is very flexible. In other words, there is no intervention of any controller that gathers data from all nodes and organizes it. All data gathering and cross-node data transfer is done by the nodes themselves.

MANETs are a major goal towards the evolution of 4G (Fourth generation) devices. In the nodes of the MANETs, computing power and network connectivity are embedded in virtually every device to bring computation to users, no matter where they are, or under what circumstances they work. These devices personalize themselves to find the information or software they need. The strife is to make use of all technologies available without making any major change to the user's behavior. There is a continuous work to make the seamless integration of various networks possible, i.e., integration of LAN, WAN, PAN and MANETs. However, there is still a lot of work to be done to make this completely possible. Figure 6.1 shows a network of three nodes, (A, B and C) where each node radio range is presented by a circle around it.

From the figure, we can conclude the following:

- Nodes A and B can communicate together, so they are called neighbors, also nodes B and C
- For A to reach C, B has to work as a router
- Packets sent out be any 2 neighbors can collide together and cancel each other.

6.2.1.1 Routing Protocols

A communication network can be modeled as a graph G = (N, E) [17] where N is a set of nodes or routers and E are the set of all possible links between the nodes. In many

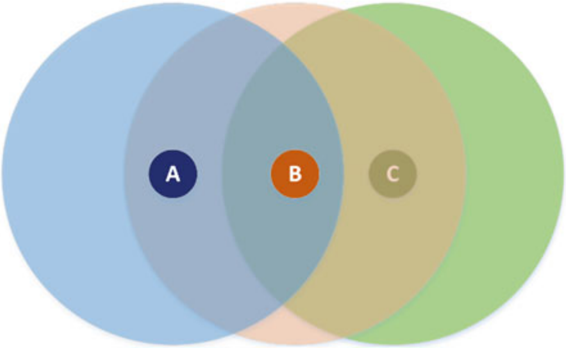


Fig. 6.1 Three wireless nodes

cases, a direct link between the source and destination node of a communication session is not possible, so a multi-hop route is required where a node receives a packet of data and transmits it to the next node in the route. Therefore, each node preserves a routing table in which all the desired destinations are listed and for each of them the next node to send the data to is indicated. These tables are created and maintained by a Routing protocol. Routing protocols are responsible for determining the best route that the data packets should take. These routing protocols allow routers (mobile nodes) to intercommunicate with other routers so that they can determine the structure of the interconnected networks. In ad-hoc networks, a direct communication between any two nodes is a possible subject to adequate radio propagation conditions and transmission power limitations of the nodes.

Routes are different from each other regarding the number of hops, the nodes involved, and the amount of data to be transmitted over each link. Finding the best route for a session of data transmission between two nodes is the sole goal of a routing algorithm. Therefore, Routing is an optimization problem.

Wired network routing protocols, as link-state and distance-vector based protocols are not adequate for MANETs. The main problem with link-state and distance-vector is that they are designed for a static topology.

Another problem is that link-state and distance-vector are highly dependent on periodic control messages. As the number of network nodes increases, the potential number of destinations will also increase. This requires large and frequent exchange of data among the network nodes. Therefore, such protocols require huge bandwidth and a tremendous CPU power that are not normally available in MANETs. Because both link-state and distance vector tries to maintain routes to all reachable destinations, it is necessary to maintain these routes and this wastes resources for the same reason as above.

Another characteristic for conventional protocols is that they assume bi-directional links, e.g. that the transmission between two hosts works equally well in both directions. In the wireless radio environment, this is not always the case.

Dedicated routing protocols for wireless mobile networks are required with a set of desired properties [18]:

1. **Distributed operation**
   The protocol should not be dependent on a centralized controlling node even for stationary networks. The difference is that nodes in an ad-hoc network can enter/leave the network very easily and because of the mobility, the network can be partitioned.
2. **Loop free**
   To improve the overall performance, the routing protocol should guarantee that the routes supplied are loop-free. This avoids any waste of bandwidth or CPU consumption.
3. **Security**
   The radio environment is especially vulnerable to impersonation attacks, so to ensure the wanted behavior from the routing protocol, some sort of preventive security measures are needed. Authentication and encryption is probably the way to go and the problem here lies within distributing keys among the nodes in the ad-hoc network. There are also discussions about using IP-sec [14] with tunneling to transport all packets.
4. **Power conservation**
   The nodes in an ad-hoc network can be laptops and thin clients, such as PDAs, that are very limited in battery power and therefore use some sort of stand-by mode to save power. It is therefore important that the routing protocol support these sleep-modes.
5. **Multiple routes**
   To reduce the number of reactions to topological changes and congestion, multiple routes could be used. If one route has become invalid, it is possible that another stored route could still be valid and thus saving the routing protocol from initiating another route discovery procedure.

Many protocols for MANETs routing had been proposed. They can be categorized in two main categories as follows.

1. Proactive (Table-Driven).
2. Reactive (On-Demand)

Proactive routing protocols attempt to maintain consistent, up-to-date routing information from each node to every other node in the network. These protocols require each node to maintain one table or more to store routing information, and they respond to changes in the network topology by propagating updates throughout the network in order to maintain a consistent network view. The areas in which they differ are the number of necessary routing-related tables and the methods by which changes in the network structure are broadcasted. Examples of proactive protocols are Destination- Sequenced Distance-Vector Routing protocol (DSDV) [19], Cluster-head Gateway Switch Routing (CGSR) [20], and Wireless Routing Protocol (WRP) [21].

Proactive protocols can be viewed as clones for wired network routing protocols and are not suitable enough for MANETs especially with high mobility and dynamic topologies. Therefore, this chapter does not focus on this type of routing protocols.

6.2.1.2 Reactive Routing for Ad-Hoc Networks

Reactive routing protocols create routes only when desired by the source nodes. When a node requires a route to a certain destination, it initiates a route discovery process within the network. This process is completed once a route is found or all possible route permutations have been examined. Once a route has been established, it is maintained by a route maintenance procedure until either the destination becomes inaccessible along every path form the source or the route is no longer desired. Two of these protocols are detailed after.

Ad-Hoc On-Demand Distance Vector Routing

Ad-Hoc On-Demand Distance Vector (AODV) routing protocol [2] builds on the DSDV algorithm previously described. AODV is an improvement on DSDV because it typically minimizes the number of required broadcasts by creating routes on a demand basis, as opposed to maintaining a complete list of routes as in the DSDV algorithm. The authors of AODV classify it as a pure on-demand route acquisition system, since the nodes that are not on a selected path do not maintain routing information or participate in routing table exchanges [2].

When a source node desires to send a message to a certain destination node and does not already have a valid route to that destination, it initiates a *path discovery* process to locate the other node. It broadcasts a route request (RREQ) packet to its neighbors, which then forward the request to their neighbors, and so on, until either the destination or an intermediate node with a *"fresh enough"* route to the destination is located. AODV uses destination sequence number (DestSeqNum) to ensure all routes are loop free and contain the most recent route information. Each node maintains its own sequence number, as well as a broadcast ID. The broadcast ID is incremented for every RREQ the node initiates, and together with the node's IP address, uniquely identifies an RREQ. Along with its own DestSeqNum and the broadcast ID, the source node includes in the RREQ the most recent DestSeqNum it has for the destination. Intermediate nodes can reply to the RREQ only if they have a route to the destination whose corresponding DestSeqNum is greater than or equal to that contained in the RREQ.

During the process of forwarding the RREQ, intermediate nodes record in their route tables the address of the neighbor from which the first copy of the broadcast packet is received, thereby establishing a reverse path. If additional copies of the same RREQ are later received, these packets are discarded. Once the RREQ reaches the destination or an intermediate node with a fresh enough route, the destination/intermediate node responds by unicasting, a route reply (RREP) packet back to the neighbor from which it was first received the RREQ. As the RREP is routed back along the reverse path, nodes along this path set up forward route entries in their route tables, which point to the node from which the RREP came. These forward route entries indicate the active forward route. Associated with each route entry is a route timer, which will cause the deletion of the entry if it is not used within the specified lifetime. Because the RREP is forwarded along the path established by the RREQ, AODV only supports the use of symmetric links.

Routes are maintained as follows. If a source node moves, it is able to reinitiate the route discovery protocol to find a new route to the destination. If a node along the route moves, its upstream neighbor notices the move and propagates *a link failure notification* message (an RREP with infinite metric) to each of its active upstream neighbors to inform them of the erasure of that part of the route [2]. These nodes in turn propagate the link failure notification to their upstream neighbors, and so on until the source node is reached. The source node may then choose to reinitiate route discovery for that destination if the route is still desired.

An additional aspect of the protocol is the use of hello messages, which are periodic local broadcasts by a node to its neighbors. Hello messages can be used to maintain the local connectivity of a node. However, the use of hello messages is not required. Nodes listen for retransmission of data packets to ensure that the next hop is still within reach. If such a retransmission is not heard, the node may use the reception of hello messages to determine if the next hop is still within communication range.

Dynamic Source Routing

Dynamic Source Routing (DSR) protocol [3] is an on-demand routing protocol based on the concept of source routing. Mobile nodes are required to maintain route caches that contain the source routes of which the mobile node is aware. Entries in the route cache are continually updated as new routes are learned.

The protocol consists of two major phases: *route discovery* and *route maintenance*. When a mobile node has a packet to send to some destination, it first consults its route cache to determine whether it already has a route to the destination. If it has a fresh enough route to the destination, it will use this route to send the packet. On the other hand, if the node does not have such a route, it initiates route discovery by broadcasting a *route request* packet. This route request contains the address of the destination, along with the source node's address and a unique identification number. Each node receiving the packet checks whether it knows of a route to the destination. If it does not, it adds its own address to the *route record* of the packet and then forwards the packet along its outgoing links. To limit the number of route requests propagated on the outgoing links of a node, a mobile node only forwards the route request if the request has not been seen yet by the mobile and if the mobile node's address does not already appear in the route record.

A *route reply* is generated when the route request reaches either the destination itself, or an intermediate node which contains in its route cache a fresh enough route to the destination [3]. If the node generating the route reply is the destination, it places the route record contained in the route request into the route reply. If the responding node is an intermediate node, it will append its cached route to the route record and then generate the route reply. In order to return the route reply, the responding node must have a route to the initiator. If it has a route to the initiator in its route cache, it may use that route. Otherwise, if symmetric links are supported, the node may reverse the route in the route record. If symmetric links are not supported, the node may initiate its own route discovery and piggyback the route reply on the new route request.

Route maintenance: if a route fails at any hop in the path during a communication between source and destination, the node encountering the error sends an *error message* to the originator of the route. The failing route is detected when a node fails to forward a packet by using periodic broadcast messages. When a route error is received, the node experiencing the error is removed from the cache of the node receiving the error message. All routes containing the node in error must be shortened at that point. If links do not work, equally well in both directions then *end-to-end acknowledgment* is used to detect the failing link. If the source knows any other available route, it uses it. Otherwise a route discovery mechanism is started.

6.2.1.3 Node Misbehavior and the Trust Problem in Ad-Hoc Networks Routing

The routing protocols discussed above and many others assume that the nodes will fully participate in the routing process. Unfortunately, node misbehavior is a common phenomenon. Misbehaving nodes at the routing level can be classified into two main categories [22].

1. Selfish node: operates normally in the Route Discovery and the Route Maintenance phases of the DSR protocol. However, it does not perform the packet forwarding function for data packets unrelated to itself. The selfish node attempts to benefit from other nodes, but refuse to share its own resources.
2. Malicious node: acts to the detriment of the network by manipulating routing. Many routing protocols use hop count as a metric. A node can falsely claim a low hop count to a destination, enabling it to intercept traffic for that destination. Node identities are not authenticated, so a node can claim to be the destination of a route.

Since such misbehaving nodes participate in the Route Discovery phase, they may be included in the routes chosen to forward the data packets from the source. The misbehaving nodes, however, refuse to forward the data packets from the source. Therefore, the existence of misbehaving nodes may paralyze the routing operation. In this section, the routing misbehavior problem is illustrated in the context of the DSR protocol; the following notations are used while describing the problem caused by routing misbehavior:

P_r: The ratio of misbehaving routes.
P_m: The ratio of misbehaving nodes.

In order to demonstrate the adverse effect of routing misbehavior, a simulation of DSR with the following parameters [15] was carried out.The data packet size is 512 bytes. The wireless transmission range $T_r = 50$ (m) and the total number of nodes $N_t = 100$ mobile nodes were randomly distributed in 100 * 100 m and 200 * 200 (m) flat area. The total simulation time was 3600 s. A random-way-point mobility model [23] was assumed with a maximum speed of $V_m = 20$ (m/s) and a pause time of 0 second. Constant Bit Rate (CBR) traffic was used [23]. Each simulation included

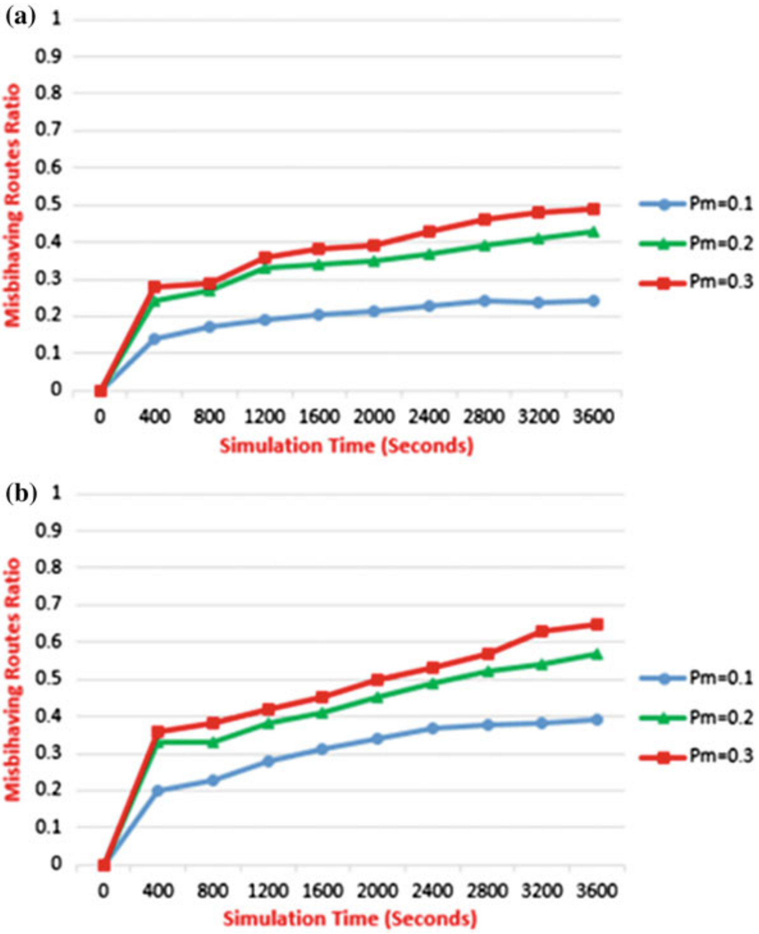


Fig. 6.2 The misbehavior problem illustrated. **a** Area of 100 × 100 m. **b** Area of 200 × 200 m

10 CBR sessions, each of which generated four packets per second. The misbehaving nodes are selected among all network nodes randomly. In our simulations, P_m ranges from 0 to 0.4.

As shown in Fig. 6.2, P_r increases with the increasing of P_m. P_r also increases with network area because the routes are longer and the probability of containing misbehaving nodes increases. The adverse effect of misbehaving nodes in MANETs motivates our development of an efficient approach for detecting and mitigating routing misbehavior.

6.2.2 Swarm Intelligence

6.2.2.1 Swarm Intelligence

Swarm Intelligence (SI) [24–26] is a term referring to the collective behavior of limited-capabilities set of similar objects. Although the individual behavior is far from achieving a goal by itself, the collective behavior archives such a goal.

The inspiration of SI comes mainly from the social behavior of insects living in colonies such as ants, bees, and termites. For example, ants scouting for food lay a chemical compound with a distinctive odor called ***Pheromone***. The following ants follow the same path by finding traces of the pheromone. Overall, the path between the food and the nest is an optimal path and in many cases is the shortestone.

Many algorithms following the SI framework like Ant Colony Optimization (ACO) had been proposed and is now a major area of research and development in optimization problems [27].

In a social insect colony, a worker performs a specific task and leaves other tasks to other worker groups. This division of labor based on specialization is believed to be very efficient. SI offers an alternative way of designing intelligent system, in which autonomy, emergence and distributed functioning replace control, preprogramming and centralization. This approach emphasizes on distributed ness, flexibility, robustness and direct or indirect communication among relatively simple agents [13, 28].

Interaction among insects in a colony is a requirement to achieve the optimization goal. Direct and indirect interaction are both used where in the latter case, the insect changes the environment and the next insect detects such change.

SI gives raise to intelligent behavior through complex interaction of thousands of autonomous swarm members. The ability of ants to self-organize is based on four principles.

- Positive feedback—When an ant follow a path by a pheromone, it increases the pheromone be depositing more to the path. This attracts more ants to follow and strengthen the same path. This is the principles of preferring a good solution.
- Negative feedback—By time, the pheromone strength decays. This is the principle of destroying bad solutions. The decay rate is a problem specific parameter where a large rate can destroy good solutions and a little rate can keep bad solutions for longer times.
- Randomness—Path to be taken by ant is completely random, hence generation of new solutions is always possible.
- Multiple interactions—The solution is found byinteraction of many agents Ant-based routing protocols use the metaphor of swarm intelligence to deploy "ants" in the form of control packets to discover routes, reinforce shorter routes via pheromone deposition, and discard longer, less-efficient routes via pheromone evaporation.

There are different types of ant algorithm, which are used for routing in networks such as Ant Based Control algorithm, AntNet algorithm [12], Mobile Ant Based

Routing, AntColony Based Routing Algorithm and Termite. ARA, the Ant Colony Based Routing Algorithm [14] is much like the DSR and is considered in details.

6.2.2.2 ARA-The Ant-Colony Based Routing Algorithm for MANETs

ARA [14] is an on-demand routing approach for mobile multi-hop ad-hoc networks. The approach is based on ant colony optimization meta-heuristic swarm intelligence.

ARA provides three phases, the route discovery, the route maintenance, and the route failure handling.

In *Route Discovery*, the demanding node broadcasts a Forward Ant (FANT) packet and waits for a Backward Ant (BANT) packet. FANT is an agent, which establishes the pheromone track to the source node. In contrast, a BANT establishes the pheromone track to the destination node. The FANT is a packet with a unique sequence number. All neighbor nodes receive the FANT and duplicates are destroyed. A node receiving a FANT for the first time creates a record in its routing table consisting of (destination address, next hop, pheromone value). The node interprets the source address of the FANT as destination address, the address of the previous node as the next hop, and computes the pheromone value depending on the number of hops the FANT needed to reach the node or any other distance parameter. The FANT is duplicated and broadcasted again. When the FANT reaches the destination node, the destination node extracts the information of the FANT and destroys it. Subsequently, it creates a BANT and sends it to the source node. The BANT has the same task as the FANT, i.e. establishing a track to this node. When the sender receives the BANT from the destination node, the path is established and data packets can be sent [14].

In *Route Maintenance* phase, improvement of the routes during the communication is carried out. Subsequent data packets are used to maintain the path. Similar to the nature, established paths do not keep their initial pheromone values forever. When a node relays a data packet toward the destinationit increases the pheromone value between the node and the next hop in its routing table by •, the path to the destinationis strengthened by the data packets. In contrast, the path to the source node is also strengthened.The evaporation process of the real pheromone issimulated by regular decreasing of the pheromone values [14].

Finally, in the *Route Failure Handling* phase ARA handles routing failures, caused by node mobility by recognizing a missing acknowledgement. The node detecting the failure deactivates this link by setting the pheromone value to zero. Then the node searches for an alternative link in its routing table. The packet is sent via this path. Otherwise, the node informs its neighbors, hoping that they can relay the packet. Either the packet can be transported to the destination node or the backtracking continues to the source node. If the packet does not reach the destination, the source has to initiate a new route discovery phase [14].

6.3 Trusted Routing Protocols for MANETs

Trusted routing is the design of a routing protocol to avoid the formation of route that contain misbehaving either malicious or selfish nodes. Global or local trust evaluation model had been propose. The main idea is to reach a numerical measure to identify the trust level of a node.

Nodes behavior is measured to evaluate its trust level. The node behavior can be monitored by the neighboring nodes and its forwarding attitude for packets is the focus in many cases. Promiscuous mode is used where a node listens and records the packets received and then forwarded by its neighbors. This of course drains a lot of battery power and hence is not preferred in MANETs. A local reliable approach has to be found.

Trust and reputation have been recently suggested as an effective security mechanism for ad-hoc networks. There are Common basic functions of the reviewed trust models. Nodes watch their neighbors during a communication and a report is sent to the members of the network. Each node updates trust of its neighbors by combining the report about neighbors and that node's experience with that neighbor. If a node's experience is less than a certain threshold, then that node is excluded from the network.

Using the trust and reputation management scheme to secure ad-hoc networks requires paying close attention to the incurred bandwidth and delay overhead, which so far have been overlooked by most research work. Searching nodes' reputation in a network with a central authority is not difficult. However, the absence of any centralized authority in ad-hoc networks and the bandwidth limitation of these networks make it challenging to trace nodes' reputation accurately. Flooding the network with request messages is a useful tool for data searching in a fully distributed environment. However, since message transfer consumes both bandwidth and energy, trust and reputation management schemes that generate large amounts of traffic by flooding the network with request messages are not. In addition, because trust and reputation information is usually requested by nodes before they start communicating with each other, trust and reputation management schemes with poor latency are not acceptable. Besides that, it complicates the routing process due to the additional computational overhead.

6.3.1 Examples of Trusted Routing Protocols

6.3.1.1 Distributed Trust Model

The Distributed Trust Model [29] makes use of a protocol that exchanges, revokes and refreshes recommendations about other entities. By using a recommendation protocol, each entity maintains its own trust database. This ensures that the trust computed is neither absolute nor transitive. The model uses a decentralized approach to trust management and uses trust categories and values for computing different levels of trust. The integral trust values vary from −1 to 4 signifying discrete levels of trust from complete distrust (−1) to complete trust (4). Each entity executes the recommendation protocol either as a recommender or as a requestor and the trust levels are computed using the recommended trust value of the target and its recommenders. The model has provision for multiple recommendations for a single target and adopts an averaging mechanism to yield a single recommendation value. The model is most suitable for less formal, provisional and temporary trust relationships and does not specifically target MANETs. Moreover, as it requires that recommendations about other entities be passed, the handling of false or malicious recommendations was ignored in their work.

6.3.1.2 Secure Routing Protocol

Secure Routing Protocol (SRP) [30] route discovery protocol mitigates the detrimental effects of such malicious behavior, to provide correct connectivity information. It guarantees that fabricated, compromised or replayed route replies would either be rejected or never reach back the querying node. The proposed protocol is capable of operating without the existence of an on-line certification authority or the complete knowledge of keys of all network nodes. Its sole requirement is that any two nodes that wish to communicate securely can simply establish a priori a shared secret, to be used by their routing protocol modules.

The routing misbehavior is mitigated by including components like Watchdog and Pathrater in the scheme proposed in [10]. Every node has a watchdog process that monitors the direct neighbors by promiscuously listening to their transmission. No penalty for the malicious nodes is implicated.

6.3.1.3 Trusted AODV

Trusted AODV (TAODV) [10] is a secure routing protocol, this protocol extends the widely used AODV routing protocol and employs the idea of a trust model in subjective logic to protect routing behaviors in the network layer of MANETs. TAODV assumes that the system is equipped with some monitor mechanisms or intrusion detection units either in the network layer or the application layer so that one node can observe the behaviors of its one-hop neighbors [31]. In the TAODV, trust among nodes is represented by opinion, which is an item derived from subjective logic. The opinions are dynamic and updated frequently. Following TOADV specifications, if one node performs normal communications, its opinion from other nodes' points of view can be increased; otherwise, if one node performs some malicious behaviors, it will be ultimately denied by the whole network. A trust recommendation mechanism is also designed to exchange trust information among nodes.

6.3.1.4 Cooperation of Nodes: Fairness in Dynamic Ad-Hoc Networks

Cooperation of Nodes: Fairness in Dynamic Ad-Hoc Networks (CONFIDANT) [32] is an extension to the DSR protocol and it aims at detecting and isolating misbehaving nodes in order to discourage the uncooperative behavior of nodes during the routing process. CONFIDANT is composed of four main components. The Monitor, the Reputation System, the Path Manager, and the Trust Manager. The components are present in every node.

(1) The Monitor (Neighborhood Watch): With the aim to detect a non-compliant participant in the network, nodes watch their neighbors when the later are forwarding packets. When the behavior observed is different from the behavior expected, a node failing to behave as expected is reported by sending the ALARM message to warn other nodes.
(2) The Trust Manager: It is the responsibility of the trust manager to send the ALARM message and to receive them. ALARM messages are sent by nodes that experience or observe misbehavior, and a node that gets an ALARM also broadcasts it to other nodes. Each node has a list of friends from which ALARM messages can be accepted. Before any reaction is taken after receiving an ALARM message, the trustworthiness of a friend sending the message is verified.
(3) The Reputation System (Node Rating): Each node has a local list of friends and their rating and a black list containing nodes with bad rating; these lists are shared among friends so that rating can be decentralized. The reputation system manages a table that contains a list of nodes and their rating. Rating of a node is only changed if there are enough evidences about the malicious behavior suspected. i.e. if the malicious behavior happened more than times that are acceptable as number of accidents. The rate is changed according to the way the behavior has been detected. The greatest weight is assigned to own experience, a small rate for the observation in the neighborhood and a smaller rate for the reported behavior.
(4) The Path Manager: The path manager ranks the routes according to reputation of the nodes in the path and it deletes routes that contain malicious nodes. The path manager also takes action when a malicious node requests a route. Usually a malicious node is ignored when it requests a route.

6.3.1.5 A Collaborative Reputation Mechanism for Node Cooperation in Ad-Hoc Networks (CORE)

CORE adopts the approach of reputation as a foundation of trust mechanism to solve problems caused by misbehaving nodes in the network [33]. As it is done in CONFIDANT, in CORE also nodes observe their neighbors or get information about the behavior of other nodes in the network from network members. Reputation is formed by combining information a node gets from its own experience and information that node gets from other nodes about a particular node of interest. A final reputation of a node is formed by a combination of three types of reputation; subjective reputation, indirect reputation and functional reputation.

Subjective reputation is got from direct current observations and past experience with a node and its neighbors. Experience is given more weight for irregular misbehavior not influenced by the final reputation value. Subjective reputation is calculated at a particular time by a particular node. In calculation of subjective reputation, a past values is more relevant. The past value is got from a number of observations, and a scale from -1 for an unexpected behavior observed to $+1$ for an expected behavior observed is used. When the observed behaviors are conclusive, enough during a certain period of time, then the value 0 is used for neutral experience.

Indirect reputation is calculated with the aid of information that a node gets from other members of the network. Negative information from members of the network are not considered, thus indirect reputation can only take positive values. The cause of consideration of positive values is to avoid denial of service attacks by providing negative reports about well-behaved nodes.

Functional reputation is the combination of direct and indirect reputation with respect to different tasks like packet forwarding or routing that nodes are usually supposed to perform for each other.

6.3.1.6 Dependable DSR

Dependable DSR without a Trusted Third Party [34] is a technique of discovering and maintaining dependable routes in MANETs even in the presence of malicious nodes. Each node in the network monitors its surrounding neighbors and maintains a direct trust value for them. These values are propagated through the network along with the data traffic. This permits evaluation of the global trust knowledge by each network node without the need of a trusted third party. These trust values are then associated with the nodes present in the DSR link cache scheme. This permits nodes to retrieve dependable routes from the cache instead of standard shortest paths.

6.3.1.7 ATDSR: Agent-based Trusted On-Demand Routing Protocol for Mobile Ad-Hoc Networks

ATDSR [15] uses a multi-agent system (MAS) consisting of two types of agents cooperating with each other to achieve the required task; specifically monitoring agent (MOA) and routing agent (ROA). MOA is responsible for monitoring its hosting node behavior in the routing process and then computing the trust value for this node. ROA is responsible for using the trust information and finding out the most trustworthy route for a particular destination.

ATDSR Assumptions

(1) Every participating node in the network will hold, install and execute its multi-agent system consisting of the MOA and the ROA. (2) All the trust evaluations

are based on objective trust and maintained locally by the nodes monitoring agent MOA. (3) Node's unauthorized modification to its locally stored trust information can be identified. (4) The agents are resilient against the unauthorized analysis and modification of their computation and messages. Obviously, nodes cannot manage and compute their own trust value.

Node's Trust Value Calculation

In ATDSR, MOA is the only source responsible for evaluating and maintaining its hosting node trust value. Thus, ATDSR guarantees that there is always a single trust value for each node in the network. It gives quite an accurate estimation about the trustworthiness of a node. There is no need for each node to operate in a promiscuous mode, which saves nodes batteries. In addition to that, the nodes themselves are able to provide their own trust information whenever requested; therefore, trust computation is done without network wide flooding and with no acquisition-latency.

To reflect this kind of node's "selective forwarding" behavior, ATDSR compute *Trust-Value* (*M*) as:

$$\text{Trust-Value (M)} = \frac{\text{HF(M)} * \text{Pkt\_Size(HF(M))}}{\text{RF(M)} * \text{Pkt\_Size(RF(XM))}} \tag{6.1}$$

where,

- $RF(M)$: The total number of packets that all nodes have transmitted to node M for forwarding.
- $HF(M)$: The total number of packets that have been forwarded by node M.
- $Pkt\_Size(RF(M))$: Total packet's size of $RF(M)$.
- $Pkt_{Size}(H_F(M))$: Total packet's size of $H_F(M)$.

Trust Values Propagation

ATDSR depends on transferring the trust values of nodes through the Route Request packets. In addition to the routing information, each Route Request (RREQ) packet contains a trust record in which is accumulated a record of the sequence of hops taken by the RREQ packet as it is propagated through MANETs during this Route Discovery process. Each sending node S builds its own trust evaluation table Teval(S) using the propagated trust values in the network. Teval(S) contains the trust value of all other nodes in the network. Using these trust information, the sending node routing agent ROA(S) is responsible for computing the most trustworthy route to a particular destination. If the most trustworthy route trusts value is found lower than a threshold value (denoted by $\text{R}_{\text{threshold}}$). The route is rejected and a new Route Discovery process is initiated. The trust value in route R by source node S is represented as $\text{T}_\text{s}(\text{R})$ and given by the following equation:

$$\text{T}_\text{s}(\text{R}) = \min(\text{Trust-Value}(\text{N}_\text{i}))\,\forall\,\text{N}_\text{i} \in \text{R} \tag{6.2}$$

ATDSR phases
In ATDSR,

- For each arbitrary node N_i, its MOA, locally maintains a trust evaluation table $T_{eval}(N_i)$. The trust evaluation table $T_{eval}(N_i)$ contains the trust value of other nodes in the network.
- Each node ROA is responsible of appending the computed trust value of its hosting node to the RREQ packet during the Route Discovery process.
- $T_s(R)$ field is appended in the RREP packet.

Route Discovery

The Route Discovery includes three processes, (a) RREQ Delivery; (b) RREP Delivery; (c) Route Selection.

RREQ Delivery

When a source node S needs to send data to the destination node D, it first checks whether there is a feasible path found between S and D. If so, S sends the data to D; otherwise, S will start a Route Discovery. First, the ROA of S appends its node ID and trust value (Trust-Value (S)) to the RREQ target D, then broadcast the RREQ packet and set a timer window tw and the route threshold value Rthreshold at the same time.

When any intermediate node receives a RREQ packet, it processes the request according to the following steps:

Step 1: If the pair {Source ID, Request ID} for this RREQ packet is found in this node's list of recently seen requests, then it discards the RREQ packet and does not process it further.
Step 2: Otherwise, if this node's address is already listed in the route record in the request, then it discards the RREQ packet and does not process it further.
Step 3: Otherwise, if the target of the request matches this node's own address, then the route record in the packet contains the route by which the request this node from the source node of the RREQ packet. Intermediate node returns a copy of this route in a RREP packet to the source node.
Step 4: Otherwise, its ROA appends the node ID and Trust-Value to the route record in the RREQ packet and re-broadcast the request to the neighbor nodes.

RREP Delivery

When the destination node receives the first RREQ packet, it sets a timer window td. If td expired, it discards the follow up RREQ packet. Otherwise, its ROA computes Ts(R) according to the formula (2), and then unicasts the RREP packet with Ts(R) to the intermediate node.

Route Selection

Step 1: When S receives the RREP packet, it checks whether the timer window tS expired. If tS expired, S discards follow-up RREP.

Step 2: Otherwise, it compares $T_s(R)$ with $R_{\text{threshold}}$. If $T_s(R) > R_{\text{threshold}}$, S discards the RREP.
Step 3: Otherwise, it adds the route R to the Route Cache.
Step 4: if there is at least a valid route to D, it picks a path with the largest Ts(R).
Step 5: Otherwise, S waits until tS expires, then it initiates a new Route Discovery process.

Route Maintenance

After each successful route discovery takes place, S can deliver its data to D through a route. However, the route may break at any time instant due to the mobility of nodes, or attacks. In order to maintain a stable, reliable and secure network connection, route maintenance is necessary to ensure the system survivability. Route Maintenance is used to detect if the network topology has changed such that the link used by this packet is broken. Each node along the route, when transmitting the packet to the next hop, is responsible for detecting if its link to the next hop has broken. When the retransmission and acknowledgement mechanism detects that the link is broken, the detecting node returns a Route Error packet to the source of the packet. The node will then search its route cache to find if there is an alternative route to the destination of this packet. If there is one, the node will change the source route in the packet header and send it using this new route. This mechanism is called "salvaging" a packet. When a Route Error packet is received or overheard, the link in error is removed from the local route cache, and all routes, which contain this hop, must be truncated at that point. The source can then attempt to use any other route to the destination that is already in its route cache, or can invoke Route Discovery again to find a new route.

6.4 Trusted Ant Colony Based Routing Algorithm for Mobile Ad-Hoc Networks

Trusted Ant Colony Based Routing Algorithm for Mobile Ad Hoc Networks (TARA) is an ant colony based trusted routing protocol for MANETs. The protocol uses an objective trust model and relies on a multi-agent system (MAS). The following assumptions are made.

1. Each node participating in the network must install the MAS and performs all routing through it.
2. The MAS is assumed resilient against modifications from the hosting node.
3. A node cannot change its trust or pheromone values within the routing table.

The multi-agent system consists of two types of agents, static agents and mobile agents.

The static agents are responsible for building the routes, maintaining the routes, evaluating the routes and selecting the optimal routes for data packets, and evaluating the node trust value.

Table 6.1 TARA set of agents

Agent name	Acronym	Type	Responsibility
Routing Agent	RA	Static	RA performs all the tasks for building the route and selecting the best route from the routing table for a specific destination
Node Evaluation Agent	NEA	Static	NEA is responsible for monitoring its hosting node behavior in the routing process and then computing the trust value for this node
Route Maintenance Agent	RMA	Static	RMA receives control and data ants. Accordingly, it updates the pheromone value of each link in the routing table. It also decreases it periodically
Route Ant Agent	RAA	Mobile	RAA travels the network to find the possible routes in both forward and backward directions.
Key Exchange Ant Agent	KEAA	Mobile	KEAA is used whenever a secure end-to-end or link encryption is to be used. Usually KEAA is combined with RAA. KEAA is not discussed here and left for future expansions
Data Carrier Agent	DCA	Mobile	DCA is the ant responsible for carrying the actual payload data. In addition, it interacts with RMA to strengthen the links it travels on.

The mobile agents are the ants scouting the network. They are responsible for carrying the data and control packets throughout the network. Control packets including route building and maintaining related ants and key exchange ants.

The list of agents are described in the Table 6.1.

Before going to the details of TARA, some constants and thresholds are to be defined.

- Trust Threshold: This is value defines the minimum trust value for a route within the MANET.
- Maximum number of hops: This value determines the maximum route length. No route to exceed this length is accepted.
- Δ_{RD}: The pheromone decrement value used within route building process.
- Δ_{RS}: The pheromone strengthen increment used when data packets are sent through a valid route.
- Δ_{RV}: The pheromone evaporation constant used during route maintenance operation.
- α: Size/Data weight factor used when calculating the node trust value.
- β: Trust/Performance weight factor used when selecting a route to a specific destination.

6.4.1 Routing Table Data Structure

In TARA, a unique routing table that contains, plus the normal next hop id, other values to indicate both the trust and link quality values. The table record is defined as follows, for route number j;

$$(R_j, N_d, N_n, T_r(R_j), P_j(N_n), T_j)$$

The fields of the routing table are described as follows:

- R$_j$: The route ID. This is a sequence number to identify each route and is local to the node.
- N$_d$: The destination node ID for this route.
- N$_n$: The node ID of the next hop for this route.
- T$_r$(R$_j$): The trust value of the route.
- P$_j$(N$_n$): The pheromone value of the link from the source node to the next hop in this particular route.
- T$_j$: The time stamp when the route parameters were last updated.

6.4.2 Node Trust Evaluation

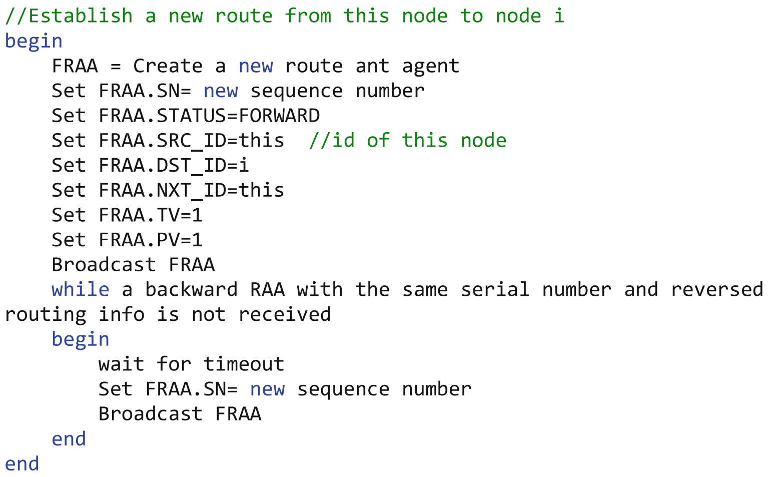
In TARA, an objective self-trust value is calculated by the NEA. This value is updated whenever routing is performed. NEA monitors its hosting node routing behavior and four quantities are measured.

- R$_f$: The number of all packets this not has received and was asked to forward to another node.
- H$_f$: The number of packets actually forwarded by this node
- SR$_f$: The total size of all packets this node has received and was asked to forward to another node.
- SH$_f$: The total size of all packets actually forwarded by this node.

The trust value is calculated as follows

$$T_r(N_i) = \alpha \frac{H_f(N_i)}{R_f(N_i)} + (1-\alpha)\frac{SH_f(N_i)}{SR_f(N_i)} \tag{6.3}$$

The trust value calculation takes into consideration both the size and number of packets forwarding behavior. This trust calculation formula also considers the selective behavior of nodes that forwards only small packets and ignores large ones. The constant is chosen to adjust the weight of (number/size) forwarding issue. It is chosen in the range of $0 \leq \alpha \leq 1$.



```
//Establish a new route from this node to node i
begin
    FRAA = Create a new route ant agent
    Set FRAA.SN= new sequence number
    Set FRAA.STATUS=FORWARD
    Set FRAA.SRC_ID=this  //id of this node
    Set FRAA.DST_ID=i
    Set FRAA.NXT_ID=this
    Set FRAA.TV=1
    Set FRAA.PV=1
    Broadcast FRAA
    while a backward RAA with the same serial number and reversed
routing info is not received
    begin
        wait for timeout
        Set FRAA.SN= new sequence number
        Broadcast FRAA
    end
end
```

Fig. 6.3 Initialize route discovery process

6.4.3 Route Discovery

In route discovery phase, a node wishing to send a packet searches first in its routing table for valid route. If it is not found, a Route Ant Agent (RAA) is generated and broadcasted to all neighbors. The source node then waits for a pre-defined timeout period and if a backward RAA with the same serial number and reversed source/destination, then it will create a new RAA and broadcast it. The RAA contains the following information:

- Sequence Number (SN): Each RAA is given a unique sequence number.
- Status (STATUS): the RAA status is set to FORWARD.
- Source Node ID (SRC_ID): The requesting node ID.
- Destination Node ID (DST_ID): The final destination for the route.
- Next Hop ID (NXT_ID): The next hop ID for the route. It is initialized by the destination ID.
- Trust Value (TV): The trust value of the route. RAA is loaded initially with 1.
- Pheromone (PV): RAA pheromone value is initialized to 1.

The following pseudo code (Fig. 6.3) represents the route request process.

The RAA in its forward status is responsible for making a back track route from the destination and all intermediate nodes to the source node. When an RAA is received by a node, it is forwarded immediately to the RA and the following processing is doe as shown in the pseudo-code (Fig. 6.4).

From Fig. 6.4, a node receiving a forward RAA checks if it had received it before, in this case it is destroyed. Otherwise, if this is an intermediate node, an entry is created in its routing table setting the RAA source as the destination, the RAA next hop, trust value, and pheromone value as their corresponding values in the route

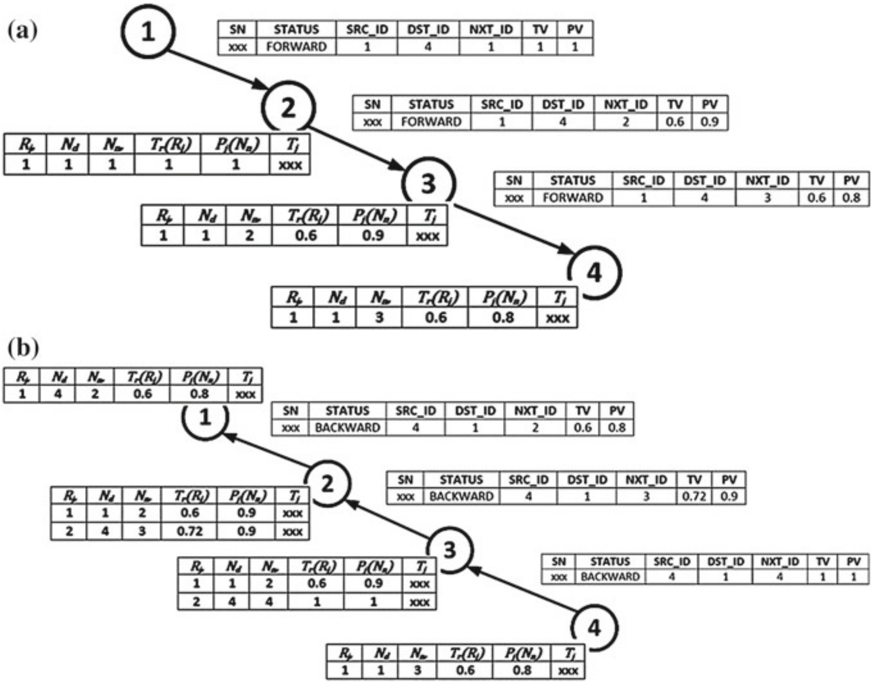
```
Begin
    Check the received ant agent RAA.SN, RAA.SRC_ID, RAA.DST_ID.
    if RAA previously received
    begin
        Destroy RAA
        Stop Processing
        exit
    end
// Create a new route and add it to the routing table
    ROUTE=Create new route entry in routing table
    Set ROUTE.Rj=new sequence number
    Set ROUTE.Nd=RAA.SRC_ID
    Set ROUTE.Nn=RAA.NXT_ID
    Set ROUTE.Tr=RAA.TV
    Set ROUTE.Pj=RAA.PV
    Set ROUTE.Tj=NOW

    // Is this an intermediate node
    if RAA.DST != this
    begin
        RAAN= Duplicate(RAA)  // Copy all fields
        // Update the new RAA
        Set NRAA.NXT_ID=this  // change the next hop to this node ID
        Set NRAA.TV=min(RAA.TV, thisnode.TV)
        Set NRAA.PV=RAA.PV- $\Delta_{R D}$
        if NRAA.PV==0 or NRAA.TV < Trust_Threshold
        begin
            Destroy RAA
            Stop Processing
            exit
        else
            Broadcast NRAA
        end
     end
    //Is this the final destination
    if RAA.DST = this
    begin
        // Create a backward ant agent
        BRAA = Create a new RAA
        Set BRAA.SN= RAA.SN
        Set BRAA.STATUS=BACKWARD
        Set RAA.SRC_ID=this
        Set RAA.DST_ID=RAA.SRC_ID
        Set RAA.NXT_ID=this
        Set RAA.TV=1
        Set RAA.PV=1
        Broadcast BRAA
        Destroy RAA
     end
   end
```

Fig. 6.4 Processing of ARA route discovery ant aget

entry and adds a time stamp. Then the node updates the RAA next hop by its own id and decreases the RAA pheromone vale by a predefined value Δ_{RD}.

$$\Delta_{RD} = \frac{1}{maximum\, no\, of\, hops}$$



(a)

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	FORWARD	1	4	1	1	1

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	FORWARD	1	4	2	0.6	0.9

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	1	1	1	xxx

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	FORWARD	1	4	3	0.6	0.8

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	2	0.6	0.9	xxx

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	3	0.6	0.8	xxx

(b)

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	4	2	0.6	0.8	xxx

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	BACKWARD	4	1	2	0.6	0.8

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	BACKWARD	4	1	3	0.72	0.9

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	2	0.6	0.9	xxx
2	4	3	0.72	0.9	xxx

SN	STATUS	SRC_ID	DST_ID	NXT_ID	TV	PV
xxx	BACKWARD	4	1	4	1	1

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	2	0.6	0.9	xxx
2	4	4	1	1	xxx

R_j	N_d	N_n	$T_r(R_j)$	$P_i(N_n)$	T_j
1	1	3	0.6	0.8	xxx

Fig. 6.5 Illustration of route discovery process. **a** Forward Pass. **b** Backward Pass

Δ_{RD} is chosen that way so no routes is created that exceeds the maximum number of hops. If t he new pheromone value is zero, the RAA is destroyed and route discovery ends.

Finally, the RAA trust value is compared with the node trust value obtained from the NEA. The minimum of these two values is used for the next forwarded ant agent. In this way, the route trust value will be the minimum trust value of all nodes along the route. In contrast to ATDSR, the trust value need not be propagated to other nodes minimizing the communication overhead and the memory needed to store the local trust value table of other nodes.

In addition, if the new RAA trust value is below the Trust Threshold value, the route discovery is also terminated. In this way, no malicious route is ever created.

A backward RAA is treated the same way but it establishes the route forward path and it can be different from the backward path. It is also used to end the route request session if it arrives back before the time out.

Figure 6.5 shows an example of a 3-node route established between nodes 1 and 4 where (a) shows the forward RAA and (b) shows the backward RAA. For demo purpose the maximum number of hops is assumed to be 10 and the trust values of nodes 1, 2, 3, and 4 is assumed 0.8, 06, 0.72 and 0.3 respectively.

6.4.4 Route Selection and Route Failure Handling

When a node wishes to send a packet to a destination, it finds all possible routes from its routing table. All routes with pheromone ($P_j\ (N_n)$) value of zero is excluded.

For each valid route, a route quality value is calculated as follows

$$Q(R_j) = \beta T_r + (1 - \beta) P_j(N_n) \tag{6.4}$$

where β is the trust/performance weight factor and is chosen in the range

$$0 \leq \beta \leq 1$$

If β is chosen to be zero then TARA becomes pure ARA. On the other hand, if β is one then TARA behaves like ATDSR.

The route with the maximum quality is selected for delivering the packet.

Again, a timeout is waited and if an acknowledgement is not received, the pheromone value ($P_j(N_n)$) is divided by 2. In contrast to ARA, the pheromone value in RA is set to zero and the route is invalidated. However, the bi-division of the pheromone is chosen to accommodate the loss of acknowledgement packet due to mobility and then the route is kept for a while hoping the mobility pattern of node will bring nodes of the route within the range of each other again.

A search for the next quality route is done and if not found a new route discovery is initiated.

6.4.5 Route Maintenance

One property of ant colony optimization for food finding is that the pheromone value of a specific path is not constant with time. Rather, each ant following the path will deposit a new amount of pheromone making the path stronger. This process is known as path strengthen. On the other hand, the pheromone evaporate with time and is weekend. This process is known as pheromone evaporation. The result of the two processes is to prefer routes with more ants on them and to discard abandoned paths and is believed to help selecting the best route.

This process is simulated in TARA by the RMA and is described by the following:

- Whenever a new data packet is sent or forwarded to the next hop over a certain route, its pheromone value $P_j(\mathrm{N}_n)$, is incremented by Δ_{RS} and its time stamp T_j is reset to the current time. This simulates the strengthening process.
- Periodical the RMA decreases all the routes pheromone value $P_j(N_n)$ by Δ_{RS} if there pheromone value had not changed since the last decrement. The timestamp value T_j is checked to select those routes to experience the evaporation.

Table 6.2 Communication and operation complexity parameters

Variable	Description
N	Number of nodes in the network
D	Diameter of the network
Y	Total number of nodes forming the route from *SRC* to *DST*
R	Number of available routes from *SRC* to *DST*
Z	Diameter of Y; the directed path where the backward ARA packets transits

6.4.6 TARA Complexity Analysis

This section introduces an analysis to the performance of the TARA protocol based on two factors; operation complexity (OC) and communication complexity (CC). OC can be defined as the number of steps required in performing the protocol operation, while CC can be defined as the number of messages exchanged in performing the protocol operation. The values represent the worst-case analysis. Besides that, the variables defined in Table 6.2 are used when referring to Operation and communication complexities.

6.4.6.1 Analysis of Route Discovery Phase

The route discovery phase in TARA protocol is accomplished by broadcasting anARA ants and waiting for the backward ARA. The maximum number of nodes receiving this ant agent is all nodes in the network (N). The destination then sends a backward ARA and back to the source along this established path in which (Y) are participating.

Thus, the initial route discovery phase will require exchanging of $\mathrm{N}+\mathrm{Y}$ messages. This is the case of ATDSR and ARA.

On the other hand, broadcasted query message will traverse the whole network diameter in searching for the destination, so it needs to perform steps equal to the network diameter D, while the reply message will traverse the diameter of all the nodes forming the route from the destination node to the source node (Z). Thus, the initial route discovery phase will require performing of $\mathrm{D}+\mathrm{Z}$ steps.

6.4.6.2 Analysis of Nodes Evaluation Phase

In TARA, the evaluation of the node trust is performed localy. The trust value need not be transferred and is embedded directly in the routing table. So there is no extra messages to be transferred to propagate the trust values. Unlike ATDSR this step CC is zero.

Table 6.3 Comparative complexity of TARA, DSR, DDSR, ATDSR and ARA

Protocol phase	Route discovery		Nodes evaluation		Routes evaluation	
	OC	CC	OC	CC	OC	CC
TARA	O(D + Z)	O(N + Y)	0	0	O(2R)	0
DSR	O(2D)	O (2N)	N/A	N/A	N/A	N/A
DDSR	O(2D)	O(N + Y)	O (N^2)	0	O(ZR)	O(RN^2)
ATDSR	O(D + Z)	O(N + Y)	O(1 + Z)	O(3ZR)	O(2R)	0
ARA	O(D + Z)	O(N + Y)	N/A	N/A	N/A	N/A

Also no extra steps are required, so this operation OC is also zero.

6.4.6.3 Analysis of Routes Evaluation Phase

The routes evaluation phase does no t require message transfer, so the communication complexity of this process is zero. On the other hand, the source node computes the route's quality value for each available route. Then, the source node picks a path with the largest quality; so OC for this process is 2R.

Table 6.3 summarizes the complexity analysis of TARA, ARA, DSR, DDSR and ATDSR.

6.5 Performance Evaluation of TARA

6.5.1 Simulation Environment

TARA is simulated and its performance is compared against DSR and ATDSR. ARA performance is very similar to DSR so one of them is enough.

For the simulation parameters:

1. Data packet size was chosen to be 512 bytes.
2. The wireless transmission range of each node was $T_r = 250$ m.
3. In the simulations, $N_t = 100$ mobile nodes randomly distributed over a 700 (m) by 500 (m) flat area.
4. The source and the destination nodes were randomly chosen among all nodes in the network.
5. The total simulation time was 900 s.
6. Both UDP and TCP traffics have been simulated to evaluate the performance of the model.
7. A random way-point mobility model was assumed with a maximum speed of $V_m = 10$ to 50 (m/s) and a pause time of 0 sec.

8. The mobility scenarios were generated by the “random trip” generic mobility model.
9. Constant Bit Rate (CBR) traffic was used. Each simulation included 50 CBR sessions, each of which generated four packets per second.

A total of three simulations were conducted to evaluate the performance of TARA under varying number of malicious nodes, mobility and node density. The parameters discussed above are common to all three simulations. To evaluate the performance of the proposed scheme, we use the following metrics [34]

- *Packet Delivery Ratio* (*PDR*): The ratio between the number of packets received by the destination nodes to the number of packets sent by the source nodes.
- *Routing Packet Overhead* (*RO*): The ratio between the total number of control packets generated to the total number of data packets received during the simulation time.
- *Average Latency* (*AL*): The mean time taken by the packets to reach their respective destinations (in ms).
- *Path Optimality* (*PO*): The ratio between the number of hops in the shortest possible path to the number of hops in the actual path taken by a packet to reach its destination.
- *Energy Consumption* (*EC*): The amount of energy (in Joules) consumed per node during the simulation time.
- *Probability of Detection* (*PD*): The ratio between the number of nodes whose behavior (malicious or benevolent) is identified correctly, to the actual number of such nodes present in the network.

Figure 6.6, below shows the simulated network where normal nodes are represented by blue circles while malicious nodes are represented by red xs.

6.5.2 Simulation Results and Analysis

(1) *Simulation 1*: Varying the number of malicious nodes

The number of malicious nodes varies from 5 nodes to 35 nodes, with a 5 nodes increment. Figure 6.7 shows the performance results of the TARA compared to ATDSR and the standard DSR.

1. The packet delivery ratio in standard DSR is lower than the other two protocols. This can be attributed to the fact that standard DSR does not take into account the benevolence levels of the nodes and prefers shorter routes by default. The malicious nodes are constantly selected in the routing process, which leads to an overall lower through put of the network. On the other hand, TARA and ATDSR select or deselect route nodes based upon their trust levels and thus attempts to avoid any malicious nodes.

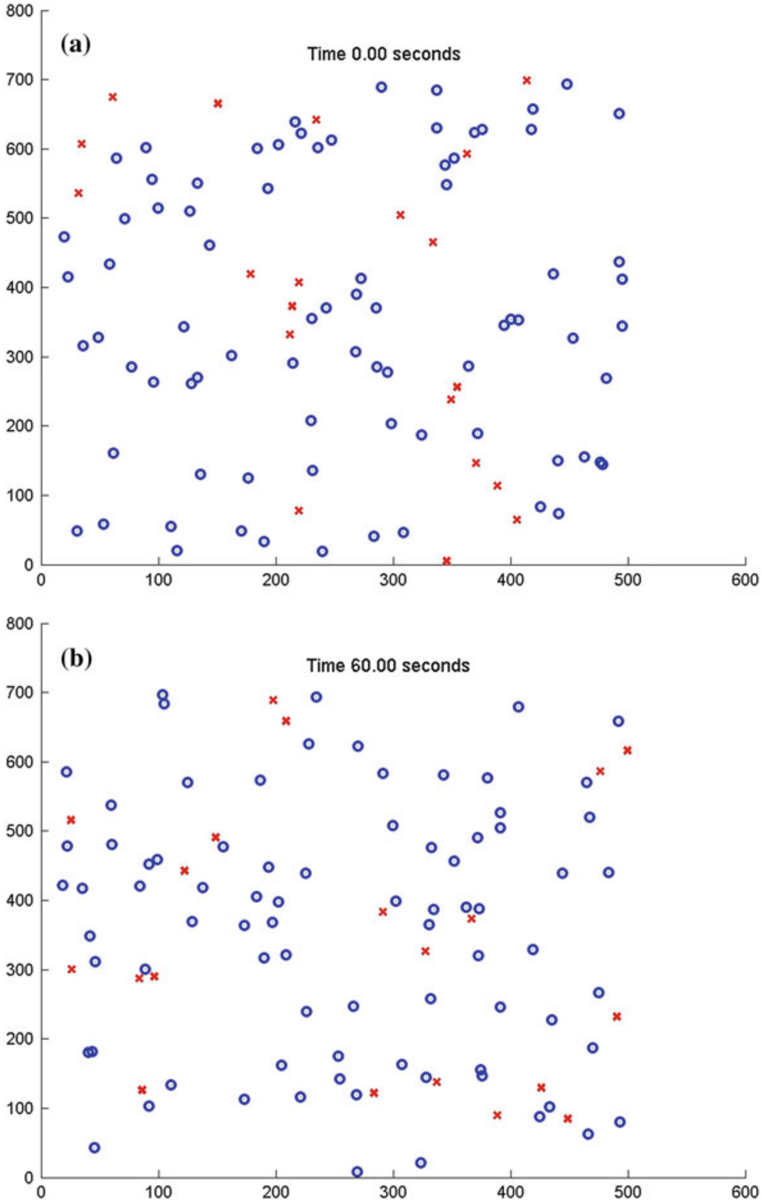


Fig. 6.6 Simulated network with 100 nodes and 20 % malicious nodes. **a** Start of simulation. **b** After 60 s

2. The routing packet overhead of ATDSR increases with the increase in the number of malicious nodes due to the additional route discoveries initiated in search of trusted routes. TARA has a better routing packet overhead than ATDSR because

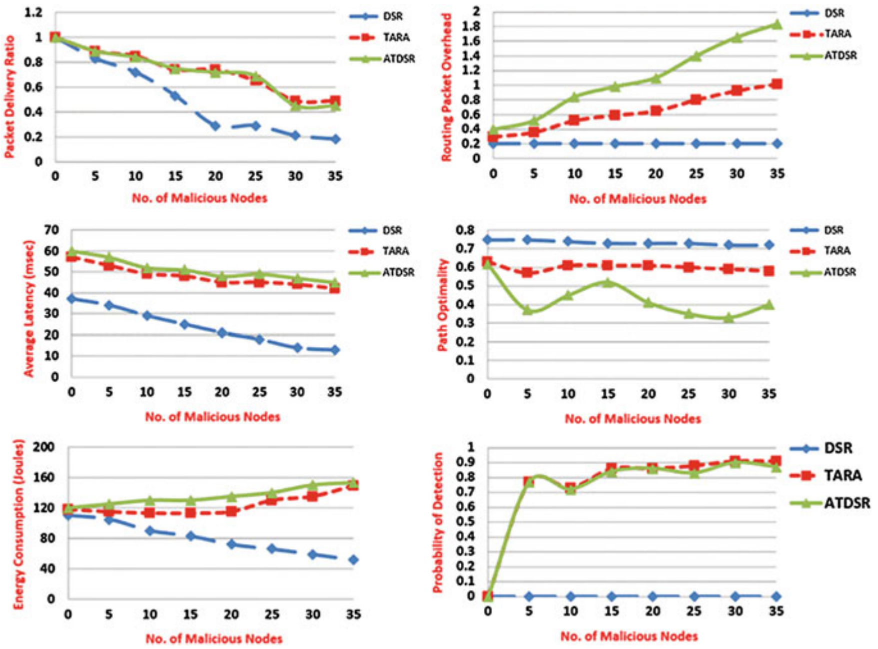


Fig. 6.7 TARA performance against number of malicious nodes

it requires less control packets and there is no trust propagation. The process of finding the trusted routes results in longer routes, takes more time, and hence increases the average latency and path optimality.

3. The energy consumption of all nodes increases when the TARA protocol is engaged. The energy consumption primarily increases due to the improved number of data packets that are now being transferred on trusted but longer routes in the network and because of installing the multi-agent system in each participating node.

(2) *Simulation 2*: Varying Pause Time

The metrics are measured by varying the pause time and thus the extent of mobility in the network, a higher pause time results in a network with lower mobility and a pause time of 900 s (same as the simulation duration) results in a completely static network. Figure 6.8 shows the performance results of the TARA compared to ATDSR and the standard DSR.

The following observations can be made from the results. The packet delivery ratio increases with increasing the pause time due to the network stability. TARA reduces the overhead by exploiting the mobility characteristics of mobile nodes. More importantly, this reduction causes similar reduction in the energy consumption per node. The average latency and path optimality have not a noticed affect.

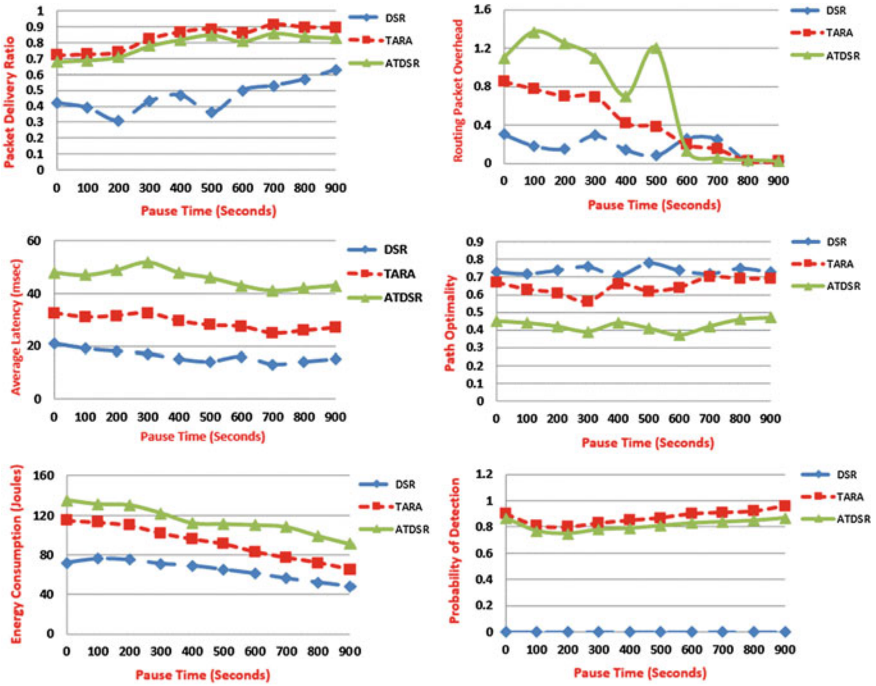


Fig. 6.8 TARA performance against pause time

(3) *Simulation 3*: Varying Network Density

This simulation is to study the impact of network density by comparing the routing protocols for various network densitiesin Fig. 6.9. A medium mobility scenario with a pause time of 100 s is chosen with constant ratio 20 % of malicious nodes.

Increasing the number of nodes leads toan increase in the number of neighbors for each forwarder node. If there are more neighbors, a data packet can generally increase the physical distance traveled at each hop. This causes a noticed increase in the path optimality and a decrease in the average latency. However, the routing packet overhead increases with the number of nodes. Therefore, the probability of collision increases as more neighbors compete to broadcast the same request, whichcauses higher energy consumption per node. The packet delivery ratios and the probability of detection do not have a noticed affect.

6.6 Conclusion

Mobile ad-hoc networks are spreading and due to their own nature, they are subject to adversaries who can compromise nodes and exploit the routing mechanism. Security is very important, but it is also difficult due to the nature of these networks. Trusted

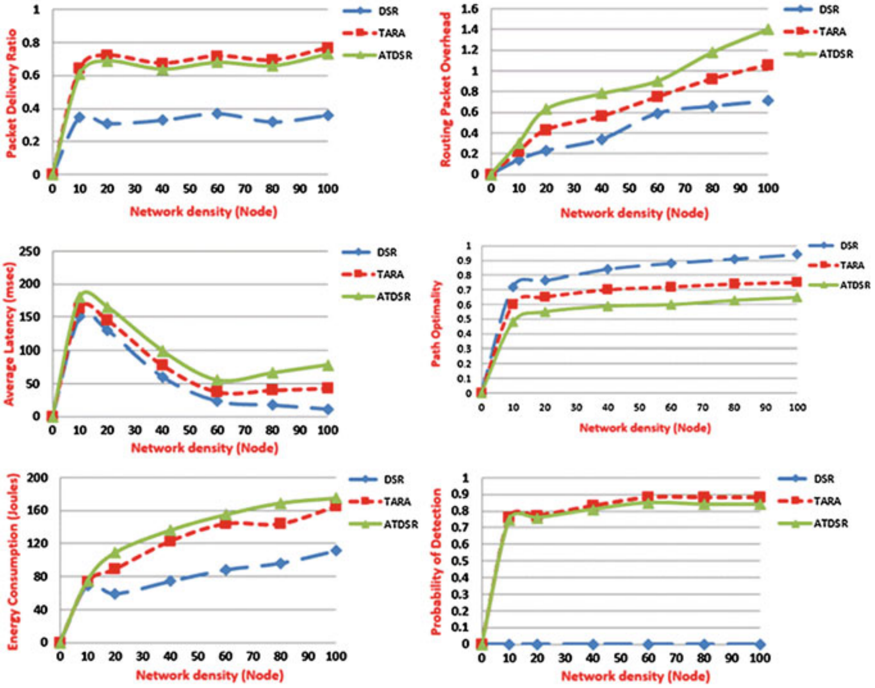


Fig. 6.9 TARA performance against network density

routing protocols are one means of providing security. This chapter presented an Intelligent Ant Colony Multi Agent based Trusted on demand Routing Protocol for MANETs called TARA. TARA uses a methodology to establish an objective trust value for each node based on self-monitoring. The protocol is tuned to minimize the number of messages to be exchanged and trust value propagation is eliminated. Ant Colony Optimization is used to find the optimal route resulting in a better performance and lower latency. The advantages and complexity of TARA are examined and it is compared with other protocols. Simulation results show that TARA works better than standard and trusted routing protocols in the presence of malicious nodes.

References

1. Raju, J., Garcia-Luna-Aceves, J.: A comparison of on-demand and table driven routing for ad-hoc wireless networks. In: IEEE International Conference on Communications, 2000, (ICC 2000), (2000)
2. Perkins, C.E., Royer, E.M.: Ad hoc on-demand distance vector routing. In: Second IEEE Workshop on Mobile Computing Systems and Applications, WMCSA'99 (1999)

3. Johnson, D.B., Maltz, D.A.: Dynamic source routing in ad hoc wireless network. In: Kluwer International Series in Engineering and Computer, Science, pp. 153–179 (1996)
4. Park, V., Corson, M.S.: Temporally-ordered routing algorithm (TORA) version 1 functional specification. Internet-Draft, draft-ietf-manet-tora-spec-00.txt (1997)
5. Yih-Chun, H., Perrig, A.: A survey of secure wireless ad hoc routing. IEEE Secur. Priv. **2**(3), 28–39 (2004)
6. Hu, Y.-C., Perrig, A., Johnson, D.B.: Ariadne: A secure on-demand routing protocol for ad hoc networks. Wireless Netw. **11**(1), 21–38 (2005)
7. Perrig, A., Canetti, R., Tygar, J.D., Song, D.: The TESLA broadcast authentication protocol. Cryptobytes **5**(2), 1–13 (2005)
8. Hu, Y.-C., Johnson, D.B., Perrig, A.: SEAD: Secure efficient distance vector routing for mobile wireless ad hoc networks. Ad Hoc Netw. **1**(1), 175–192 (2003)
9. Sun, Y.L., Yu, W., Han, Z., Liu, K.R.: Information theoretic framework of trust modeling and evaluation for ad hoc networks. IEEE J. Sel. Areas Commun. **24**(2), 305–317 (2006)
10. Li, X., Lyu, M.R., Liu, J.: A trust model based routing protocol for secure ad hoc networks. In: 2004 IEEE Aerospace Conference (2004)
11. Pirzada, A., McDonald, C.: Reliable routing in ad hoc networks using direct trust mechanisms. In: Cheng, M., Li, D. (eds.) Advances in Wireless Ad Hoc and Sensor, Networks, pp. 133–159. Springer, New York (2008)
12. Di Caro, G., Dorigo, M.: AntNet: Distributed stigmergetic control for communications networks. J. Artif. Intell. Res. **9**, 317–365 (1998)
13. Di Caro, G., Dorigo, M.: Mobile agents for adaptive routing. In: Proceedings of the Thirty-First Hawaii International Conference on System Sciences (1998)
14. Gunes, M., Sorges, U., Bouazizi, I.: ARA-the ant-colony based routing algorithm for MANETs. In: Proceedings of International Conference on Parallel Processing Workshops, 2002 (2003)
15. Halim, I.T.A., Fahmy, H.M., Bahaa-ElDin, A.M., El-Shafey, M.H.: Agent-based trusted on-demand routing protocol for mobile ad-hoc networks. In: 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM) (2010)
16. Chlamtac, I., Conti, M., Liu, J.J.-N.: Mobile ad hoc networking: imperatives and challenges. Ad Hoc Netw. **1**(1), 13–64 (2003)
17. Tanenbaum, A.S., Wetherall, D.J.: Computer Networks, 5th edn. Pearson (2011)
18. Macker, J.: Mobile Ad hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Considerations, IETF (1999)
19. Perkins, C.E., Bhagwat, P.: Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. ACM SIGCOMM Comput. Commun. Rev. **24**(4), 234–244 (1994)
20. Chiang, C.-C., Wu, H.-K., Liu, W., Gerla, M.: Routing in clustered multihop, mobile wireless networks with fading channel. In: Proceedings of IEEE SICON (1997)
21. Murthy, S., Garcia-Luna-Aceves, J.J.: An efficient routing protocol for wireless networks. Mobile Networks Appl. **1**(2), 183–197 (1996)
22. Liu, K., Deng, J., Varshney, P.K., Balakrishnan, K.: An acknowledgment-based approach for the detection of routing misbehavior in MANETs. IEEE Trans. Mob. Comput. **6**(5), 536–550 (2007)
23. Le Boudec, J.-Y., Vojnovic, M.: Perfect simulation and stationarity of a class of mobility models. In: Proceedings of IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies (2005)
24. Bonabeau, E., Dorigo, M., Theraulaz, G.: Swarm Intelligence: From Natural to Artificial Systems. Oxford university press, New York (1999)
25. Engelbrecht, A.P.: Computational Intelligence: An Introduction. Wiley, New York (2007)
26. Kennedy, J.F., Kennedy, J., Eberhart, R.C.: Swarm Intelligence. Morgan Kaufmann, San Mateo (2001)
27. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Trans. Evol. Comput. **1**(1), 53–66 (1997)

28. Cauvery, N., Viswanatha, K.: Enhanced ant colony based algorithm for routing in mobile ad hoc network. In: World Academy of Science, Engineering and Technology (2008)
29. Abdul-Rahman, A., Hailes, S.: A distributed trust model. In: 1997 ACM workshop on New security paradigms (1997)
30. Papadimitratos, P., Haas, Z.J.: Secure routing for mobile ad hoc networks. In: Proceedings of the SCS Commnication Networks and Distributed Systems Modeling and Simulation Conference (CNDS) (2002)
31. Marti, S., Giuli, T.J., Lai, K., Baker, M., et al.: Mitigating routing misbehavior in mobile ad hoc networks. In; The 6th Annual International Conference on Mobile Computing and Networking (2000)
32. Buchegger, S., Le Boudec, J.-Y.: Performance analysis of the CONFIDANT protocol. In; Proceedings of the 3rd ACM International Symposium on Mobile Ad hoc Networking and Computing (2002)
33. Michiardi, P., Molva:Core, R.: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In: Advanced Communications and Multimedia Security, Springer, pp. 107–121 (2002)
34. Pirzada, A.A., McDonald, C., Datta, A.: Dependable dynamic source routing without a trusted third party. In: Proceedings of the Twenty-Eighth Australasian Conference on Computer Science-Volume 38 (2007)

Chapter 7
An Overview of Self-Protection and Self-Healing in Wireless Sensor Networks

Tarek Gaber and Aboul Ella Hassanien

Abstract Wireless sensor networks (WSNs) are currently considered as one of the key technology enablers for the monitoring systems. Since WSNs are facing large amount of emerging network applications, techniques, protocols, and attacks, it is very recommended to minimize the intervention of human in the defense process which could be inefficient and error-prone. To build such realtime system using the WSN technologies while minimizing the human's intervention, it is required to provide autonomy, scalability, self-protection, and self-healing features in an WSN application. This chapter will give an overview of the autonomic computing paradigm which is the key concept of self-protection and self-healing. It then describes the need of these features for the WSN application. It also gives an overview of the self-protection and self-healing of WSNs.

7.1 Introduction

Sensors are considered as the fine link between the physical and the digital world. They capture and reveal real-world phenomena and then convert these phenomena into a digital form which can then be processed, stored, and acted upon. The sensors can be integrated into various devices and environments. This enables the sensors to provide a numerous societal benefit. They can be used to avoid infrastructure failures, increase productivity, preserve precious natural resources, enhance security

T. Gaber (✉)
Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt
e-mail: tmgaber@gmail.com

A. E. Hassanien
Faculty of Computers and Information Science, Cairo University, Cairo, Egypt

T. Gaber · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Cairo, Egypt

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_7, © Springer-Verlag Berlin Heidelberg 2014

(surveillance and broader), and support new applications, e.g. smart home technologies and context-aware systems [1].

A Wireless Sensor Network (WSN) is a network consisting of a collection of small devices with sensing, processing, and communication capabilities to monitor the real-world environment. These devices are low-power wireless devices connected through radio communication facilities. They can be deployed in a distributed environment sensing, monitoring or collecting data, processing and communicating the data and coordinating actions with other peers or nodes at a higher hierarchy in an infrastructural setting. As the devices of the WSNs are low cost, robust, low power, and reliable, these WSNs have a broad rang of applications, including utilities, military systems, transportation system automation, and patients in medical condition monitoring, and others [2, 3].

These applications are typically used either in hostile or highly dynamic conditions. Unlike conventional data networks, the human existence in the WSNs is very rare. Hence, the WSN of these applications should be tolerant to failures and loss of nodes connectivity. In other words, the sensor nodes must be provided with intelligent agent to recover from failures with less human intervention as possible. This could be achieved if the WSN is supported with the autonomic computing techniques [4]. The chapter will focus on two (self-protection and self-healing) properties of autonomic computing in the context of WSN.

The remaining of the chapter is organized as follows. Section 7.2 introduces a brief overview of the main architecture of WSNs while in Sect. 7.3 gives an overview of the autonomic computing which is the base of the self-protection and self-healing. Section 7.4 highlights the problem of the self-protection in WSN and its existed solutions while Sect. 7.6 introduces the problem of the self-healing and its importance in the WSN. Section 7.7 discuss the biologically-inspired architecture for Sensor NETworks (BiSNET) and its limitation. The chapter is concluded in Sect. 7.8.

7.2 Wireless Sensor Networks Architecture

A sensor network consists of a large number of tiny sensor nodes and a few powerful control nodes called base station. Sensor nodes are supplied with limited battery power, small memory size and limited computational ability. WSN architecture, as shown in Fig. 7.1, consists of four nodes which are described as follows [5]:

1. Sensor nodes: They are the filed devices that must be capable of routing packets, control the process and control of equipment. In this case, a router is another special device that does not responsible for sensing process sensor or control equipment. It only routes the collected data to the selected path.
2. Gateway or Access point: It enables communication between host application and field devices.

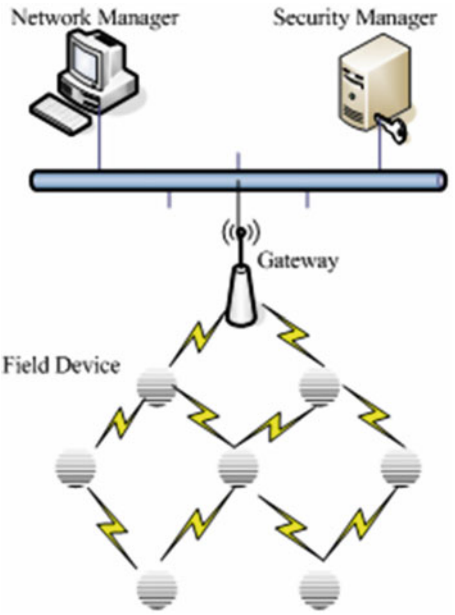


Fig. 7.1 The conceptual architecture of WSN [5]

3. Network manager: It is responsible for configuring network, scheduling communication between devices, management of routing tables, monitoring and reporting the health of the network.
4. Security manager: It is responsible for generation, storage and management of keys.

A WSN has the following characteristics [6]:

- Power consumption constrains for nodes using batteries or energy harvesting
- Ability to cope with node failures
- Mobility of nodes
- Dynamic network topology
- Communication failures
- Heterogeneity of nodes
- Scalability to large scale of deployment
- Ability to withstand harsh environmental conditions
- Easy of use
- Unattended operation.

7.3 Autonomic Computing

The main idea of autonomic computing is the concept of technologies managing technologies. The autonomic computing is originally inspired by the biological system, e.g. the autonomic human nervous system. It provides the ability of self-managing of computing systems/applications. This is emerged due to the fact that the complexity of computer systems including networking is increasing with the evolution. Hence, the systematic architecture of the software has changed, leading to gradual loss of the effective management of the system by the system manager. To solve this problem, autonomic computing lets the system manager to transfer the management issues to computer system which will automatically choose an appropriate solution if it meets a different problem. This solution does not require any involvement of the IT manager during the execution of the system operations [3, 7].

The autonomic computing systems enjoy a number of characteristics and consists of a number of elements. These are given in the next sections.

7.3.1 Autonomic Computing Characteristic

Autonomic systems or applications are able to manage their own behaviors and their relationships with other systems/ applications in accordance with high-level policies. Autonomic applications exhibit eight defining characteristics [8, 9]:

- Self Configuring: This means that an autonomic application has the ability of configuring and reconfiguring itself under variable and unpredictable conditions.
- Self-Awareness: This means that an autonomic application (a) can know itself and (b) is aware of its behaviors and its state.
- Self Optimizing: An autonomic system should be automatically able to properly achieve the configuration of its resources such that it gets high efficiency of the business or the customers needs.
- Self-Healing: It refers to the ability of the system to automatically detect and recover potential problems so that the application continues to function smoothly.
- Self Protecting: It is the feature of an autonomic system enabling it to identify, detect and protect its resources from potential internal and external attacks. This allows the system to maintain the its overall security and integrity.
- Context Aware: It refers to the ability of a system to be aware of it environment and be able to respond to any change in this environment.
- Open: An autonomic application should be able to function in an heterogeneous world. In addition, it must be portable across different hardware/ software architectures. In other words, it should not be a proprietary solution.
- Anticipatory: An autonomic application/system must have the ability of anticipating the optimized resources required to keep its complexity hidden from the users.

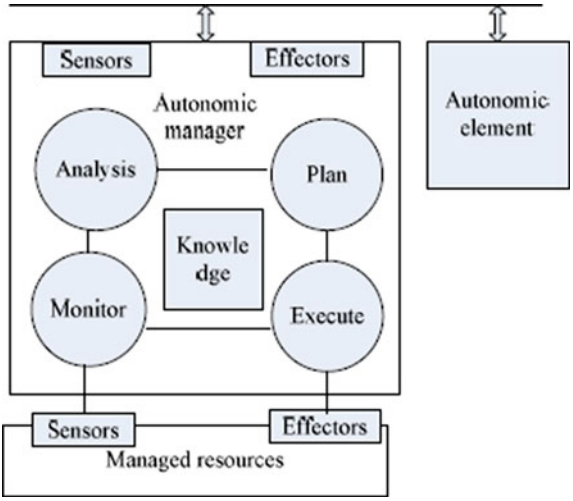


Fig. 7.2 The conceptual architecture of autonomic element [10]

7.3.2 *Autonomic Computing Elements*

As depicted in Fig. 7.2, an autonomic system consists of two major components: autonomic manager and managed resources. Sensors and effectors are two entities which are shared between the autonomic manager and the managed resources. The system makes use of the sensors to collect data relating to the state of a given element. This is achieved either polling-based or notification-based. On the other hand, the system uses the effectors to provide ways of changing the state of an element. Both of the sensors and effectors are collaboratively used to provide a manageability interface [2, 3, 7, 10]. The autonomic manager implements the control loop and consists of four parts, each of which shares a common knowledge base. To achieve the autonomic computing, the autonomic manager defines a control loop to perform functions concerning to Monitoring, Analyzing, Planning and Executing (MAPE) of processes [7, 8]. These function are defined below.

- Analyze: Show a relationship of collected data.
- Plan: Specifies the actions required to accomplish pre-defined goals.
- Monitor: Collect, aggregate, filter, and report about data from managed resources.
- Execute: Enable changes to be applied to managed resources in conjunction with a plan.

Autonomic managers continuously (a) monitor the managed resources and their environment and (b) deal with events on which actions could be executed upon. Sensors and effectors then provide reflections, manage, and control interfaces to the managed resources.

7.4 Self-Protection

Self-protection is considered as an essential characteristic of the autonomic computing. It is closely related to the other characteristics, see Sect. 7.3, e.g. self-configuration and self-optimization. From one side, the system integrity of a self-configuring and self-optimizing depends on self-protection functions which make sure that the system integrity is still in intact during any dynamic change. On the other side, the implementation of self-protection functions could use the same methods used for the reconfiguration and adaptation of the system.

As reported in [4], the self-protection can be seen from two perspectives: reactive and proactive. When it is seen as a reactive, the system should be automatically able to defend against cascading failures or malicious attacks. When it is taken as proactive, the system should be able to anticipates any security problem, may happen in the future, and takes actions to deal with it.

A self-protecting system should be able to (a) detect and identify hostile behaviour and then (b) take autonomous steps to protect itself from intrusive behavior. Self-protection mainly aims to defend system against malicious intentional actions. This is achieved though scanning untrusted activities and then reacting accordingly to these activities. This protection process is achieved without the user's awareness of it [11]. As discussed in [8], to build a self-protected, the following principles are required:

1. A self-protected system should be able to detect intrusions. Sure, such intrusions are different from system to another. It is based on the definition of each systems' operations.
2. A self-protected system must have the ability to take actions to attacks. This means that the system must be able to block the attack a whenever such attack is detected.
3. A self-protected system must protect the self-protection components themselves, i.e. prevent them from being compromised.

An autonomic system protects itself from malicious attacks but also from end users who inadvertently make software changes, e.g. by deleting an important file. The system autonomously tunes itself to achieve security, privacy and data protection. Thus, security is an important aspect of self-protection, not just in software, but also in hardware (e.g. TCPA [23]). A system may also be able to anticipate security breaches and prevent them from occurring in the first place. Self-protecting systems anticipate, detect, identify, and protect themselves from attacks from anywhere. Self-protecting systems must have the ability to define and manage user access to all computing resources within the enterprise, to protect against unauthorized resource access, to detect intrusions and report and prevent these activities as they occur, and to provide backup and recovery capabilities that are as secure as the original resource management systems. Systems will need to build on top of a number of core security technologies already available today, including LDAP (Lightweight Directory Access Protocol), Kerberos, hardware encryption, and SSL (Secure Socket

Layer). Capabilities must be provided to more easily understand and handle user identities in various contexts, removing the burden from administrators.

7.4.1 Self-Protection in WSN

WSNs have been widely used in monitoring the real-world environments such as object monitoring, path protection, or military applications. Such environments are under hostile conditions. This requires the providence of fault tolerance or a degree of protection for the senor network to be resistible to the attacks mounted from outsides. Since the sensors themselves are the most important part of the network, it is crucial to provide a kind of protection to them. It is believed that the sensors themselves are the best object to provide this protection. Wang et al. [12] were the first authors who have named this problem with self-protection problem in WSNs [12–14].

The self-protection in WSN concerns with the use of sensor nodes for providing protection to themselves, so that they can detect and thwart attacks directly targeting them. A WSN is said to be p-self-protected, if at any time, at least there are p active sensors which can monitor any sensor node (active or non-active) of this WSN. An active node is the one which has the ability to carry out protections; otherwise it is called a non-active sensor [12–14].

7.4.1.1 Problem of Self-Protection in WSN

As Wang reported in [12] the self-protection problem in WSN is defined as follows. The sensor network is viewed as a graph G(V;E), where V denotes to the set of wireless sensor nodes, both active and non-active, while E refers to the set of directed links on the graph, (u; v), where the nodes u; v ∈ V and v is in the sensing range of u. Given this view, the problem of self-protection in WSNs could be seen as the problem of finding the Minimum p-Self-Protection (MSP_p) which aims to find a subset of V to serve as active sensors in the sense that the WSN is p-self-protected and the number of active sensor nodes ($|MSP_p|$) is minimized. The following definitions are also needed while giving an overview of the Self-Protection in WSN.

- Independent set: A set of vertices is called an independent set if there is no edge between any pair of vertices.
- Maximum independent set (MIS) is independent set with the largest number of vertices.
- Dominating set : A subset S of $V$is called a dominating set if every node u in V is either in S or is adjacent to some node v in S. From this definition, it is clear that any MIS is a dominating set.
- Minimum dominating set (MDS): A dominating set having minimum cardinality is known as minimum dominating set (MDS).
- Connected dominating set (CDS): A subset K of V is a connected dominating set (CDS) if K is a dominating set and K induces a connected sub-graph.

7.4.1.2 Existed Solutions for Self-Protection in WSN

Although the self-protection is very important to provide autonomic structure in WSN, there are few chapters in the literature addressing this feature. Wang et al. [12] were the first researchers introducing a formal study on the self-protection problems in WSNs. They proved that the focus was only on improving the quality of field or object being covered, there is no need to provide self-protection to the sensors, thus makes the system extremely vulnerable. In other words, they argued that only focusing on the quality of area or object being covered does not certainly provide self-protection to the sensors. In [12] Wang et al. proved that the problem of finding 1-self-protection is NP-complete problem [15] by reducing the minimum set cover problem. Also, based on the approximation algorithm for minimum dominating set, they proposed two methods addressing 1-self-protection problem. The first method is called a centralized method which is much suitable for small-scale WSN. The second method is known as distributed method to be used in large-scale WSN.

In [13] Wang et al. have proposed an solution to the p-self-protection[1] problem in WSN. They firstly proved that the problem of finding the p-self-protection is NP-complete problem. They also introduced two efficient algorithms (centralized and distributed) with constant approximation ratio. In addition to achieving the p-self-protection, the distributed algorithm also maintains the connectivity of all active sensor nodes. Wangs algorithms are designed based on maximum independent set (MIS). The work in [12–14], introduced two main methods addressing the self-protection problem in WSN. These methods are highlighted below.

7.4.1.3 Centralized Method

Wang et al. [12] suggested a centralized method which depends on the approximation algorithm concept. They first proved that the cost of finding the minimum 1-self-protection is the double of the cost of finding the minimum dominating set. They then applied the $(1 + \log n)$ approximation algorithm [16] to find the minimum dominating set. This is achieved with the cost of $2(1 + logn)$ approximation.

Yu Wang et al. [13] have proved that the centralized method in [12] is very difficult to be extended for solving the p-self-protection problem. So, Yu has proposed another centralized method depending on constant approximation ratio. This method can solve 1-self-protection and p-self-protection problems. For 1-self-protection problem, maximal independent set MIS is first computed and then one neighbor for each node in MIS is chosen. The entire nodes in MIS along with the chosen neighbors are then set as active sensors. Where every node outside the set MIS is protected by its selected neighbor, then 1-self-protected is achieved [13, 14].

For *p-self-protection* problem, the algorithm starts by generating *k* MISs in *k* rounds. In each round, an MIS, Mk, is created based on ranks of nodes. The ranks

[1] A WSN is p-self-protected, if at any moment, for any wireless sensor, there are at least p active sensors that can monitor it [13].

are then updated to prevent the selected MISs used in the early rounds to be used again in future rounds of MISs. After creating k MISs, the entire nodes in these MISs are put in the active set M. For each node n contained in these MISs, if it is surrounded by less than p neighbors, a neighbor v is added into M. The node v is then used to protect the node u.

7.4.1.4 Distributed Method

Unlike Centralized Method, Distributed Method is useful for a lot of applications that there is no centralized Control Center and all sensor nodes are selforganized. Therefore, each one of sensor node has to make decisions based on little information. Thus, it is recommended to establish a simple distributed method to handle the self-protection problem for self-organized sensors.

At the beginning of the algorithm, each node is in undecided state and the entire nodes are in the first round. In this undecided state, each node has information about its neighbors. This information includes: ID of each node, protection level for each node showing that this node is covered by which sensors in MIS, node round-counter indicating that the node is in which round of MIS construction, and the node status providing that role (undecided, active, non-active) of the node.

Where each node u knows information about its neighbors, the node can know which round is currently performing. For example, suppose that a node u is in *Undecided state* and in round r. If all u's neighbors are indeed in round r and in the same round, its ID is the largest between all non-MIS nodes, the node will then become a node in M_r. The node then send protection message, *Protect(u, r)*, to its neighbor. In addition to its neighbors received, *Protect(u, r)*, the node enters a new round $r + 1$.

The node u can not make a decision whether it must be marked as active or nonactive until it and all its neighbors complete p rounds. At this step, the nodes existed in $U^p(i = 1)M_i$ will be marked as active whereas nodes with *Undecided* state will become nonactive. At the end of p rounds, if the nodes in MIS have less than p-protection, each of these nodes will randomly chose a nonactive node to use it for its protection, and send *Request-Protection* message to notify that node. Upon the reception of this *Request-Protection* message, this node will become active and then inform its neighbors.

7.5 Self-Healing

Basically, self-healing is one of properties defining the autonomic system introduced by IBM [4]. However, the literature contains different aspects of Self-healing. According to [2, 17], self-healing is defined as a self-healing system should recover from the abnormal (or unhealthy) state and return to the normative (healthy) state, and function as it was prior to disruption. In [18], Ganek et al. have defined a self-healing application as “an organized process of detecting and isolating a faulty component,

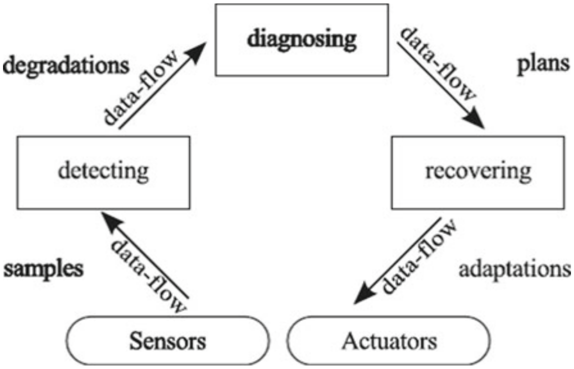


Fig. 7.3 Schematic view of self-healing process stages [19]

taking it off line, fixing the failed component, and reintroducing the fixed or replacement component into the system without any apparent disruption". They also proved that a self-healing system should be able to predict conflicts and try to prevent possible failures. For more details about the other definitions of self-healing, see [19].

To achieve the self-healing aim, as depicted in Fig. 7.3, a loop among three stages *(detecting, diagnosing, recovery)* and the environmental interfaces should be executed. This loop works as follows:

- Detecting: With this stage, any suspicious information received from reports or samples is filtered to identify the malicious information or attacks.
- Diagnosing: Based on the identified attacks or threats, analysis of the root cause is conducted and then an appropriate recovery plan is prepared according to policies predefined.
- Recovery: The planned adaptations are then carefully applied to the system such that the system constraints are met and any unpredictable side effects are avoided.

Self- Healing States Authors in [9] reported that the robustness of the self-healing system should not rely on a single element of the system but the entire system must have the ability to recover from its failures. In other words, the failures of single element should leave only minor impact on the entire system. In some systems, a fine line could be detected to clearly separate between acceptable and an unacceptable state. A temporary transmission zone could be only observed between such states. For such systems, Ghosh et al. [17] defined a fuzzy transition zone known as *Degraded State* reflecting the fact that a number of system conditions making self-healing systems to drift in are still acceptable state but are close to failure. This is illustrated in Fig. 7.4.

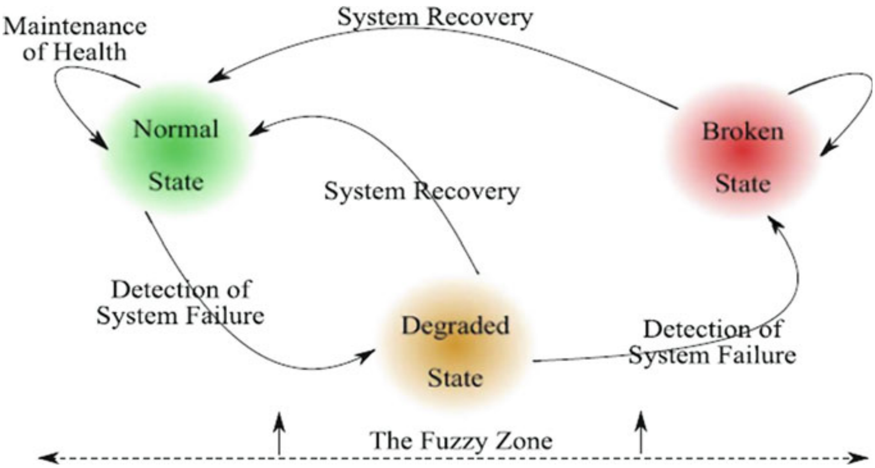


Fig. 7.4 Schematic view of self-healing states [19]

7.6 Self-Healing in WSNs

It is believed that WSNs should have a self-healing property. This is because the WSNs are subject to complex perturbations such as node join or leave, node movement or crash, and state corruption. So, the WSNs should have the ability of self-healing to repair their selves when they face a problem from one or more of their nodes which lose the connection with the rest of a network. This ability is very critical to ensure the coverage of network and continued functionality network after intentional (e.g. node malfunction) or unintentional (e.g. battery drainage) disconnecting of one or multiple nodes from the network. In addition, as the WSNs could be a large network, the self-healing is very essential for the availability, stability, and scalability of this network [20].

Self-healing has been studied in relation to personal computing [21]. Nonetheless, it is very difficult to directly apply the suggested solutions to the WSNs context. This is because these solutions require too much weight from the sensor nodes with limited resources. The literature contains a number of efforts [22, 23], on self-healing in WSNs. Some were targeting the architectural design of self-healing, others explored self-healing for a particular system or protocol. Next, an overview of both proposals of [22, 23] will be highlighted.

7.6.1 Existing Solution of WSN with Self-Healing

SASHA [22] is a self-healing mechanism for WSNs. Its design is inspired by the immunology mechanisms. It identifies the fault sensor nodes for base stations of the

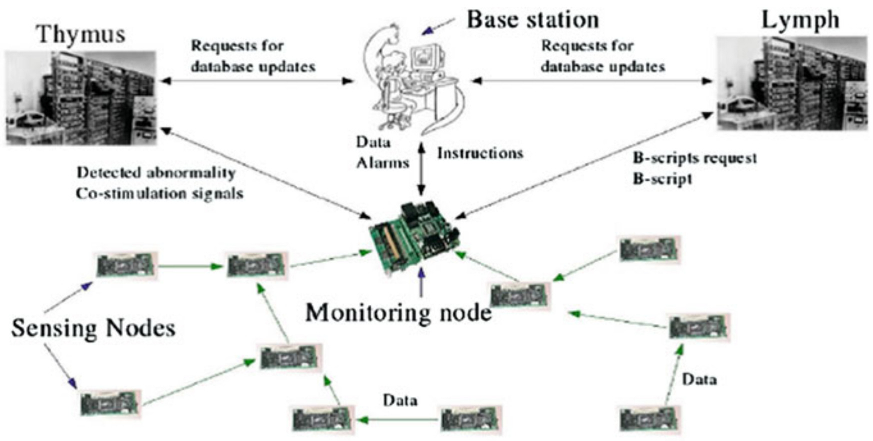


Fig. 7.5 Schematic view SASHA architecture [22]

WSNs. The fault nodes are detected by having the base station to compare collected data from multiple sensor nodes. The authors of SASHA, have identified three feature which should be in a self-healing WSN:

1. correct sensor readings,
2. appropriate behavior of a running application event, and
3. authenticated set of nodes.

They then proposed an architecture to achieve these features. The architecture and the functions of this system are described below.

7.6.1.1 SASHA Architecture

As shown in Fig. 7.5, SASHA architecture consists of five components: *Sensing Nodes, Monitoring Nodes, Base Stations, Thymus, and Lymph (database) machines.*

Sensing nodes: These are sensor nodes which are small, resource constrained, e.g. Mica mote.[2] The sensing nodes performs the following functions:

1. Sensing and transmitting measurements under study to the closest monitoring nodes.
2. Authenticating its neighbors and packets received.
3. Learning what establishes the self-set in terms of sensor readings.
4. Maintaining connectivity to monitoring nodes.
5. Responding to the commands received from the monitoring nodes.

[2] It is a node in a WSN that is capable of performing some processing, gathering sensory information and communicating with other connected nodes in the network.

Monitoring Nodes: They contain enhanced capabilities for communication, sensing, and processing, e.g. the Stargate. These nodes have two channels of communication: one for the communication among themselves and the other for the communication with sensing nodes. Every monitoring node has a portion of the network to cover. The sensing node performs the following tasks:

1. Authenticating (a) the nodes in its topology (usually the tree topology), and (b) its neighbors of the monitoring nodes.
2. Monitoring (a) the level of the noise behavior comes from its tree, (b) the data periodicity, and (b) the set of nodes.
3. Reviewing the sensor readings in its topology and validates their correctness.
4. Informing others nodes with an appropriate action responding to an identified attack.
5. Forwarding received data and notifying attack to a designated base station.
6. Query on an appropriate action to be taken from other monitoring nodes or base stations, in case of identified attacks or discovered anomalies.

Lymph: This is inspired by the natural immune system where B cells,a form of white blood cell, are programmed to search for specific types of disease-causing pathogens to destroy both the infected cells and the pathogens. In a sensor network, the B-cells could be achieved to detect anomalies by mobile scripts executing on all monitors. In designing a WSN for fault tolerant temperature data collection, the Lymph machine, as an example, should generate and send a monitoring script to the monitoring node to detect abnormal sensor readings. So, the Lymph machine contains a database storing signatures of identified attacks and other threats, as well as possible solutions to them.

Thymus: This component is designed to mimic the Thymus gland of the immune system. In this gland, T-cells go through a maturation process before releasing the circulation. With this process, the T-cells can develop a crucial attribute called self-tolerance. In the WSN, Thymus machine is designed to do the following tasks:

1. Storing the representation of a self-set. In case of detecting any abnormality, the representation of this abnormality will be transmitted to the Thymus machine.
2. Respond with co-stimulation signals to the monitor node to confirm the presence of faults and take a proper actions.

Base Station: This is designed to collect sensor data and provide solutions to the detected attacks at monitoring nodes.

7.6.1.2 SASHA Functions

SASHA is designed to achieve three objectives:

1. **Automatic fault recognition**: SASHA should be automatically and efficiently detect sensor faults. To accomplish this function, SASHA is embedded with a lightweight and distributed learning algorithm to identify faulty sensor readings, and other forms of anomalies.

2. **Adaptive network monitoring**: SASHA has realized this objective by developing a system to distribute mobile scripts to the WSN. This system is able to generate, maturate and migrate mobile monitoring scripts. This is achievable by using genetic algorithms which could generate continuous and changeable set of scripts based on the Lymph database including the representation of a non-self set.
3. **Coordinated Response**: SAHSA is able to respond to a malicious behavior. This is done by building a coordination between monitoring nodes themselves or Thymus, Lymph, and Base station.

SASHA Limitation This the self-healing protocol proposed in SASHA system only considers faulty sensor readings and does not look into approach of reducing energy conservation by means of self-healing. As reported in [24, 25], SASHA system gave the full attention to faulty sensor readings and overlooked the energy consumption problem while addressing the self-healing problem.

7.7 BiSNET: Biologically-Inspired Architecture for Sensor NETworks

The idea of BiSNET(Biologically-inspired architecture for Sensor NETworks) [23, 26] is inspired by the bees behaviors and characteristics. It is noticed that bees act autonomously to its local environmental conditions and with other bees. All activities of a bee colony are done without centralized control. This makes the colony able to scale to a huge number of bees. In addition, a bee colony can adjust its feature to dynamic environmental conditions. For example, in a hive, when it is noticed that the amount of honey is below a specific level, a number of bees go out to bring nectars from flowers. When it is noticed that the hive is filled with the honey, bees have a rest or expand the hive. Another feature of bee colony is that, when a group of bees lose their pheromone traces to flowers, they can recover or self-heal this problem. It can be seen that through shared behaviors and interactions among bees, a group of bees can autonomously develop a number of desirable system characteristics, e.g. adaptability and self-healing. Based on this remark, Boonma et al. have proposed a framework called BiSNET (Biologically-inspired architecture for Sensor NETworks). This framework employs key biological mechanisms to design WSN applications.

7.7.1 Overview of BiSNET

The BiSNET framework, shown in Fig. 7.6, composes of two software modules: *agents and middleware platforms*. These modules are modeled such as bees and flowers in the bees colony, respectively. Using BiSNET, a WSN application will be

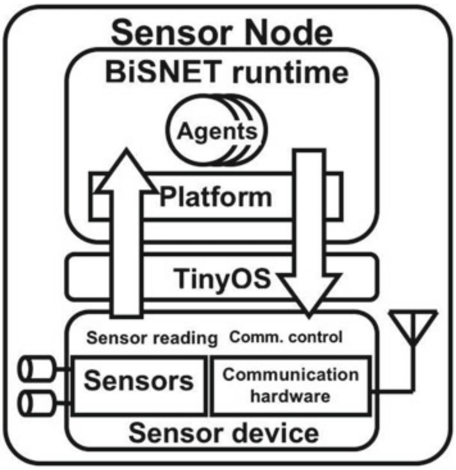


Fig. 7.6 Schematic view of the architecture of the BiSNET runtime [23]

designed as a distributed collection of multiple agents. As an example for application in WSN, BiSNET comprises multiple bees (agents in WSN) similar to a bee colony. An agent is designed to follow autonomy, decentralization, and food gathering or consumption as key biological principles. Like bees collecting nectars on flowers, the agents collect sensor data on platforms and carry the data to base stations. These data are then modeled as nests for bees. Agents accomplish these functionalities through autonomously invoking biological behaviors such as pheromone emission, replication, reproduction and migration. In each sensor node, a middleware platform runs atop of TinyOS and hosts one or more agents. It controls the state of the underlying node such as sleep and listen states. It also gathers required information for node localization and transmits it to a base station. In addition, it offers runtime services by agents to make their functionalities and behaviors.

7.7.2 Self-Healing in BiSNET

The self-healing capability of BiSNET is provided through two capabilities: *intra-node and inter-node self-healing*. The *Intra-node* self-healing permits agents to sense local node malfunctions and avoid false positive data to be transmitted to neighboring nodes. For example, reading sensor may swing between very low and very high values on a malfunctioning node. The *Inter-node* self-healing allows agents to distinguish malfunctions in neighboring nodes and avoid false positive data from the node. Based on the simulation results conducted in [23], using the self-healing of both *intra-node and inter-node*, BiSNET enables sensor nodes/gents to autonomously detect and eliminate (self-heal) false positive data and avoid wasting power. In other words, the two self-healing capabilities (intra-node and inter-node) can significantly save power consumption by reducing unnecessary data transmissions.

7.7.2.1 BiSNET Limitation

As reported in [24], self-healing property of BiSNET is designed based on the biologically architecture of bees; however, the self-heal ability of the BiSENT is not well illustrated in general. It is only considered the power consumption while addressing the self-healing. It has overlooked the relation between self-healing and performance of the system. It has also focused on solution of detection and elimination of false positive data and it does not considered solutions for detecting false negative data.

7.7.2.2 Other Self-Healing Solution

The literature also contains a number of papers [27–31] that proposed solutions for self-healing in WSN environment. The solutions in [27, 28] attempt to address the self-healing in terms of bridging holes in routing or node coverage. These solution addressed node mobility which could be anticipated by the authors. Nonetheless, this proposal is still a luxury for current sensor network practices in general.

Also, in [30], based on infusion of secure randomness, Di et al. have proposed a self-healing protocol allowing sensors to collectively recover from compromises. Nevertheless, they only give attention to backward secrecy, i.e. recovering from prior compromise. Furthermore, in [28, 31], a sefl-healing processes are suggested for a ZigBee-based sensor network. This process enables the network to repair itself in case of node failure or communication breakdown. This is achieved thought permitting a disconnected subnet to rejoin the network. The two proposals [28, 31] have added an improvement to the standard of ZigBee self-healing scheme described in specification [26]. However, solving the network connectivity problem was the main focus of these two proposals.

7.8 Conclusions

A self-protecting system is defined as a system which is be able to detect and identify hostile behaviour and then take autonomous steps to protect itself from intrusive behavior. And a self-healing system is the one that can detect and isolate a faulty component, take it off line, fix the failed component, and reintroduce the fixed or replacement component into the system without any apparent disruption. WSNs are currently being used in many applications, e.g., utilities, military systems, transportation system automation, and patients in medical condition monitoring. Depending on human interventions in the running of these applications could be inefficient and error-prone. So, it is highly recommended to design WSN technologies for these applications such that they provide the self-protection and self-healing properties. This chapter has given an overview of the autonomic computing paradigm which is the key concept of these two properties. It also has discussed the need of these properties in the WSN application. In addition, it has presented an overview of the existed solutions proposed to achieve the self-protection and self-healing in WSNs environment. It finally highlights a number of open issues about these two issues in the WSN environment. As pointed out by Boonma in [23], in order to effectively direct agents to base stations of the BiSNET, other biological-inspired routing protocols could be investigated.

References

1. Mukhopadhyay, S.: Wireless sensors and sensors network. In: Intelligent Sensing, Instrumentation and Measurements of Smart Sensors, Measurement and Instrumentation. vol. 5, pp. 55–69, Springer, Berlin (2013)
2. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Comput. Netw. **52**(12), 2292–2330 (2008)
3. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Comput. Netw. **38**(4), 393–422 (2002)
4. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer **36**(1), 41–50 (2003)
5. Kalita, H.K., Kar, A.: Wireless sensor network security analysis. Int. J. Next Gen. Netw. (IJNGN) **1**(1), 1–10 (2009)
6. Ponnusamy, V.A.A.: Wireless sensor network. Int. J. Comput. Eng. Sci. (IJCES) **2**(3), 55–61 (2012)
7. Liao, B.S., Li, S.J., Yao, Y., Gao, J.: Conceptual model and realization methods of autonomic computing. J. Softw. **19**(4), 779–802 (2008)
8. Parashar, M., Hariri, S.: Autonomic computing: An overview. In: Unconventional Programming Paradigms, pp. 257–269, Springer (2005)
9. White, S., Hanson, J., Whalley, I., Chess, D., Kephart, J.: An architectural approach to autonomic computing. In: Proceedings International Conference on Autonomic Computing, pp. 2–9 (2004)
10. Duan, F., Li, X., Liu, Y., Fang, Y.: Towards autonomic computing: a new self-management method. In: Artificial Intelligence and Computational Intelligence. pp. 292–299, Springer (2011)
11. Chopra, I., Singh, M.: Shape–An approach for self-healing and self-protection in complex distributed networks. J. Supercomput. **67**(2), 1–29 (2013)
12. Wang, D., Liu, J., et al.: Self-protection for wireless sensor networks. In: 26th IEEE International Conference on Distributed Computing Systems, 2006. ICDCS 2006, pp. 67–67 (2006)
13. Wang, Y., Li, X.Y., Zhang, Q.: Efficient self protection algorithms for static wireless sensor networks. In: Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE, pp. 931–935 (2007)
14. Wang, Y., Li, X.Y., Zhang, Q.: Efficient self protection algorithms for wireless sensor networks http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.9174. Accessed 27 May 2014
15. Garey, M.R., Johnson, D.S.: Computers and Intractability; A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York (1990)
16. Johnson, D.S.: Approximation algorithms for combinatorial problems. J. Comput. Syst. Sci. **9**(3), 256–278 (1974)
17. Ghosh, D., Sharman, R., Rao, H.R., Upadhyaya, S.: Self-healing systems survey and synthesis. Decis. Support Syst. Emerg. Econ. **42**(4), 2164–2185 (2007)
18. Ganek, A.G., Corbi, T.A.: The dawning of the autonomic computing era. IBM Syst. J. **42**(1), 5–18 (2003)
19. Psaier, H., Dustdar, S.: A survey on self-healing systems: approaches and systems. Computing **91**(1), 43–73 (2011)

20. Zhang, H., Arora, A.: Gs3: scalable self-configuration and self-healing in wireless sensor networks. Comput. Netw. **43**(4), 459–480 (2003). Wireless Sensor Networks
21. Sterritt, R., Bantz, D.: Personal autonomic computing reflex reactions and self-healing. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **36**(3), 304–314 (2006)
22. Bokareva, T., Bulusu, N., Jha, S.: Sasha: Toward a self-healing hybrid sensor network architecture. In: Proceedings of the 2Nd IEEE Workshop on Embedded Networked Sensors. EmNets '05, IEEE Computer Society, Washington, USA, pp. 71–78 (2005)
23. Boonma, P., Suzuki, J.: Bisnet: A biologically-inspired middleware architecture for self-managing wireless sensor networks. Comput. Netw. **51**(16), 4599–4616 (2007). (1) Innovations in Web Communications Infrastructure (2) Middleware Challenges for Next Generation Networks and Services
24. Li, J., Wu, Y., Stankovic, J., Son, S., Zhong, Z., He, T., Kim, B.W., Joo, S.S.: Predictive dependency constraint directed self-healing for wireless sensor networks. In: 2010 Seventh International Conference on Networked Sensing Systems (INSS), pp. 22–29 (2010)
25. Ponnusamy, V.A.A.: Energy efficient mobility in wireless sensor network. Int. J. Multimedia Image Process. **1**(1), 53–62 (2011)
26. Boonma, P., Champrasert, P., Suzuki, J.: Bisnet: A biologically-inspired architecture forwireless sensor networks. In: 2006 International Conference on Autonomic and Autonomous Systems, 2006. ICAS '06, pp. 54–54 (2006)
27. Vlajic, N., Moniz, N.: Self-healing wireless sensor networks: Results that may surprise. In: 2007 IEEE Globecom Workshops, pp. 1–6 (2007)
28. Qiu, W., Hao, P., Evans, R.: An efficient self-healing process for zigbee sensor networks. In: International Symposium on Communications and Information Technologies, 2007. ISCIT '07, pp. 1389–1394 (2007)
29. Du, X., Zhang, M., Nygard, K., Guizani, M., Chen, H.H.: Distributed decision making algorithm for self-healing sensor networks. In: IEEE International Conference on Communications, 2006. ICC '06, vol. 8, pp. 3402–3407 (2006)
30. Di Pietro, R., Ma, D., Soriente, C., Tsudik, G.: Posh: Proactive co-operative self-healing in unattended wireless sensor networks. In: IEEE Symposium on Reliable Distributed Systems, 2008. SRDS '08, pp. 185–194 (2008)
31. Wan, J., Chen, W., Xu, X., Fang, M.: An efficient self-healing scheme for wireless sensor networks. In: Second International Conference on Future Generation Communication and Networking, 2008. FGCN '08, vol. 1, pp. 98–101 (2008)

Chapter 8
Cybercrime Investigation Challenges: Middle East and North Africa

Mohamed Sarrab, Nasser Alalwan, Ahmed Alzahrani and Mahdi Kordestani

Abstract Cybercrime is the use of IT in any suspicious criminal activities.Currently, our life becomes increasingly depending on modern information technology; however, it has become very necessary to improve the cybercrime investigation techniques especially in when processing very secret and sensitive information such as government and military intelligence, financial or personal private information. Cybercrime investigation attempts to detect unauthorized access to information in digital source with the intent to steal, modify or destroy that digital information. Such suspicious activities can cause financial damages or government information disseminates; moreover, it might destroy military high secret and confidential information. Therefore, this chapter focuses mainly on highlighting the main challenges of Middle East's and North Africa's countries in cybercrime investigation system by considering the recent developments in the continents Internet infrastructure and the need of information security laws in these particular countries. This chapter mainly focuses on Internet infrastructure development in these particular to show how they might become vulnerable to cybercrime attacks.

M. Sarrab (✉) · M. Kordestani
Communication and Information Research Center, Sultan Qaboos University,
Muscat, Oman
e-mail: sarrab@squ.edu.om

M. Kordestani
e-mail: amiri@squ.edu.om

N. Alalwan · A. Alzahrani
Computer Science Department, Riyadh Community College, King Saud University,
Riyadh, Saudi Arabia
e-mail: nalalwan@ksu.edu.sa

A. Alzahrani
e-mail: ahmed@ksu.edu.sa

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_8, © Springer-Verlag Berlin Heidelberg 2014

8.1 Introduction

Due to the revolution of communication and information technology the use of internet has been enlarged which increases the rate of digital crime all over the world. Digital crime can be defined as crimes directed at digital devices or their application systems. It is very important to secure sensitive information that are processed over the web (Internet) such as government and military intelligence, banking information or personal private information. Nowadays, our life becomes increasingly depending on modern information technology; however, it becomes very important to improve the cybercrime investigation procedure especially in case of processing very important and sensitive information. Cybercrimes may include different suspicious activities such as machine unauthorized access, digital frauds, system interference as well as computer misuse which might not involve any type of machine physical damage. In fact, computer crime is not just unauthorized access to a computer system with intent to delete, modify or damage computer data and information but it's so far more complex. It can have the form of simple snooping into computer system without authorization. It can also be theft of money, data or secret information. These different and complex types of computer crime make the work of cybercrime investigator became very hard to trace detect and prevent any type of cybercrimes.

The main challenge in this study is that the Middle East's and North Africa's countries are in different stages or level of electronic management implementation such as E-commerce, E-government, E-business and E-services overall. The motivation of this research is raised from the rapid increase of the computer crimes. Some countries have a good progress in the investigation procedures as well as in the law of cybercrime. In some other countries the investigators have a good experience in different kinds of cybercrimes and they modified their law to protect the personal private and government secret information. In all of these countries, the law is enforced to be responsible for criminal penalties for any identified computer crimes. This may lead the other countries to be as a worth field of computer crime and attract criminal to steal, destroy or distribute any type of information. The purpose of this chapter is to discuss the main challenges of the Middle East and North Africa's countries in cybercrime investigation system, by considering the recent developments in the continents Internet infrastructure and the need of information security laws in these particular countries [1–3].

The reminder of the chapter is structured as follows. Sections 8.2 and 8.3 provide general overview about cybercrimes and introduce the cybercrime conception. Section 8.4 discusses the state of information security in Middle East and North Africa's countries. Section 8.5 presents the Internet users and population statistics for Middle East and 6 provides Internet users and population statistics for North Africa's countries. Section 8.7 presents case study includes four different examples Stuxnet, Win32.Disttrack, Start and Flame. Section 8.8 discusses the Internet users and population statistics for Middle East North Africa's countries. Section 8.9 highlights the challenges of cybercrime investigation system in these particular countries.

8.2 Related Work

The telecommunication and information technologies has increased the number of users as well as raising the number of cybercrimes. To secure data or information from cyber criminals, it is very important to have a database to prevent unauthorized access based on confidentiality [4]. Cybercrime and traditional crime investigation have similar investigation procedure including (collecting evidences, inspection and analyzing evidences). In addition to that, in both traditional and cybercrimes the investigators trying to answer [5].

- What was the type of the crime?
- When did it happen?
- Where did it happen?
- How did it happen?
- Who did it?
- Why did it happen?

Cybercrimes are dealing with specific areas such as computer devices, network, and storage devices and might include any other digital communication Medias. In cybercrimes investigation it's very important to have huge record of any available devices catalogs, manuals or any logging files which can help in tracing or can be used as evidence for detecting the computer crime perpetrator [6, 7]. The strategic plan is the most important step in cybercrimes which can be as a long term plan or map that is concerned with national data network infrastructure [8]. The investigation team is another important factor in discovering cybercrime. In fact it's very difficult to have one investigation team with different skills such as a good experience in information technology, network, computer machines and software tools. Moreover, the guide or team leader should be the most expert in forensic and cybercrimes investigation [4]. Digital forensics can be defined as the science of identifying, collecting, documenting, preserving, analyzing, examining and presenting evidence from computer devices, networks, and other electronic devices. In general, digital forensics classifies and deals with the digital evidence in such way that is officially acceptable by courts [9, 10]. The court accepted digital evidence is a necessary part of the computer crime investigation procedure where they might involve computer hardware, software, manuals, or phone numbers [11–13]. Despite a long history and many work has been done on the cybercrime investigation [14–16]; to the best of our knowledge no work has been done about cybercrime investigation in this particular area generally Middle East's and North Africa's countries.

8.3 Conceptualizing Cybercrime

Cybercrimes are any suspicious criminal activities which are committed against computer hardware machines or software tools. In cybercrime, the computer device is the target of any suspicious criminal activities. In fact, computer crime types are not only associated to the software, data, information or any other program applications or tools. The criminal actions in the context of computer is often refers to the computer functions; such as file transfer facilities, social media applications, audio or visual conferencing tools and electronic mailing etc. However, computer crimes are any suspicious criminal activities committed using computer and network (Internet) to violate the existing legislation laws or forensic roles. Cybercrime might also involve the use of digital resources to commit any type of normal crimes such as theft of identifiable card information and other forms of proprietary information or property in both digital and physical form [15, 17]. The following are types of cyber attacks:

- Viruses and worms are programs that affect the storage a computer or network devices, which may destroy, steal or replicate information.
- Spam emails are junk postings which sent without the consent of the receiver.
- Malware is a software that have the power to control any individuals computer to spread a bug to other peoples devices.
- Trojan is a computer program that appears legitimate. But, once it runs, locates password information or makes the system more vulnerable to future entry. Trojan may destroy programs or data on the storage device.
- Scareware is tactics that used to compel users to download certain program. While such program is usually presented as antivirus software, after while these programs start attacking the users system. The user then has to pay the criminals to remove such viruses.
- Denial-of-service (DoS) occurs when criminals attempt to bring down or cripple individual websites, computers or networks, often by flooding them with messages.
- State cyber attacks is the use of cyber attacks as a new means of warfare between governments.
- Fiscal fraud targeting official online payment channels, cyber attackers can hamper processes such as tax collection.
- Carders Stealing bank or credit card details to withdraw cash at ATMs or in shops.
- Phishing attacks are designed to steal a persons login and password.

8.4 State of Information Security in North Africa and Middle East

The state of cybercrime in Middle East and North Africa is different from other countries, where the state of information technology security in Middle East and North Africa regions are affected by many factors such as IT infrastructure, growth of IT user and lack of regulation and training of law enforcements.

Table 8.1 World Internet Usage and Population Statistics June 30, 2012

Region	Population	Users 2000	Users	Growth (%)
Africa	1,073,380,925	4,514,400	167,335,676	3,606.7
Asia	3,922,066,987	114,304,000	1,076,681,059	841.9
Europe	820,918,446	105,096,093	518,512,109	393.4
Mid East	223,608,203	3,284,800	90,000,455	2,639.9
N America	348,280,154	108,096,800	273,785,413	153.3
S America	593,688,638	18,068,919	254,915,745	1,310.8
Oceania	35,903,569	7,620,480	24,287,919	218.7
Total	7,017,846,922	360,985,492	2,405,518,376	566.4

8.4.1 Information Technology Infrastructure

In the past, it was rare to have Internet connectivity in Africa [18] and Asia. However, recently Middle East and North Africas countries show signs of becoming a major player in the information and communication technology (ICT) arena. United Nation members, including Asia's and Africa's countries, have agreement to reach some common objectives by year 2015, involving the development of global partnerships [19]. This goal concentrates on the cooperation with the private sector to maintain the benefits of new technologies, especially information and communication technologies [20]. The Middle East and North African governments have agreed to cooperate with private sector companies to provide information and communication technology (ICT) services to all of their citizens. Different foreign organizations and companies have already started investing in these particular areas, helping and supporting these regions in developing their infrastructure. In 2007, SEACOM built the Africas first undersea fiber-optic cable to connect Africa's eastern and southern countries with the rest of the world [21]. Africa and specially the North Africas countries are now well equipped with Internet high speed, giving the local Internet Server Provides the ability to offer faster and cheaper Internet access types to the customers [22].

8.4.2 Growth of IT User

With currently cheap network broadband connections the number of IT users is growing every day. According to Internet World statistics, Asias Internet penetration rate as of June 2012 was 841.9 %. Where, this percentage refers to the number of users divided by the population. In 2012, Asias population nearly 4 billion [23] as indicated in Table 8.1 the world internet usage and population statistics in June 2012. The number of users in Asia has been changed from 114,304,000 in 2000 to 1,076,681,059 in 2012 has made many new online business opportunities. In Africas Internet penetration rate as of June 2012 was 3,606.7 %. In 2009, Africas population reached 1 billion but on June 2012 was increased to 1,073,380,925 [23, 24] as indicated in Table 8.1. In terms of being connected with the rest of the world, only a few Africa's countries can be considered as emerging or developed country. Many Africa's countries have yet to garner high-enough penetration rates, due to the fact that most of them are not politically stable or in case of North Africa's countries have many work to do in terms of infrastructure. This penetration growth in Middle East and North Africa has increase the potential for misuse of IT. Because of IT and information security policy enforcement many Internet and IT users have become victims of computer crime attacks [1].

8.4.3 Lack of Regulations and Training of Law Enforcements

Most North Africas and some Middle East countries suffer from the lack of IT and information security awareness and different kind of security policy enforcement among IT users. The security awareness and security policy enforcement in these countries are far behind the main players of the dangerous game Europe, Russia, China and USA. Because of that North Africas and some Middle East countries spend less effort to raise awareness and security policy enforcement among their IT users. Generally, these countries are lack of training and regulations of law enforcements. These countries need strong security awareness training, targeting native speakers to educate users, employees and law enforcers to understand the dangers and risks of attacks and hackers [25]. Moreover, the computer criminal activities in North Africa are not well reported. Hacktivist attacks and scams are very common in North Africa's countries, following the unknown attacks defacing different sites for political reasons. For instance, an Algerian hacking attack that defaced several Romanian sites, including PayPal and Google [23].

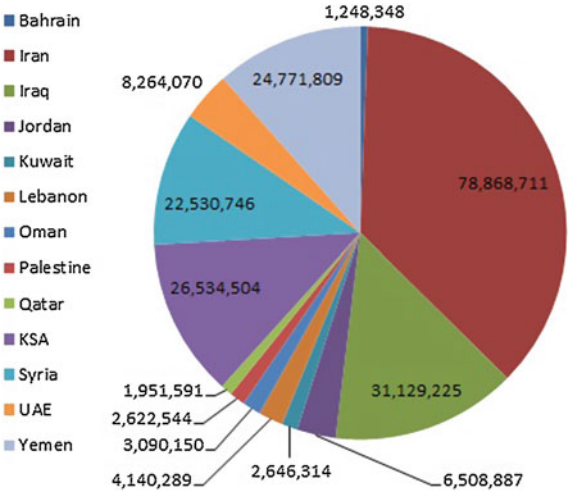
8.5 Internet Users and Population Statistics for Middle East

With currently high speed Internet connection and cheaper cost of internet services the number of internet users will be increased daily. However, that the Internet and IT security context is not known to the vast majority of Middle east users [2]. Table 8.2 indicates the Internet users and population in Middle East (Bahrain, Iran, Iraq, Jordan, Kuwait, Lebanon, Oman, Palestine, Qatar, KSA, Syria, UAE and Yemen). As shown in Fig. 8.1 that Iran has the highest population with 78,868,711 but it was the third Middle East country in the number of Internet users by December 2000 after UAE with 735,000 users and Lebanon with 300,000 users. Iraq has the second highest population with 31,129,225 but it has the lowest number of Internet users by December 2000. Iraq was moved to be the fourth middle East country with 2,211,860 users in June 2012 indicated in Fig. 8.2. As shown in Fig. 8.2 that Qatar is the second smallest countries in Middle East with population of 1,951,591 and its internet users are moved from 30,000 to be 1,682,271 users in June 2012 with the highest penetration

Table 8.2 Internet users and population in Middle East countries

Country	Population	Users 2000	Users 2012	Penetration (%)
Bahrain	1,248,348	40,000	961,228	77.0
Iran	78,868,711	250,000	42,000,000	53.3
Iraq	31,129,225	12,500	2,211,860	7.1
Jordan	6,508,887	127,300	2,481,940	38.1
Kuwait	2,646,314	150,000	1,963,565	74.2
Lebanon	4,140,289	300,000	2,152,950	52.0
Oman	3,090,150	90,000	2,101,302	68.8
Palestine	2,622,544	35,000	1,512,273	57.7
Qatar	1,951,591	30,000	1,682,271	86.2
KSA	26,534,504	200,000	13,000,000	49.0
Syria	22,530,746	30,000	5,069,418	22.5
UAE	8,264,070	735,000	5,859,118	70.9
Yemen	24,771,809	15,000	3,691,000	14.9
Total	214,307,188	2,014,800	84,686,925	93.5

Fig. 8.1 The population of Middle East countries



population 86.2 %, whereas sultanate of Oman is the fifth in the users penetration population with 68.8 %. The population of UAE is eight times less than Iran but their Internet users in December 2000 were 735,000 users as indicated in Fig. 8.2 however, by June 2012 UAE reached nearly 6 Million users being as the fourth penetration population 70.9 in Fig. 8.4. Bahrain is the smallest among the Middle East countries with population 1,248,348 and 961,228 Internet users in June 2012 as indicated in Fig. 8.3. However, Bahrain has the second biggest Internet penetration population 77 % just after Qatar with 86.2 % and before Kuwait 74.2 %. Table 8.1 indicated that Middle East had 90,000,455 Internet users in June 2012. Whereas, Iran and Saudi Arabia with the biggest Internet user bases. The exponential growth of Middle East

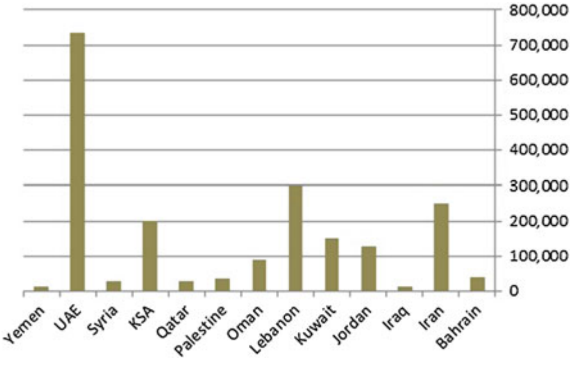


Fig. 8.2 The Middle East users in December 2000

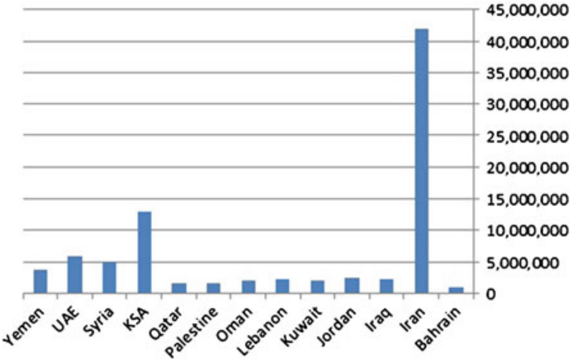


Fig. 8.3 The Middle East internet users in June 2012

user's base will force Internet provider to provide better services with cheaper cost, which might become potential hacking area. Because this growth of Internet users might benefits both users and attackers. The estimated number of Internet users in Middle East as of June 2012 was 84,686,925 as shown in Table 8.2. In December 2000, the Internet users were only 2,014,800 and by June 2012 become 84,686,925 users As can be seen from Fig. 8.4 that Iraq which is the second biggest country in Middle East has the smallest penetration percentage 7.1 %. However, nine from the thirteenth Middle East countries in Table 8.2 have penetration percentage over 49 % and the other four countries less than 49 % such as Yemen, Syria, Jordan and Iraq.

8.6 Internet Users and Population Statistics for North Africa

With nowadays cheaper and faster Internet access, more North Africans go Internet online or continually connected which increases the number of new users. But IT and internet security is a concept that it's not known to the vast majority of North African users [2]. Table 8.2 indicates the Internet users and population in North Africa (Egypt, Algeria, Morocco, Tunisia, Libya, and Mauritania).

Fig. 8.4 The Middle East countries internet penetration percentage

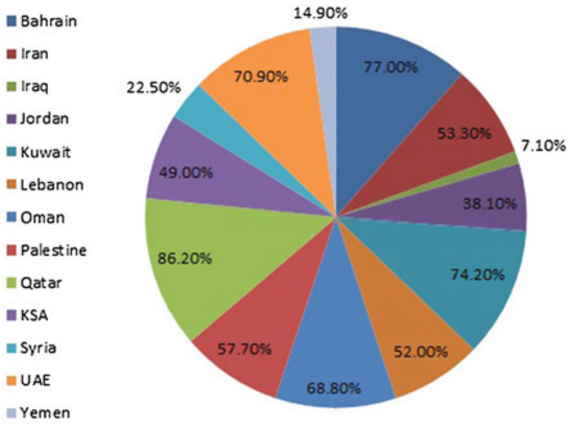
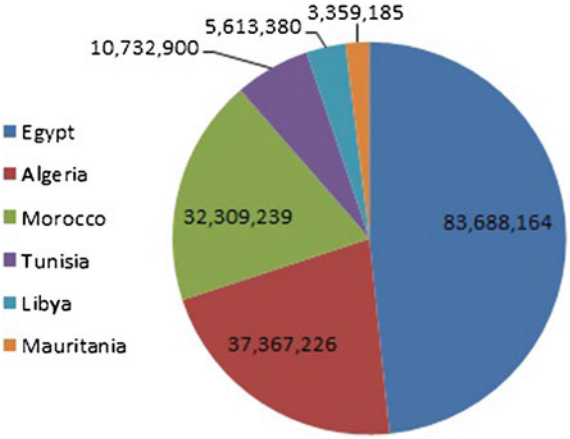


Fig. 8.5 The population of North Africa's countries



As shown in Fig. 8.5 Egypt has the highest population with 83,688,164 and it is the first North Africa's country in the number of Internet users by December 2000. Algeria has the second highest population with 37,367,226 but it was the fourth country in the number of Internet users by December 2000. Algeria was moved to be the third North Africa's country with 5,230,000 users in June 2012.

As shown in Fig. 8.5 Morocco is the third biggest North Africa's country with population of 32,309,239 and its Internet users are moved from 100,000 to be more than 16 million users with the highest penetration in using Internet 51 %.

The population of Tunisia is three times less than Morocco but their Internet users in December 2000 were 100,000 users as indicated in Fig. 8.6 and by June 2012 reached more than 4 Million users scoring the second penetration 39.1 in North Africa's countries. Mauritania has the smallest among the North Africa's countries with population 3,359,185 and 151,163 Internet users in June 2012 as indicated in Fig. 8.5.

Fig. 8.6 The North African internet users in December 2000

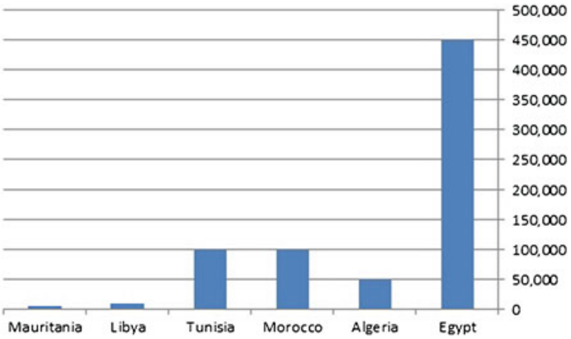


Fig. 8.7 The North African internet users in June 2012

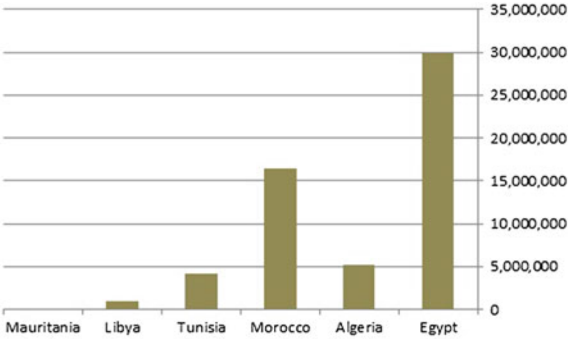


Fig. 8.8 The North Africa's countries internet penetration percentage

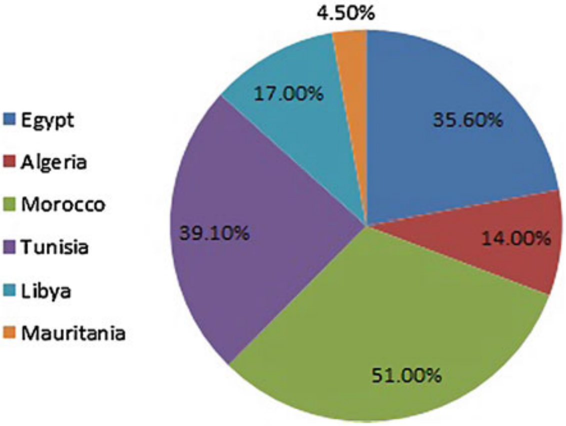


Table 8.1 indicated that Africa had 167,335,676 Internet users in June 2012. Whereas, Egypt, Algeria, Morocco and Tunisia with the ten Africa's countries with the biggest Internet user bases. The exponential growth of North Africa's user's base force Internet provider to reduce their service cost, which will benefit both users and attackers (Fig. 8.7).

Table 8.3 Internet users and population in North Africa's countries

Country	Population	Users 2000	Users 2012	Penetration (%)
Egypt	83,688,164	450,000	29,809,724	35.6
Algeria	37,367,226	50,000	5,230,000	14.0
Morocco	32,309,239	100,000	16,477,712	51.0
Tunisia	10,732,900	100,000	4,196,564	39.1
Libya	5,613,380	10,000	954,275	17.0
Mauritania	3,359,185	5,000	151,163	4.5
Total	173,070,094	715,000	56,819,438	32.8

The estimated number of Internet users in North Africa as of June 2012 was 56,819,438 as shown in Table 8.3. In December 2000, the Internet users were only 715,000. As can be seen from Fig. 8.8 that Mauritania has the smallest in the Internet users penetration percentage 4.5 % and it is the smallest country among the North Africa's countries. However, Algeria is the second biggest country in North Africa and it has the second smallest penetration percentage 14 % in Internet users.

8.7 Case Study

8.7.1 Stuxnet

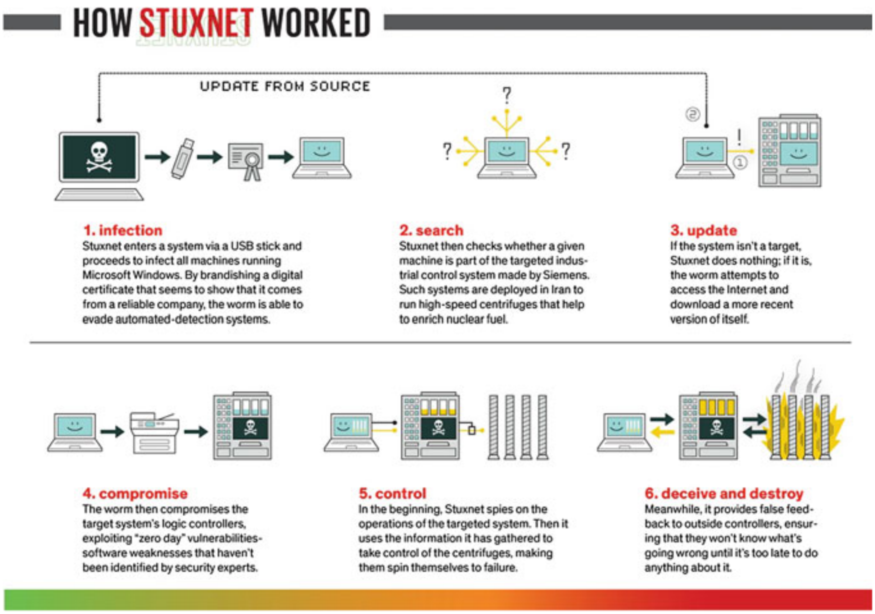
Stuxnet is described as a very complex computer worm which was first discovered around June 2010. Symantec has reported early Stuxnet variant first went live in 2005 with several versions of this software. Another report from Symantec indicated that Stuxnet is a threat that was primarily written to target an industrial control system or set of similar systems. Industrial control systems are used in gas pipelines and power plants. Its final goal is to reprogram industrial control systems (ICS) by modifying code on programmable logic controllers (PLCs) to make them work in a manner the attacker intended and to hide those changes from the operator of the equipment. In order to achieve this goal the creators amassed a vast array of components to increase their chances of success. This includes zero-day exploits, a Windows rootkit, the first ever PLC rootkit, anti virus evasion techniques, complex process injection and hooking code, network infection routines, peer-to-peer updates, and a command and control interface. We take a look at each of the different components of Stuxnet to understand how the threat works in detail while keeping in mind that the ultimate goal of the threat is the most interesting and relevant part of the threat [26, 27]. The Real Story of Stuxnet is an article published on February 2013, described Stuxnet as This worm was an unprecedentedly masterful and malicious piece of code that attacked in three phases. First, it targeted Microsoft Windows machines and networks, repeatedly replicating it, then it sought out Siemens Step7 software, which is also Windows-based and used to program industrial control systems that operate equipment, such as centrifuges. Finally, it compromised the programmable logic controllers. The worms authors could thus spy on the industrial systems and even cause the fast-spinning centrifuges to tear themselves apart, unbeknownst to the human operators at the plant. Stuxnet affected at least 14 industrial sites in Iran, including a uranium-enrichment plant [28, 29].

Symantec says Stuxnet was the first worm to exploit the Microsoft Windows Shortcut ‘LNK/PIF’ Files Automatic File Execution Vulnerability (BID 41732) in order to spread; in fact when Stuxnet was first discovered, this vulnerability was an unknown, or zero-day, vulnerability and it was not until Stuxnet was analyzed that we discovered this vulnerability. Normally when one thinks of a vulnerability in software, one would think of a coding error that an attacker discovers and then exploits. However, while this does indeed fit the definition of a vulnerability, specifically it is a design flaw as Windows is doing exactly what it was designed to do. and also reports other than Iran, countries like Indonesia, India, Azerbaijan, Pakistan and other are also infected [30]. The Register reported NSA whistleblower Edward Snowden has confirmed that the Stuxnet malware used to attack Iranian nuclear facilities was created as part of a joint operation between the Israelis and the NSA’s Foreign Affairs Directorate (FAD). [31]. Stuxnet is the first weapon completely made from the source code against another country. The Washington Times reported The 2009 cyber attack by the U.S. and Israel that crippled Irans nuclear program by sabotaging industrial equipment constituted ‘an act of force’ and was likely illegal under international law, according to a manual commissioned by NATOs cyber defense center in Estonia [32].

8.7.2 W32.Disttrack

The case study is about Saudi Arabian Oil Company (Saudi Aramco) where it confirmed the attack of its network occurred due to virus infection. Saudi Aramco is one of the largest energy and petroleum companies all over the world. This virus attack could lead information stealing, destroy or any other financial damage. In that time, Symantec announced the discovery of a new malware called W32.Disttrack or Shamoon (Figs. 8.9, 8.10, 8.11 and 8.12).

The malware infects a PC, steals certain data, send the data to another infected PC inside the compromised network and then overwrites the PCs Master Boot Record, which makes the system useless. The way this malware works might be linked to the Wiper malware which infected Iranian oil terminals in April 2012. The Wiper malware is also considered new variant of Flame as the investigation of the Wiper led to the discovery of Flame, according to Kaspersky Lab. Kaspersky also provides new analysis of how Shamoon is coded. This type of malware might be used to physically access to a computer device that is connected to Saudi Aramco network then data and information propagation started. The infected device might not be inside Aramco but it can be connected with the company remotely from any other place.



1. infection
Stuxnet enters a system via a USB stick and proceeds to infect all machines running Microsoft Windows. By brandishing a digital certificate that seems to show that it comes from a reliable company, the worm is able to evade automated-detection systems.

2. search
Stuxnet then checks whether a given machine is part of the targeted industrial control system made by Siemens. Such systems are deployed in Iran to run high-speed centrifuges that help to enrich nuclear fuel.

3. update
If the system isn't a target, Stuxnet does nothing; if it is, the worm attempts to access the Internet and download a more recent version of itself.

4. compromise
The worm then compromises the target system's logic controllers, exploiting "zero day" vulnerabilities-software weaknesses that haven't been identified by security experts.

5. control
In the beginning, Stuxnet spies on the operations of the targeted system. Then it uses the information it has gathered to take control of the centrifuges, making them spin themselves to failure.

6. deceive and destroy
Meanwhile, it provides false feedback to outside controllers, ensuring that they won't know what's going wrong until it's too late to do anything about it.

Fig. 8.9 The real story of Stuxnet-IEEE Spectrum: how Stuxnet works

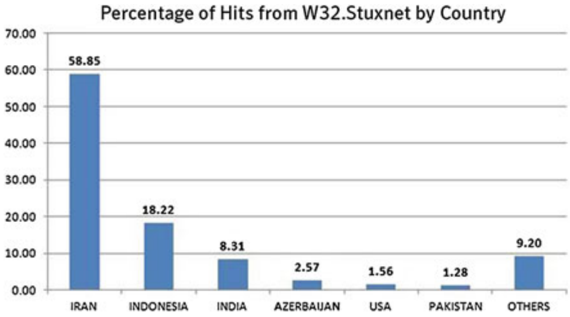


Fig. 8.10 Symantec report about Stuxnet spread

In this situation, Aramco needs to conduct thorough investigation to figure out from where this malware accessed their network not only focusing on the recovery from that attack. To identify the identity of the attacker a lot of work and collaboration needs to be done between GCC together and with other countries over the world specially the major players of the dangerous game such as USA, Europe, China and Russia [16, 18]. However, the import question after this Aramco attack is: Are Middle East and North Africa's countries ready for the 21st century threats? This case of Saudi Aramco were the companys machines are infected with the Shamoon virus which requires the kind of Saudi Arabia to co-ordinate typical of state-sponsored attacks, and the targeting of critical infrastructure shortens the list of suspects.

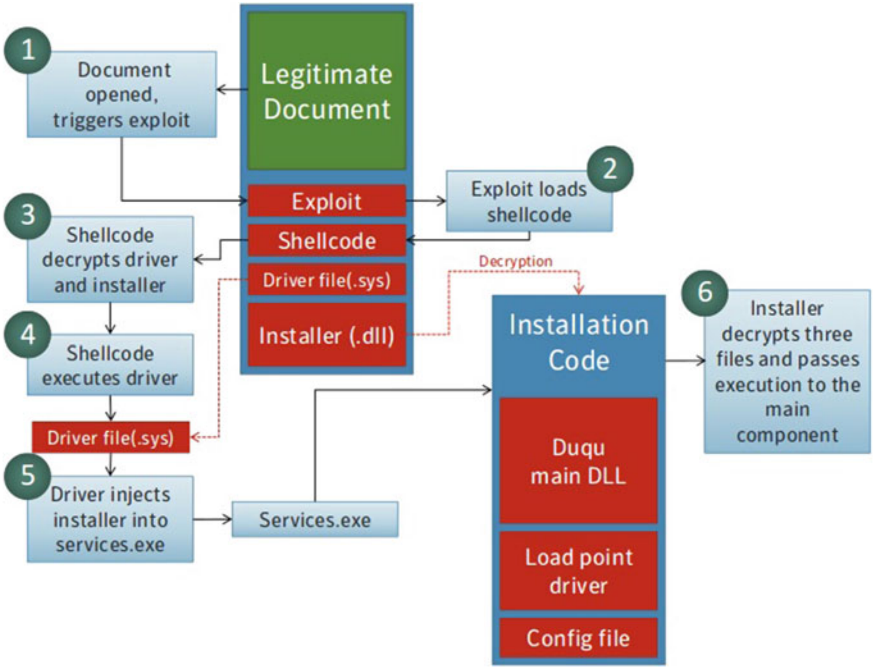

1
Document opened, triggers exploit
Legitimate Document
Exploit
Shellcode
Driver file(.sys)
Installer (.dll)
2
Exploit loads shellcode
3
Shellcode decrypts driver and installer
4
Shellcode executes driver
Driver file(.sys)
5
Driver injects installer into services.exe
Services.exe
Decryption
Installation Code
Duqu main DLL
Load point driver
Config file
6
Installer decrypts three files and passes execution to the main component

Fig. 8.11 Symantec: W32.Duqu installation process

8.7.3 Duqu

Star is another virus which was discovered by Iranian security specialists on April 2011. Iran government claim this is the second cyber attack against Iran [8]. None of the international security companies have had a chance to have a sample copy of this virus yet and some people have even doubt the actual virus does exist. Researchers from Kaspersky company believe that this virus can be a module of Duqu [33]. Duqu is the second globally recognized malware related to Stuxnet which was discovered around September 2011. Laboratory of Cryptography and System Security (CrySyS) of Budapest University of Technology and Economics have wrote a technically report on Duqu on October 2011 [34]. They have summarized their findings very briefly in two items:

- Stuxnet code is massively re-used in targeted attacks.
- A new digitally signed windows driver is used by attackers that was signed by another hardware manufacturer in Taiwan.

They have also mentioned ... we believe that the creator of Duqu had access to the source code of Stuxnet. Stuxnet and Duqu had digital signs from well known companies like RealTek, JMicron and C-Media. As we all know, digital sign will

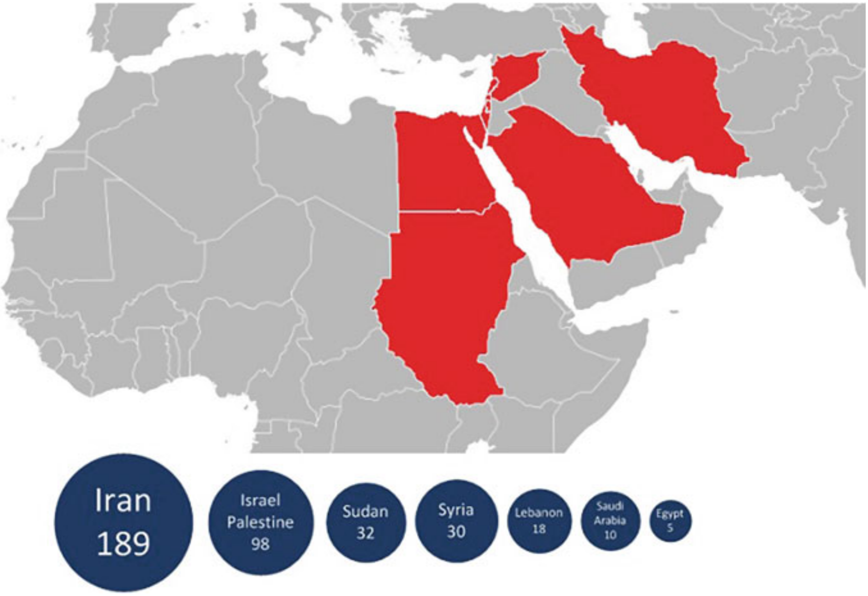


Fig. 8.12 Kaspersky: Flame malware infected countries

lead to a trust between anti virus programs and the virus, but what we don't know is who has had access to those digital signs and how?

Kaspersky lab has a page about Duqu. They say Duqu is sophisticated Trojan that seems to have been written by the same people who created the infamous Stuxnet worm. But unlike Stuxnet, whose main purpose was performing industrial sabotage, Duqu was created to collect intelligent about its targets. Basically it can steal anything on a targeted system; however, it looks like the attackers were particularly interested in collecting passwords, taking desktop screen shots (to spy on the users activity), and stealing various kind of documents. This heralds the entry into an era in which technology used by cyber criminals has advanced to a such a degree that they are capable of successfully carrying out industrial espionage and by extension, blackmail and extortion [35]. Kaspersky calls Duqu as a mystery because of the language which Duqu was using to communicate with command-and-control servers, the complexity level, different versions and modules. Part of Kaspersky blog, about the new version of Duqu they say the fact that the new driver was found in Iran confirms that most of Duqu incidents are related to this country. This means new Duqu variants were more active in Iran compared to other countries. The specialists believe Duqu installer uses the zero-day type of vulnerability in Win32k TrueType font-parsing engine and was wildly separated inside various office documents like Microsoft Word documents.

Symantec described this virus as Duqu is essentially the precursor to a future Stuxnet-like attack. The threat was written by the same authors, or those that have access to the Stuxnet source code, and the recovered samples have been created after the last-discovered version of Stuxnet. Duqu's purpose is to gather intelligence data and assets from entities such as industrial infrastructure and system manufacturers, amongst others not in the industrial sector, in order to more easily conduct a future attack against another third party. The attackers are looking for information such as design documents that could help them mount a future attack on various industries, including industrial control system facilities [36]. Duqu was initially a spy on targeted computers and industries, the virus was forwarded what it could found via secure protocols like https and ssh. Symantec described this feature of software as Duqu uses HTTP and HTTPS to communicate with a command and control (C&C) server. Duqu also has proxyaware routines, but these do not appear to be used by default. Each attack used one or more different C&C servers. The C&C servers were configured to simply forward all port 80 and 443 traffic to other servers. These servers may have forwarded traffic to further servers, making identification and recovery of the actual C&C server difficult. The traffic-forwarding C&C servers were scrubbed on October 20, 2011, so limited information was recovered. Even if the servers were not scrubbed, little actionable information would likely have been found due to their limited purpose of simply forwarding traffic. There are very limited information available around this software. The main reason is due to different versions of this virus and the virus where configured to automatically remove itself after 30 days of activity however this virus can download and install other modules as well.

Different security vendors have reported the existence of this virus in Iran, Egypt, Algeria, Sudan and some European countries in 15 different variants and 6 different modules.

8.7.4 Flame

Flame (also known as sKyWIper) was first discovered in Middle East on May 2012. It has been described as The name Flame comes from one of the attack modules, located at various places in the decrypted malware code. In fact this malware is a platform which is capable of receiving and installing various modules for different goals. Currently, none of the 43 tested anti viruses could detect any of the malicious components [37]. CrySyS describes Flame as Our first insight suggests that sKyWIper is another info-stealer malware with a modular structure incorporating multiple propagation and at tack techniques, but further analysis may discover components with other functionalities. In addition, sKyWIper may have been active for as long as 5–8 years, or even more. sKyWIper uses compression and encryption techniques to encode its files. More specifically, it uses 5 different encryption methods (and some variants), 3 different compression techniques, and at least 5 different file formats (and some proprietary formats too). It also uses special code injection techniques. Quite interestingly, sKyWIper stores information that it gathers on infected systems in a highly structured format in SQLite databases. Another uncommon feature of sKyWIper is the usage of the Lua scripting language. sKyWIper has very advanced functionality to steal information and to propagate. Multiple exploits and propagation methods can be freely configured by the attackers. Information gathering from a large network of infected computers was never crafted as carefully as in sKyWIper.

The malware is most likely capable to use all of the computers functionalities for its goals. It covers all major possibilities to gather intelligence, including keyboard, screen, microphone, storage devices, network, wifi, Bluetooth, USB and system processes. The results of our technical analysis support the hypotheses that sKyWIper was developed by a government agency of a nation state with significant budget and effort, and it may be related to cyber warfare activities. sKyWIper is certainly the most sophisticated malware we encountered during our practice; arguably, it is the most complex malware ever found [38]. According to Kaspersky security specialists: Currently there are three known classes of players who develop malware and spyware: hacktivist, cyber-criminals and nation states. Flame is not designed to steal money from bank accounts. It is also different from rather simple hack tools and malware used by the hacktivist. So by excluding cyber-criminals and hacktivist, we come to conclusion that it most likely belongs to the third group. In addition, the geography of the targets (certain states are in the Middle East) and also the complexity of the threat leaves no doubt about it being a nation state that sponsored the research that went into it [39].

8.8 Discussion

For all mentioned examples in the case study, there are many other cybercrime case attached in the region such as, the UAEs e-government sites that have been attacked by hackers, which caused financial loss and propagation of secret and confidential information to the public; Moreover, the Al-Jazeera website is another example of hacking big names inside the region. In fact that the investments in IT infrastructure have increased the value of e-business and e-governments and have created great opportunities for small and medium size businesses in the region, which helps in solving the unemployment problem. The lack of security awareness and security policy enforcement is one of the biggest problems inside IT companies in the Middle East and North Africa as most important IT users and decision makers are not aware of the growth of the cybercrime problems. Poor security policy enforcement means that investments and chance to fight in the level of cybercrime are minimal which leaves the business across the Middle East and North Africa vulnerable to cybercrime or online attacks.

As can be seen in Table 8.1 that African Internet users were 4,514,400 in 2000 which was 32 times less than Asian users which were 114,304,000 users. However, in 2012 the number of Asian Internet users has been increased to be 1,076,681,059 which is 6.4 time greater than African users 167,335,676. Thus, the penetration rate of African 3,606.7 % which is greater than the penetration rate of Asian 841.9 %. Tables 8.2 and 8.3 indicate that Qatar has the best penetration rate with more than 85 % of its population among Middle East and the North Africa countries. Most of Middle East countries have penetration rate less than 50 % except Iraq, Jordan, KSA, Syria and Yemen with 7.7, 38.1, 49.0, 22.5 and 14.9 % respectively. However, in North Africa only Morocco has penetration rate greater than 50 %. As consequence the Internet users are growing in Middle East faster than North Africa. Therefore, in these growth continues, there will be millions of future Internet user's potential cybercrime victims in these particular regions. In fact Bahrain, Iran, Kuwait, Lebanon, Oman, Qatar, KSA, UAE, Qatar, Egypt, Algeria, Morocco and Tunisia have accepted this risk and started working on a national cyber security policy. Each of these regions countries need to identify and mitigate their unique cyber security vulnerabilities and threats through joint initiatives and sharing of best practices.

Finally, the cyber criminals are not just subjected to individual people and hackers. There are governmental cyber wars which countries in Middle East and North Africa should be aware of them and prepare their own infrastructure not just for detect and trace a small number of cyber criminals, instead to detect a giant cyber-attack and do appropriate cyber defense.

8.9 Challenges

There are many challenges need to be defeat in order to improve the cybercrime investigation system in Middle east and North Africa such as lack of comprehensive study of the main influencing factors of cybercrime investigation system in Middle East and North Africa. To the best of our knowledge, there has not been any scientific study on the main factors or barriers that influence the work of cybercrime investigators in these particular regions. One of the important factors is the cultural and social considerations such as user behaviors and the lack of knowledge exchange about cybercrime between policy officers of these countries. Moreover, the most of these countries are still in the earlier stages of their E-services implementation. In fact cybercrime occurs in a different context and therefore presents different issues. The investigation in cybercrime entails special challenges not encountered in many traditional investigation includes:

- Cybercrimes and their victims often require a proactive enforcement strategy of contacting potential victims and improving investigation capabilities.
- Committing cybercrime can be made easier through specialist software tools, hardware and Internet access.
- Availability of information on the Internet can be used for both legitimate and criminal purposes.
- Legislation, that involves the criminal offenses, requirements to open an investigation, evidences.
- Public and private sector cooperation should assist and exchange of information with each other of any related information to computer crime victim, evidence, legislation, ...etc.
- Middle East and North Africa should internationally cooperate with other countries to exchange any information related to computer crime victims.

- Dedicated Unit involves the legal framework, field offices, and competence offices, trained and skilled officers.
- Criminal investigative procedures should allow computer and Internet access, data preservation and supports procedure complaint.
- The criminal investigative procedures should support an investigation, surveillance, identifying IP, and phone users and monitoring of phone conversation.
- Criminals need not be present at the same location as the target. As the criminal location can be completely different from the crime site.
- Skilled cybercrime police investigators and up to date computer forensic examiners with cybercrime familiarity.
- The popularity of the Internet and its services are growing fast, with over 2 billion Internet users worldwide by 2012. Where Africa and Middle East have the best penetration rate 3,606.7–2,639.9 % respectively Table 8.1.

Finally, the most important challenge in the computer crime investigation procedure is to understand the suspicious criminal activity and prove it [40].

8.10 Conclusion

The North Africa's and some Middle East countries are still in the earlier stages of their E-services implementation. Therefore, it is very important for these countries governments and organizations to maintain and improve their cybercrime investigators quality. There is no doubt that the context of cybercrime is feasible to the cybercrime investigators but our chapter discussed the many challenges to improve the cybercrime investigation system in Middle East and North Africa and highlighted the most important factors that effects the state of information technology security which are the IT infrastructure, growth of IT users and the lack of regulation and training of law enforcements. This chapter concentrated on the lack of cybercrime investigation experiences exchange between, and the need of international cooperation with other countries to exchange any information related to cybercrime victims. In case of this growth in volume continues, there will be millions of future Internet users' potential cybercrime victims in these regions. The exponential growth of the users in Middle East and North Africa will force Internet service providers to lower service prices and improve their services, benefiting both end users and attackers. The investigation and prosecution of cybercrime presents a number of challenges for law-enforcement agencies. It is vital not only to educate the people involved in the fight against cybercrime, but also to draft adequate and effective legislation. Worse than that cyber criminals are moving beyond computers and attaching different kind of mobile devices. Thus, it is time for these countries governments to comprehensively study all related issues in their cybercrime investigation system that analysis and discusses all challenges and barriers to improve the cybercrime investigators procedure. Whereas, this study should emphasize the need of single information security law in these particular regions (Middle East and North Africa) and urgently ensure adequate analytical and technical capabilities for law enforcement.

References

1. Loucif, K.: Africa, A New Safe Harbor for Cybercriminals? Trend Micro Incorporated (2013)
2. Wolf, P.: The South African Cyber Threat Barometer, A strategic public-private partnership (PPP) initiative to combat cybercrime in SA (2013)
3. Alalwan, N. Alzahrani, A. and Sarrab, M.: Cybercrime investigation challenges For gulf cooperation council governments: a survey. In: Proceedings of the Eighth International Conference on Forensic Computer Science—Icofcs, Brasilia, Brazil, pp. 33–36 (2013)
4. Timothy, E.: The field guide for investigating computer crime, part eight: information discovery. Symantec (2001)
5. Bahar, A.: Computer crime investigation. MSc Thesis, De Montfort University, Leicester, UK (2010)
6. Richter, J., Kuntze, N., Rudolph, C.: Securing digital evidence. In: Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering (2010)
7. David, I., Karl, S.: Computer Crime: A Crime fighters Handbook. OReilly Associates Inc, Sebastopol (1995)
8. Loia, V., Mattiucci, M., Senatore, S., Veniero, M.: Computer crime investigation by means of fuzzy semantic maps. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (2009)
9. Simundic, S., Franjic, S., Susic, T.: Databases and computer crime. In: 52nd International Symposium ELMAR, pp. 195–201 (2010)
10. Yousef, H., Iqbal, A.: Digital forensics education in UAE. In: 6th International Conference on Internet Technology and Secured Transactions, Abu Dhabi, UAE (2011)
11. Sammes, A., Jenkinson, B.: Forensic Computing a Practitioners Guide. Springer, New York (2000)
12. Casey, E.: Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet, 1st edn. Academic Press, London (2000)
13. Casey, E.: Error, uncertainty, and loss in digital evidence. Int. J. Digital Evid. **1**, 2 (2002)
14. McLean, S.: Basic considerations in investigating computer crime, executing computer search warrants and seizing high technology equipment. In: 14th BILETA Conference: CYBERSPACE. Crime, Criminal Justice and the Internet (1999)
15. Prosise, C., Mandia, K.: Incident Response: Investigating Computer Crime. Mcgraw-Hill Osborne Media, San Francisco (2001)
16. Stephenson, P.: Investigating Computer Related Crime. CRC Press, Boca Raton (1999)
17. Warren B.: Challenges to criminal law making in the new global information society: a critical comparative study of the adequacies of computer-related criminal legislation in the United States, the United Kingdom and Singapore. www.law.ed.ac.uk/ahrc/complaw/docs/chik.doc visited 28 October (2011)
18. International Telecommunication Union, ICTs in Africa: Digital Divide to Digital Opportunity (2008)
19. Digital divide, Internet in Africa, wikipedia (2013)
20. United Nations, The Millennium Development Goals Report, We Can end poverty 2015 Millennium Development Goals New York (2013)
21. The Emerging Africa Infrastructure Fund: Sub-Saharan Africa (Multiple Countries), Telecoms, Seacom Undersea Cable (2007)
22. Fiber Optic Link Around the Globe, wikipedia (2013)
23. Internet World Stats, Usage and Population Statstics. http://www.internetworldstats.com/list2.htm (2013)
24. Kingsley, K.: AfricaNews reporter in Abidjan, Ivory Coast, Africa's population now 1 billion. africanews.com (2009)
25. El-Guindy M.: Saudi Aramco cyber-attack, are we ready Net Safe, Middle East. http://netsafe.me2012-08-27saudi-aramco-cyber-attack-are-we-ready/more-535 (2012)
26. Kim Z.: Legal experts: stuxnet attack on Iran was illegal act of force. http://www.wired.com (2013)

27. Symantec Employee, Stuxnet 0.5: The Missing Link, Symantec security response. http://www.symantec.com (2013)
28. Nicolas, F., Liam, M., Eric, C.: W32.Stuxnet Dossier, Symantec security response (2011)
29. David, K.: The Real Story of Stuxnet, How Kaspersky Lab tracked down the malware that stymied Irans nuclear-fuel enrichment program. http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet (2013)
30. Jarrad, S.: W32.Stuxnet, Symantec. http://www.symantec.com (2010)
31. Thomson, I.: The Register, Snowden: US and Israel did create Stuxnet attack code, UK is 'radioactive' and 'Queen's selfies to the pool boy' slurped. http://www.theregister.co.uk/ (2013)
32. Waterman, S.: The Washington Times, U.S.-Israeli cyberattack on Iran was act of force, NATO study found, Strike devastated nuclear program. http://www.washingtontimes.com/news/2013/mar/24/ (2013)
33. Alexander, G.: Securelist, The Duqu Saga Continues: Enter Mr. B. Jason and TVs Dexter (2011)
34. Boldizsar, B., Gabor, P., Levente, B., Mark, F.: Duqu: A Stuxnet-like malware found in the wild, Laboratory of Cryptography and System Security. CrySyS), Budapest University of Technology and Economics, Department of Telecommunications (2011)
35. Kaspersky.: Duqu: Steal Everything, Kaspersky Labs investigation. http://www.kaspersky.com (2012)
36. Symantec.: W32.Duqu, The precursor to the next Stuxnet, The Laboratory of cryptography and system security (CrySyS) (2011)
37. MAHER.: Identification of a New Targeted Cyber-Attack, MAHER, Copmuter Emergency Response Team Coordination Center, Information Technology Organization of Iran, Ministry of Information and Communication Technology (ICT) (2012)
38. sKyWIper Analysis Team, sKyWIper (a.k.a. Flame a.k.a. Flamer): A complex malware for targeted attacks, Laboratory of Cryptography and System Security (CrySyS), Budapest University of Technology and Economics, Department of Telecommunications (2012)
39. Alexander, G.: Securelist, The flame: questions and answers. https://www.securelist.com (2012)
40. Kabay, S.: Computer Security Handbook, Wiley, Hoboken (2002)

Chapter 9
Multilayer Machine Learning-Based Intrusion Detection System

Amira Sayed A. Aziz and Aboul Ella Hassanien

Abstract Almost daily we hear news about a security breach somewhere, as hackers are constantly finding new ways to get around even the most complex firewalls and security systems. This turned the security into one of the top research areas. Artificial Immune Systems are techniques inspired by biological immune system—specifically the human immune system—which basic function is to protect the body (system) and defend against attacks of different types. For this reason, many have applied the artificial immune system in the field of network security and intrusion detection. In this chapter, a basic model of a multi-layer system is discussed, along with the basics of artificial immune systems and network intrusion detection. An actual experiment is included, which involved a layer for data preprocessing and feature selection (using Principal Component Analysis), a layer for detectors generation and anomaly detection (Using Genetic Algorithm with Negative Selection Approach), and finally a layer for detected anomalies classification (using decision tree classifiers). The principle interest of this work is to benchmark the performance of the proposed multilayer IDS system by using NSL-KDD benchmark data set used by IDS researchers. The obtained results of the anomaly detection layer shows that up to 81 % of the attacks were successfully detected as attacks. The results of the classification layer demonstrated that naive bayes classifier has better classification accuracy in the case of lower presented attacks such as U2R and R2L, while the J48 decision tree classifier gives high accuracy up to 82 % for DoS attacks and 65.4 % for probe attacks in the anomaly traffic.

Keywords Artificial immune systems · Anomaly intrusion detection · Machine learning · Computational intelligence

A. S. A. Aziz (✉) · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Cairo University, Cairo, Egypt
e-mail: amira.aly.fci@gmail.com
URL: http://www.egyptscience.net/

A. S. A. Aziz
Faculty of Business Administration and Information Systems,
Université Française d'Egypte (UFE), Cairo, Egypt

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_9, © Springer-Verlag Berlin Heidelberg 2014

9.1 Introduction

Networks are becoming more vulnerable by time to intrusions and attacks, from inside and outside. Cyber-attacks are making news headlines worldwide, as threats to networks are getting bolder and more sophisticated. Reports of 2011 and 2012 are showing an increase in network attacks, with Denial of Service (DoS) and targeted attacks having a big share in it, as reported by many web sites like [1–3].

Internal threats and Advanced Persistent Threats (APT) are the biggest threats to a network, as they are carefully constructed and dangerous, due to internal users' privileges to access network resources. With this in mind, and the increasing sophistication of attacks, new approaches to protect the network resources are always under investigation, and the one that is concerned with inside and outside threats is the Intrusion Detection System.

Nature has always been an inspiration to researchers with its diversity and robustness of its systems, and Artificial Immune Systems are one of them. Many algorithms were inspired by ongoing discoveries of biological immune systems techniques and approaches. One of the basic and most common approach is the Negative Selection Approach, which is simple and easy to implement. It was applied in many fields, but mostly in anomaly detection for the similarity of its basic idea.

In general, a single technique is not enough any more to work by itself for intrusion detection, as attacks and intrusions are becoming more complex and diverse. Hence, hybrid systems and multi-layer systems are emerging as a way to combine more than a technique in a single system to give the best results.A multi-layer artificial immune system implemented and presented at the end of this chapter, composes of two phases—implemented as three layers. The first phase is to classify the traffic into normal and anomaly ones. The second phase involves the usage of classifiers to label the detected anomalies with their right class or find out that it was originally normal but wrongfully classified as anomaly. Classifiers are simply a set of tools that is used to classify some given data into their correct classes, on the basis of some analysis and learning through a training process on previously labelled data. The detected anomalies are then the input to the classifiers mentioned in this chapter to be classified and labelled.

Many works combined machine learning techniques for intrusion detection and classification in the past. In [4], they combined decision trees with SVM as a hierarchical hybrid intelligent system model to maximize detection accuracy and minimize computational complexity. In [5] they combined Naive Bayesian (NB) classifier and decision trees (DT) to perform balance detections and keep false positives within acceptable levels for different types of network attacks. Both techniques were also combined in [6] as supervised (DT) and unsupervised (NB) classifiers in a multiple-level hybrid classifier. Supervised and unsupervised classification techniques were also combined in [7]. First, a pre-classification phase is applied to improve attack detection by obtaining the normal patterns of the users activities. Clustering techniques used in this phase are K-means, fuzzy C-means, and Gravitational Search Algorithm (GSA). Then, a hybrid classification approach was applied to enhance the detection accuracy as a second phase. Classification techniques applied in this phase were SVM and GSA. Our last example is [8], where they applied neural networks and SVM to discover useful patterns (features) that best describe user behavior on a system. Then, the relevant feature set is used to build classifiers that would recognize anomalies and known intrusions in (almost) real time.

The rest of this chapter is organized as follows. Section 9.2 presents a background of different techniques implemented and investigated in the proposed model. In Sect. 9.3, the proposed model is introduced and discussed, while in Sect. 9.4 the actual experiment is explained, with a review of the tested dataset (NSL-KDD) and the results obtained. Finally, a conclusion is given in Sect. 9.5.

9.2 Background

9.2.1 Intrusion Detection Systems

Intrusion Detection Systems (IDSs) are security tools used to detect anomalous or malicious activities from inside and outside intruders [9]. Such activities that violate the security policies of the system are considered anomalous and an alert should be raised by the IDS. An intrusion can be an attack from the Internet, attempts from authorized users of the system to gain more privileges, or an authorized user who misuse their privileges. Hence, an IDS has three basic functions [9, 10]:

1. Monitoring information sources: monitor activities concerning sources such as computers or networks for unauthorized access activities.
2. Analysis: detect unauthorized activities using events and data collected in the monitoring process. Misuse and anomaly detection analysis approaches are the most common.
3. Response: which is a set of actions the system takes when an intrusion is being detected.

Looking into more details of an IDS process, Fig. 9.1 gives a deeper look into this process [9].

Based on the approach, IDSs are classified to signature-based and anomaly-based. The former detects attacks by comparing the data to patterns stored in a signature database of known attacks. The later detects anomalies by defining a model of normal behaviour of the monitored system, then considers any behaviour lying outside the model as anomalous or suspicious activity. Signature-based IDS can detect well-known attacks with high accuracy but fails to detect or find unknown attacks. Anomaly-based IDS has the ability to detect new or unknown attacks but usually has high false positives rate (normal activities detected as anomalous). There are three types of anomaly detection techniques: statistical-based, knowledge-based, and machine learning-based [9]. IDS performance can be measured by two key aspects: the detection process efficiency and the involved cost of the operation [10, 11].

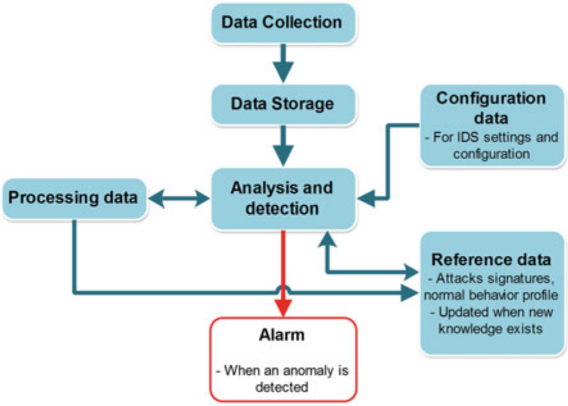

Data Collection
Data Storage
Configuration data
- For IDS settings and configuration
Processing data
Analysis and detection
Reference data
- Attacks signatures, normal behavior profile
- Updated when new knowledge exists
Alarm
- When an anomaly is detected

Fig. 9.1 Intrusion detection systems detailed process

9.2.2 Genetic Algorithm

Machine learning techniques are used to create rules for the intrusion detection systems, and genetic algorithms is a common algorithm that is been used for such purpose. Genetic Algorithms (GA) are search algorithms inspired by evolution and natural selection, and they can be used to solve different and diverse types of problems. The algorithm starts with a group of individuals (chromosomes) called a population. Each chromosome is composed of a sequence of genes that would be bits, characters, or numbers. Reproduction is achieved using crossover (2 parents are used to produce 1 or 2 children) and mutation (alteration of a gene or more). Each chromosome is evaluated using a fitness function, which defines which chromosomes are highly-fitted in the environment. The process is iterated for multiple times for a number of generations until optimal solution is reached. The reached solution could be a single individual or a group of individuals obtained by repeating the GA process for many runs [12, 13]. The whole process is shown in Fig. 9.2.

9.2.3 Principal Component Analysis

For each problem with some sample, there is a maximum number of features where performance degrades instead of improves which is called the curse of dimensionality. An accurate mapping of lower-dimensional space of features is needed so no information is lost by discarding important and basic features. The feature selection is an essential machine learning technique that is important and efficient in building classification systems. When used to reduce features, it results in lower computation costs and better classification performance.

Principal Component Analysis (PCA) is a technique that is usually used for classification and compression, as it reduces the data set dimensionality by extracting

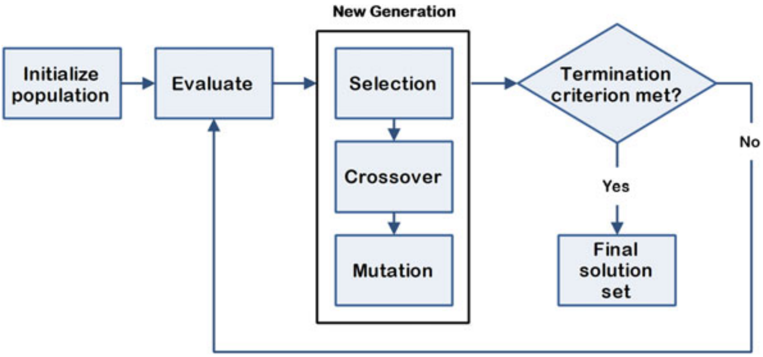
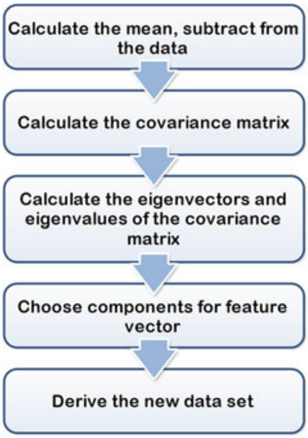


Fig. 9.2 The genetic algorithm process

Fig. 9.3 The principal components analysis feature reduction process


Calculate the mean, subtract from the data
Calculate the covariance matrix
Calculate the eigenvectors and eigenvalues of the covariance matrix
Choose components for feature vector
Derive the new data set

a new feature set that's smaller than the original one. The new extracted feature set includes most of the sample data information, that is the present variation given by the correlations between the original variables. PCA helps identifying the patterns in data in a way that highlights their similarity and differences, by the feature reduction process. The new features—called Principal Components (PCs)—are ordered by the amount of total information retained, and they are uncorrelated [14, 15]. It calculates the eigenvectors of the covariance matrix to find the independent axes of the data. Figure 9.3 shows the steps of the PCA process for features reduction.

PCA is based in the idea that most information about classes are within the directions with the largest variations. It works in terms of standardized linear projection that maximizes the variance in the projected space. It is a powerful tool in the case of high-dimensional data. The main problem with PCA is that it does not take into consideration the class label of the feature vector, hence it does not consider class separability.

9.2.4 Artificial Immune System

Artificial Immune Systems (AIS) [16–18] are a set of methodologies inspired by the Human Immune System (HIS), and are considered a branch of computational intelligence bio-inspired technology. AIS connects immunology with computer science and engineering. Attention was drawn to immune system as an inspiration to new approaches to solve complex problems. It mimics the HIS which is adaptive, distributed, tolerant, self-protective, and self-organizing with its many naturally-embedded techniques such as learning, feature extraction, and pattern recognition.

There are many methodologies within the immune system that can form an inspiration to a wide range of techniques. Those methodologies are Negative Selection Approach (NSA) [19], Artificial Immune Networks (AIN) [20, 21], and Clonal Selection Algorithm (CSA) [22]. Recent theories have also emerged such as Danger Theory (DT) [23, 24], Dendritic Cells Algorithms (DCA) [25] and Pattern Recognition Receptor Model (PRRM) [26, 27].

Some main features should exist in any AIS [16, 17] in order for its immunity mechanisms to function properly and successfully. An AIS should be embodied within other systems to provide similar properties to those of natural immune systems. AISs should provide homeostasis in which the system tries to stay within an optimal range for healthy functioning. While an AIS include innate and adaptive immune models, they should benefit from the interactions between these models to provide maximum security and reliability. An AIS also should consist of multiple, heterogeneous, interacting and communicating components, just like normal immune system and these components should be easily and naturally distributed. Finally, the most important feature of AISs is that they should be acquired to perform life-long learning.

The AIS is designed using a number of algorithms inspired by the HIS [18, 20]. There exists no single algorithm from which all immune algorithms are derived. The AISs are classified into first and second generation. The first generation AISs are simplistic models of immunology, i.e. negative and clonal selection; the second generation AISs apply the interdisciplinary research that allows for a much finer-grained encapsulation of underlying immunology, for example, the Dendritic Cell Algorithm.

9.2.5 Negative Selection Algorithm

In the beginning, scientists were concerned with understanding and implementing the computational and mathematical models in order to simulate HIS functions. Then, different approaches were uncovered that helped the scientists understand how all these mechanisms of the immune system work and function. The first of these approaches was the Negative Selection Approach (NSA), which explains how T-cells are being selected and their maturation in the system [28].

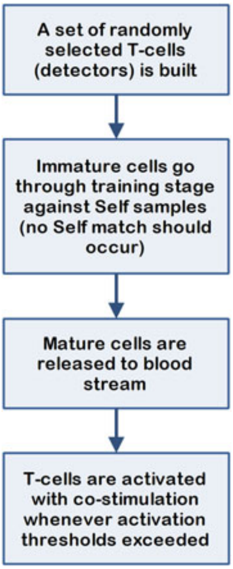

A set of randomly selected T-cells (detectors) is built
Immature cells go through training stage against Self samples (no Self match should occur)
Mature cells are released to blood stream
T-cells are activated with co-stimulation whenever activation thresholds exceeded

Fig. 9.4 Negative selection approach process

Perelsen et al. emerged with the theoretical model of the selection of T-cells in the Thymus, where the NSA was developed as a technique within AISs [18, 19]. Since 2004, there has been more collaboration between immunology practitioners and scientists with computer specialists for more understanding and development of practical immune algorithms than theoretical. The purpose of this approach is to provide tolerance for self cells, where it deals with the immune system's ability to detect unknown antigens while not reacting to the self cells.

The maturation of T-cells is in a 4-step process (as shown in Fig. 9.4): First, receptors (non-mature T-cells) are generated by a pseudo-random genetic rearrangement process. Then, they undergo a censoring process in the thymus, called the negative selection. In this process, the T-cells that react to self are destroyed, and those that do not bind to self-proteins are allowed to leave the thymus. These mature T-cells then circulate throughout the body to perform their immunological functions for the body protection.

9.3 The Proposed Multilayer Machine Learning-Based IDS

Some experiments were executed of multilayer anomaly intrusion detection systems inspired by immunity. Some implementations view a multilayer system as embedding innate and adaptive techniques, as done in [29] when they implemented a host-based multilayer AIS for positive and negative selection. Another team proposed a multilayer IDS targeting many levels of the system (host and network) that was under the development at the moment [30]. Their proposed system is inspired by immunity concepts to protect the user level, system/resource level, process level, and packet level. In [31] they implemented a dynamic multilayer model that is inspired by immunity for intrusion detection. They implemented both misuse and anomaly modules, where the misuse module stores and recalls specific signatures associated with previously detected attacks, and the anomaly module detect deviations from normal patterns.

There were some experiments executed by the authors of this chapter, where a multilayer AIS was developped for anomaly network intrusion detection. Two main layers were implemented: anomalies detection layer and anomalies classification layer, in addition to a preceding layer for data-preprocessing, where results were published and discussed in [32, 33]. Basically, an IDS is built using Genetic Algorithm for detectors generation and Principal Component Analysis (PCA) for feature selection, then some classification techniques are applied on the detected anomalies to define their classes including Naive Bayes, decision trees and multilayer perceptron neural networks. The experiment was run using the NSL-KDD IDS benchmark data set which contains four types of attacks including Denial of Service, Probe, User to Remote, and Remote to Local, beside the normal connections data.

The advantages of such system is that deviations from the normal behaviour (anomalies) are detected in the first layer, then they are forwarded to the next layer where the anomalies are either labelled with their right attack class (whether previously detected or new) or classified as normal, in that case a false alarm is being avoided by correctly classifying it as normal behavior. The proposed multi-layers machine learning techniques for anomalies detection and classification system is composed of the following three layers:

(1) **Feature selection based on principal component analysis**
(2) **Anomaly detection based on genetic algorithm with negative selection**
(3) **Label the detected anomalies using decision trees**

The overall architecture of the introduced approach is described in Fig. 9.5. These three layers are described in detail in the following subsections along with the steps involved for each layer.

9.3.1 Layer I: Feature Selection Based Principal Components Analysis Layer

In this layer, feature selection is applied to the data, using PCA. As mentioned before, PCA is a feature extraction algorithm, which tends to create a new subset of features by combining existing features. Reversing the operation of PCA to obtain the original features back, it can lead you to select features with the highest covariance values as the selected feature set.

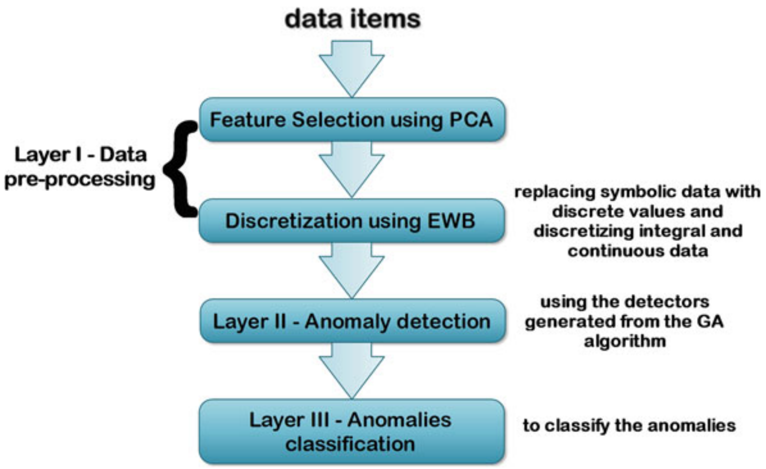

data items
Feature Selection using PCA
Layer I - Data pre-processing
Discretization using EWB
replacing symbolic data with discrete values and discretizing integral and continuous data
Layer II - Anomaly detection
using the detectors generated from the GA algorithm
Layer III - Anomalies classification
to classify the anomalies

Fig. 9.5 The proposed three-layer machine learning-based IDS

9.3.2 Layer II: Anomaly Detection-Based Genetic Algorithm with Negative Selection Layer

Data preprocessing is applied first to prepare the values for the GA operations. Preprocessing includes replacing symbolic feature values (service, protocol, flag) with discrete values. Discretization is applied after that to transform the continuous values of the features to discrete ones. Equal-Width Binning was used for discretization, where it divides the data into k intervals of equal size, without repeating values inside a single bin. The number of bins k for each feature calculated using the following equation:

$$k = max(1, 2 * \log l) \tag{9.1}$$

The GADG (Genetic Algorithm for Detectors Generation) was originally suggested and applied in [34, 35], where The algorithm sequence is shown in Algorithm 1. The purpose of this layer is to generate a set of detectors which are able to discriminate between self/normal and non-self/anomalous.

The detectors are first selected randomly from the self space (which is a population of normal individuals). A detector simply is a set of values representing the selected features. Then, using GA, new detectors are generated in which if they are better fitted, they replace the old ones. The fitness function is based on the ability of an individual to match normal records. This approach is the maturation process, where by the end the detectors know the representation of self/normal and assumes different representation is non-self/anomalous. Finally, the detectors are exposed to the data set to start the anomaly detection process, labelling the detected individuals as either normal or anomaly.

Algorithm 1: Genetic algorithm for detectors generation

```
Input: Initialize population by selecting random individuals from the Self space $S$;
for The specified number of generations do
    for The size of the population do
        Select two individuals (with uniform probability) as $parent_1$ and $parent_2$;
        Apply crossover to produce a new individual ($child$);
        Apply mutation to child;
        Calculate the distance between $child$ and $parent_1$ as $d_1$, and the distance between
        $child$ and $parent_2$ as $d_2$;
        Calculate the fitness of $child$, $parent_1$, and $parent_2$ as $f$, $f_1$, and $f_2$ respectively;
        if $(d_1 < d_2)$ and $(f > f_1)$ then
            replace $parent_1$ with $child$;
        end
        if $(d_2 <= d_1)$ and $(f > f_2)$ then
            replace $parent_2$ with $child$;
        end
    end
end
Result: Extract the best (highly-fitted) individuals as your final solution;
```

9.3.3 Layer III: Detected Anomalies Classification Layer

In anomaly NIDS, traffic is usually classified into either normal or a specific attack category. Hence, a multi-category classifier is needed for such type of classification. Multi-category classifiers are either direct or indirect. Direct classifiers generally extend binary classifiers to deal with multi-category classification problems, while indirect classifiers decomposes the multi-category problem into multiple binary classification problems. For indirect classifiers, a base classifier is used to train the binary classification problems set, and results are merged using a combining strategy which works on the collected results of the binary classifiers [36, 37].

So, after the anomaly detection process done using the generated detectors, we should have the connections classified to either normal or anomaly. The anomaly connections are then lunched into the classifiers to either confirm it is an attack and get more information about the attack type, or to find it is not an attack and it is a normal connection. A Decision Tree (DT) [38, 39] is a structure of layered nodes (a hierarchical organization of rules), where a non-terminal node represents a decision on a particular data item and a leaf (terminal) node represents a class. Advantages of DTs are that they are easy to interpret, can be modified when new scenarios appear, and they can be combined with other decision techniques. In this paper four types of decision trees were tested: J48 (C4.5) decision tree, Naive Bayes Tree (NBTree), Best-First Tree (BFTree), and Random Forrest (RFTree). Naive Bayes (NB) and Multilayer Perceptron neural network (MLP) were applied for comparison purposes.

Table 9.1 Distributions of atacks and normal NSL-KDD records

	Total records	Normal	DoS	Probe	U2R	R2L
Train_20 %	25,192	13,449	9,234	2,289	11	209
		53.39 %	36.65 %	9.09 %	0.04 %	0.83 %
Train_All	125,973	67,343	45,927	11,656	52	995
		53.46 %	36.46 %	9.25 %	0.04 %	0.79 %
Test	22,544	9,711	7,458	2,421	200	2,754
		43.08 %	33.08 %	10.74 %	0.887 %	12.22 %

Table 9.2 Known and new attacks in NSL-KDD test set

	DoS	Probe	U2R	R2L
Known	5,741	1,106	37	2,199
	76.98 %	45.68 %	18.50 %	79.85 %
New	1,717	1,315	163	555
	23.02 %	54.32 %	81.50 %	20.15 %

9.4 Experimental Results and Discussion

9.4.1 Data Set

The experiment was executed using the NSL-KDD IDS benchmark data set [40, 41], which solves some problems and deficiencies in the KDD Cup'99 data set [42]. The advantage of NSL KDD dataset are:

- No redundant records in the train set, which used to lead to unbiased results for the tested classifiers.
- Better reduction rates in the test set, again due to no redundant records.
- The number of selected records for each difficult level group is inversely proportional to the percentage of records in the original KDD data set.

The data set contains four types of attacks:

Denial-of-Service (DoS): makes a resource too busy to respond or handle requests.

Probe: to gather information about a computer or a network.

User-to-Root (U2R): an attacker logs as a normal user, then try to gain root access to the system.

Remote-to-Local (R2L): an attacker with no access to a certain machine, try to exploit some vulnerability to gain local access on that machine.

Tables 9.1 and 9.2 shows the distributions of anomalous and normal records in the dataset, and distributions of known and new attacks in the test set respectively.

There are 41 features in the data set, of different types: categorical, binary, real, and integral. These features are divided into three groups:

1. **Basic features:** which encapsulate all attributes that can be extracted from a TCP/IP connection. Most of them may lead to implicit delay in detection.
2. **Traffic (time-based) features:** they are computed using a window interval (2 s), and they are divided into "same host" features and "same service" features.
3. **Content features:** for R2L and U2R attacks, which are basically behavioral and do not contain sequential patterns like DoS and Probe attacks. These features are concerned with suspicious behavior in the data.

9.4.2 Results

Twenty two features out of 41 are selected based on PCA, which are listed in Table 9.3 along with their description and categories.

Figures 9.6, 9.7, 9.8, 9.9, 9.10, 9.11, 9.12, 9.13, 9.14, 9.15, 9.16 and 9.17 show the results of the anomaly detection process by detectors generated using Minkowski distance and detectors generated using Euclidean distance. The horizontal axis shows different population sizes with different number of generations tested and compared. *t* is the threshold value used to adjust results, and the selected values for test are 0.8, 1.5, 2.0. The performance measurements used are: Detection Rates (percentage of truly detected normal and anomalous traffic), False Positives Rates (percentage of truly detected anomalies), and False Negatives Rates (percentage of truly detected normal connections). Figures 9.6, 9.7, 9.8 show the detection results on the Train Set using the Minkowski detectors.

One can realize that higher values of threshold give better detection rates in general, as shown in Fig. 9.6. This is because as we can see in Fig. 9.8, True Negatives Rates (TNR) are very high and they are at their best values when higher threshold value is set (all 95 % and above with threshold value 2.0), the highest values are obtained by detectors generated by larger populations. On the other hand,we can realize in Fig. 9.7 that True Positives Rates (TPR) are at their best values with lower threshold values (88 % and higher with threshold value of 0.8). The highest values are obtained with detectors generated using lower population sizes. Figures 9.9, 9.10 and 9.11 show the detection results on the Test Set using Minkowski detectors.

The detection rates presented in Fig. 9.9 show that using a low threshold value give the best results, and detectors generated using lower population size give better results. The TPRs are at their best with detectors generated by lower populations, with 70 % and more of the anomalies detected as shown in Fig. 9.10. Although as shown in Fig. 9.11 that TNRs are at their best with higher threshold values, they are not that low with lower threshold, especially when it is needed to maintain high TPRs values. The following set of figure presents the anomaly detection results using Euclidean detectors. Figures 9.12, 9.13 and 9.14 show the detection results on the Train set.

As realized in Fig. 9.12, higher threshold values lead to better detection rates, also detectors generated with higher number of generations give the best results. On the

Table 9.3 Selected features by PCA

Feature	Description	Type
1. Duration	Duration of the connection	Integer
2. Protocol type	Connection protocol (e.g. tcp, udp)	Categorical
3. Service	Destination service (e.g. telnet, ftp)	Categorical
4. Flag	Status flag of the connection	Categorical
5. Source bytes	Bytes sent from source to destination	Integer
6. Destination bytes	Bytes sent from destination to source	Integer
7. Land	1 if connection is from/to the same host/port; 0 otherwise	Binary
8. Wrong fragment	Number of wrong fragments	Integer
9. Urgent	Number of urgent packets	Integer
11. Failed logins	Number of failed logins	Integer
13. Num compromised	Number of “compromised” conditions	Integer
14. Root shell	1 if root shell is obtained; 0 otherwise	Binary
17. Num file creations	Number of file creation operations	Integer
18. Num shells	Number of shell prompts	Integer
22. Is guest login	1 if the login is a “guest” login; 0 otherwise	Binary
27. Error rate	% of connections that have “REJ” errors	Real
28. srv error rate	% of connections that have “REJ” errors	Real
29. Same srv rate	% of connections to the same service	Real
31. srv diff host rate	% of connections to different hosts	Real
32. dst host count	Count of connections having the same destination host	Integer
34. dst host same srv rate	% of connections having the same destination host and using the same service	Real
37. dst host srv diff host rate	% of connections to the same service coming from different hosts	Real

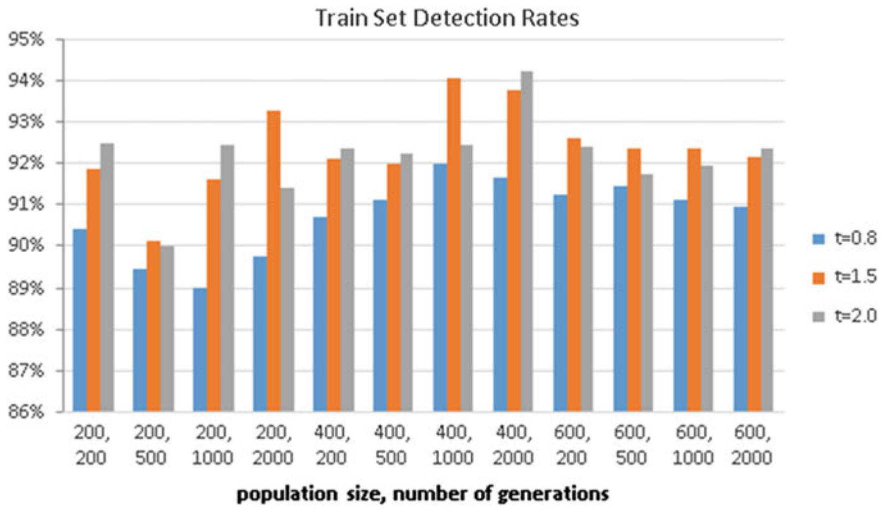


Fig. 9.6 The train set detection rates—Minkowski detectors

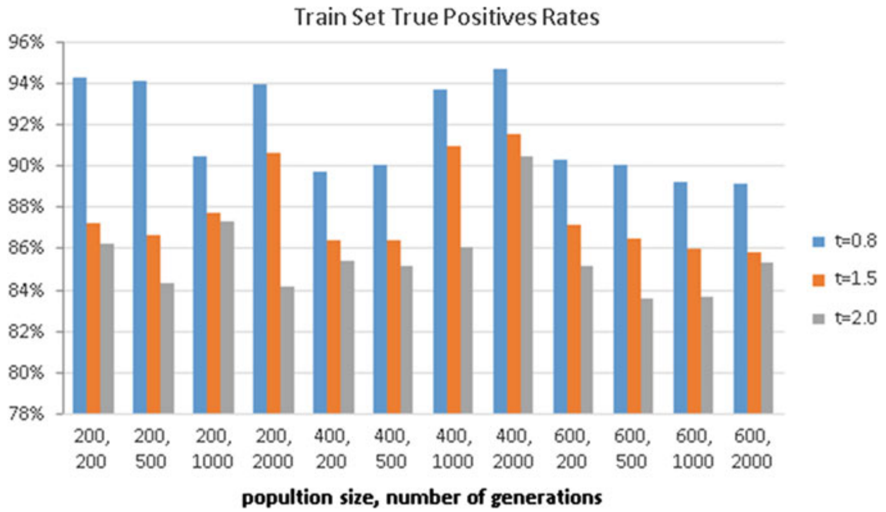


Fig. 9.7 The train set true positives rates—Minkowski detectors

other hand, with detectors generated by higher population sizes, lower threshold value give better results. TPRs in Fig. 9.13 are the highest with detectors generated by lower populations, and they are a bit higher than the results of the Minkowski detectors. For TNRs in Fig. 9.14, they are at their best with higher threshold values, and almost the same for all groups of detectors. Figures 9.15, 9.16 and 9.17 show the detection results on the Test Set using the Euclidean detectors.

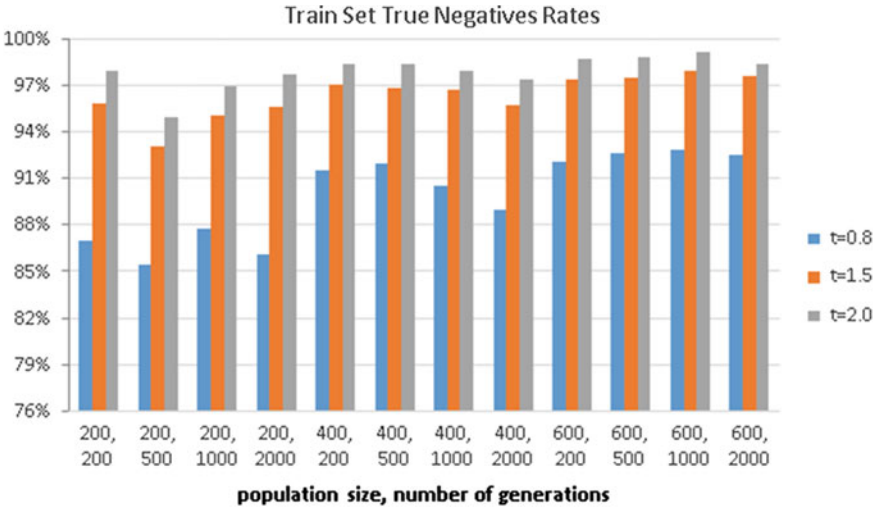


Fig. 9.8 The train set true negatives rates—Minkowski detectors

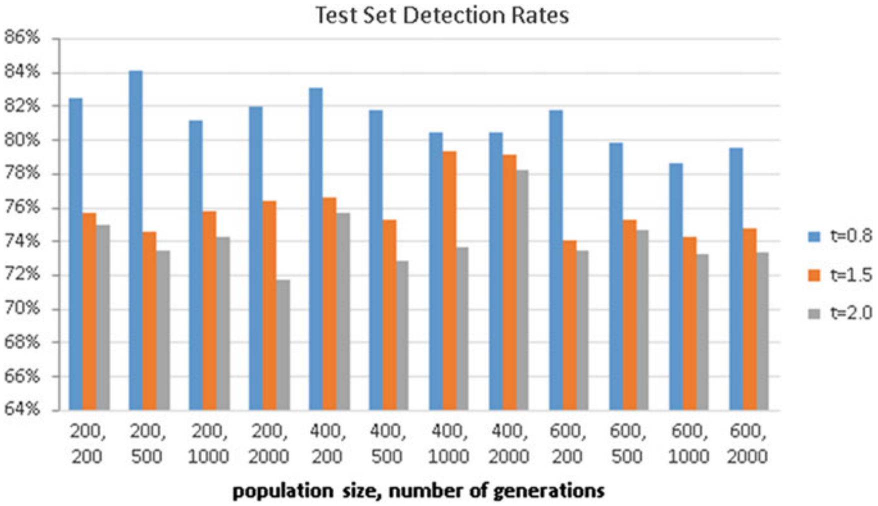


Fig. 9.9 The test set detection rates—Minkowski detectors

As shown in Fig. 9.15, the best detection rates obtained by detectors generated using lower population size and low threshold value. The TPRs too are higher with detectors obtained by low population size, also with low threshold value as seen in Fig. 9.16. Finally, we can see in Fig. 9.17 TNRs are all higher than 98 % with high threshold value, but using a low threshold value still give high rates with 90 % or more of the normal traffic detected successfully.

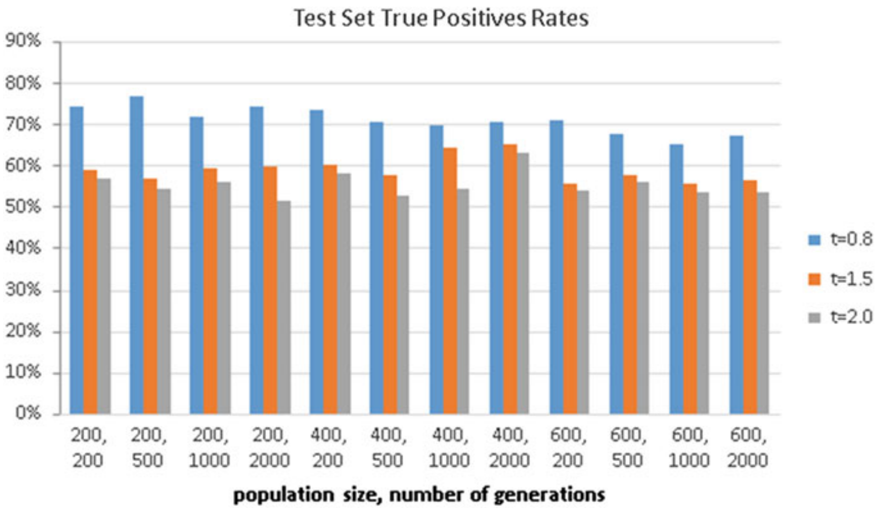


Fig. 9.10 The test set true positives rates—Minkowski detectors

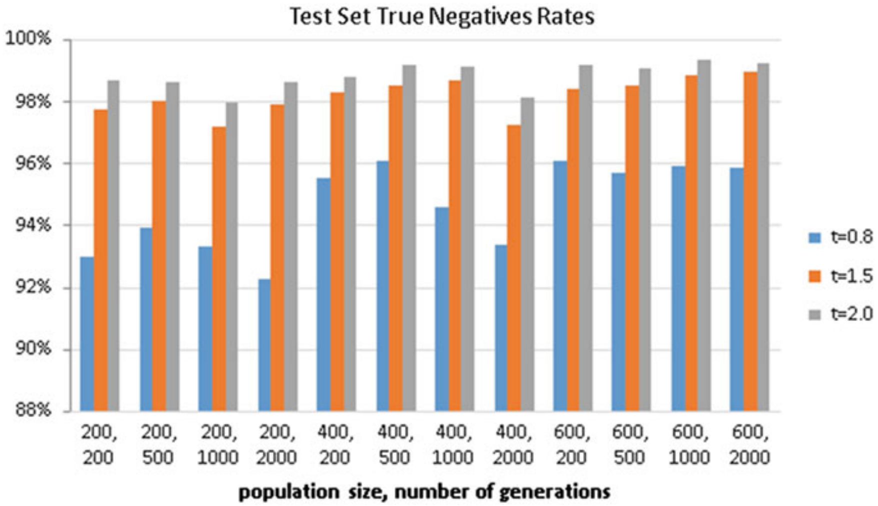


Fig. 9.11 The test set true negatives rates—Minkowski detectors

The best detection results obtained using the features selected by PCA were the ones generated using population size 200 for number of generation 200, applying the Euclidean distance measure for discrimination, and population size 200 for number of generations 500, applying the Minkowski distance measure for the discrimination. The anomaly detection results for each attack type are shown in Table 9.4, where it shows how much of each attack was detected successfully as anomalies. These group of results are chosen for the next detection layer, which is the classification phase.

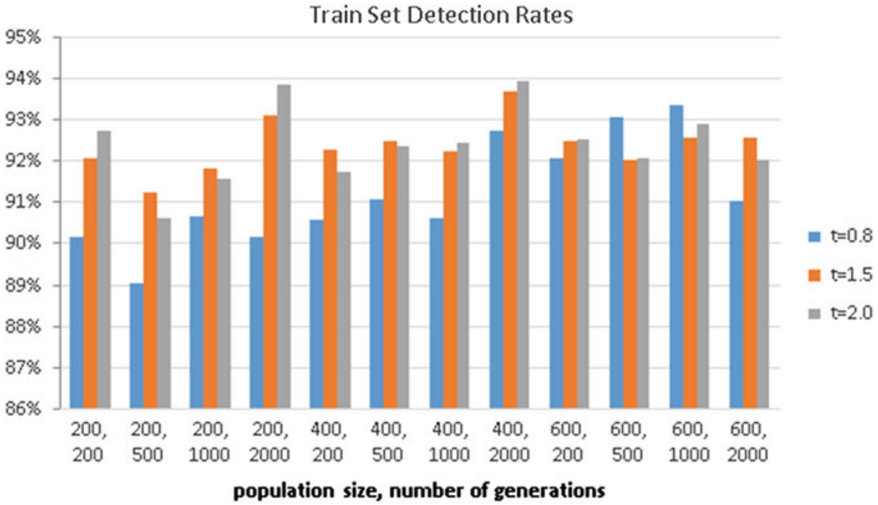


Fig. 9.12 The train set detection rates—Euclidean detectors

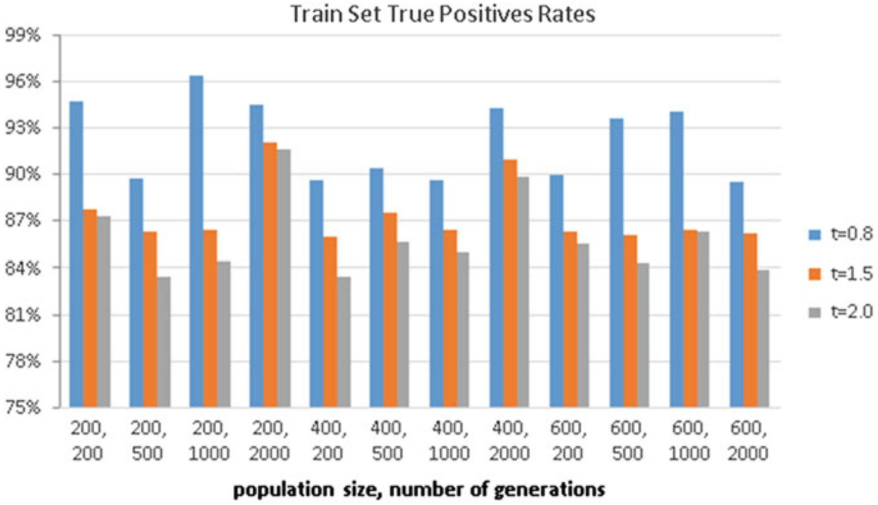


Fig. 9.13 The train set true positives rates—Euclidean detectors

Below are the results of the classification process using the classifiers described above. The classifiers were trained one time using 20 % of the training set, and another time using the whole training set. Tables 9.5, 9.6, 9.7, 9.8, 9.9 and 9.10 show the statistics of the classification results of both models of each classifier.

Looking into Table 9.5, one can realize that as an overall, the J48 classifier give the best results, then the MLP, followed by BFTree. Viewing the results from the detectors point of view, the anomalies detected using the Minkowski detectors are

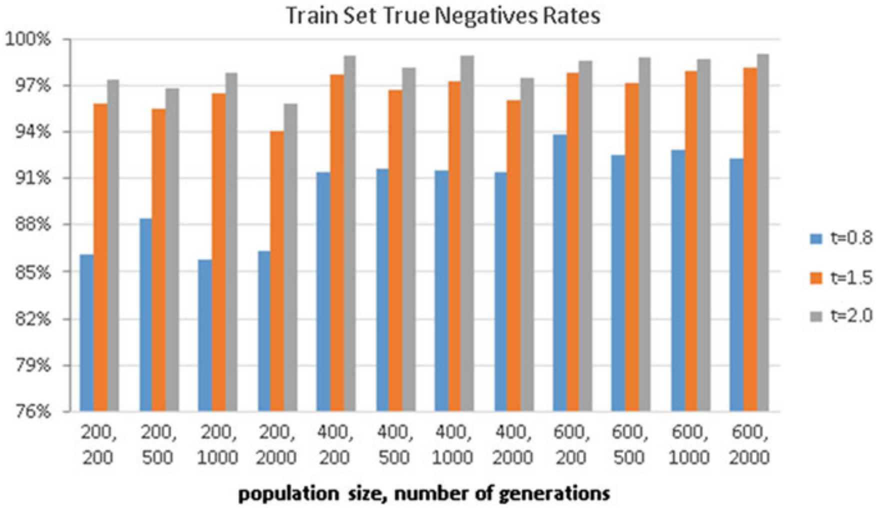


Fig. 9.14 The train set true negatives rates—Euclidean detectors

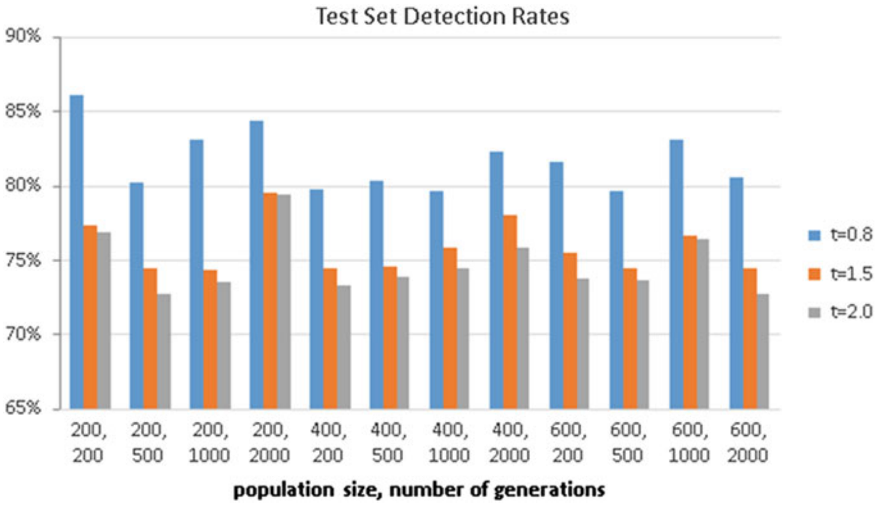


Fig. 9.15 The test set detection rates—Euclidean detectors

better classified than those detected using the Euclidean detectors. Based on the training set, NB and J48 classifiers give better results in general when trained using the whole train set, while other classifiers give better results when trained using 20 % of the train set.

In Table 9.6, the results of the DoS (Denial of Service) attack classification show that the J48 give the best results, using all the data in the train set, while the BFTree give close results using only 20 % of the train data to build the model. Surprisingly,

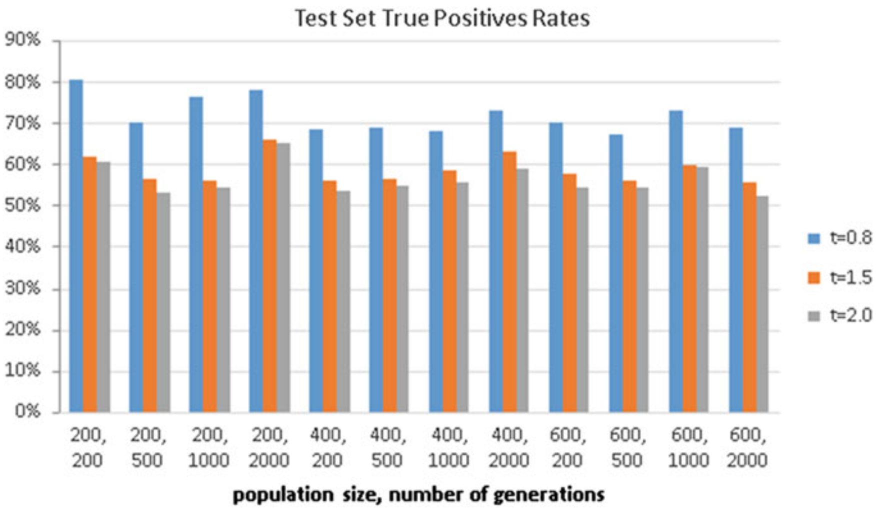


Fig. 9.16 The test set true positives rates—Euclidean detectors

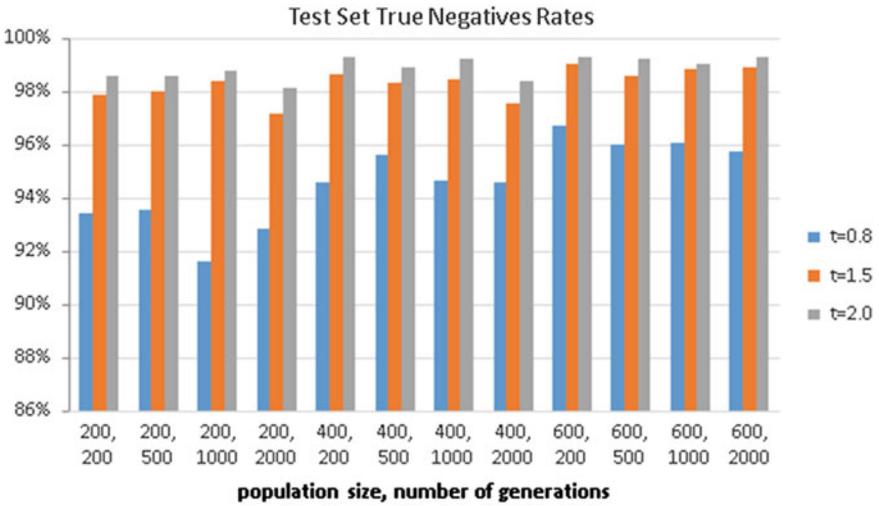


Fig. 9.17 The test set true negatives rates—Euclidean detectors

the NB classifier give close enough results of the correctly classified attacks, using all the train data to build the model.

As for the Probe attack results shown in Table 9.7, again the J48 gives the best classification results using the whole train set, while following are the BFTree using the whole train set and the NBTree using 20 % of the train set data.

Tables 9.8 and 9.9 show the classification results of R2L and U2R attacks (respectively), and due to their low representation in the data set the results are not very

Table 9.4 Layer II detection results in numbers

	DoS	Probe	U2R	R2L	Total
Euclidean detectors	6,873	2,174	172	1,111	10,330
	92.2 %	89.8 %	86 %	40.3 %	80.5 %
Minkoski detectors	6,891	2,140	150	693	9,844
	92 %	88.4 %	75 %	25.2 %	76.7 %

Table 9.5 Results of classification process

	Train data	NB (%)	BFtree (%)	J48 (%)	MLP (%)	NBtree (%)	RFtree (%)
Euclidean	All	63.49	65.41	70.07	66.20	65.25	63.91
	20	62.33	66.99	68.91	66.74	65.49	65.05
Minkowski	All	65.90	68.28	72.88	69.12	67.01	66.71
	20	64.81	69.54	71.96	69.82	67.41	67.95

Table 9.6 The DoS attack classification results

	Train data	NB	BFTree	J48	MLP	NBTree	RFTree
Euclidean	All	80.13	77.51	81.97	79.62	76.37	76.23
	20	76.79	81.32	81.74	78.92	75.02	77.81
Minkowski	All	80.16	77.48	81.96	79.76	76.34	76.20
	20	76.81	81.30	81.72	79.06	75.03	77.80

Table 9.7 The probe attack classification results

	Train data	NB	BFtree	J48	MLP	NBtree	RFtree
Euclidean	All	53.31	60.99	64.63	60.03	54.60	57.50
	20	54.46	56.12	61.04	54.55	61.22	58.56
Minkowski	All	54.11	61.45	65.42	60.47	55.23	58.18
	20	55.14	56.68	61.64	55.00	61.78	59.25

high. For the R2L classification, the NB classifier in general give the best results, followed by the MLP with Minkowski anomalies, using 20 % of the train set to build the model. The NB succeeds again to correctly classify 20 % of the U2R attacks while other classifiers almost fails to do so.

Finally for the normal data that were incorrectly detected as anomalies, as shown in Table 9.10 all classifiers, except the NB, were successful to correctly classify up to 80 % of the normal data, with the NBTree and RFTree giving the best results than others.

Table 9.8 The R2L attack classification results

	Train data	NB	BFtree	J48	MLP	NBtree	RFtree
Euclidean	All	18.00	3.15	13.95	0.99	18.63	1.35
	20	24.30	3.51	11.88	20.79	16.92	1.17
Minkowski	All	21.07	5.05	17.17	1.59	14.86	1.59
	20	32.90	0.58	18.33	33.33	12.41	1.73

Table 9.9 The U2R attack classification results

	Train data	NB	BFtree	J48	MLP	NBtree	RFtree
Euclidean	All	20.35	4.65	3.49	0.58	4.65	2.33
	20	16.28	5.81	1.74	0.00	2.91	0.58
Minkowski	All	13.33	2.00	2.67	0.67	2.67	1.33
	20	11.33	1.33	1.33	0.00	2.00	0.67

Table 9.10 The normal classification results

	Train data	NB (%)	BFtree (%)	J48 (%)	MLP (%)	NBtree (%)	RFtree (%)
Euclidean	All	9.89	75.20	76.30	74.10	79.43	78.81
	20	12.09	76.92	75.20	75.35	79.12	78.49
Minkowski	All	8.98	77.29	77.80	73.73	78.81	80.51
	20	11.53	77.97	76.95	76.95	80.51	80.00

9.5 Conclusion and Future Work

From the experimental results shown above, we can come up with few notifications. In general, decision trees give the best results. NBTree and J48 give better results in most cases if the whole train set is used to build the classification model, and BFTree gives better results in most cases when the model is built using 20 % of the training data. RFTree almost failed to detect attacks that have low representation in the data set, and for the other attacks it gave better results when 20 % of the train set was used to build the decision tree. So, in conclusion, when 20 % of the train set is used to build the classification model, obviously it takes less time for it to be built. Hence, if it gives the best results or even close to the best then it is better to use this portion of the train set, especially when time is a critical issue. Decision Trees have the advantage of being able to deal with different types of attributes, with transparency of knowledge, and are fast classifiers. Hence, they give best results when data is well represented in the training set. NBTree and BFTree take less time in training and classification than RFTree, and they give better results, so, they are better to use.

In conclusion, a multilayer system has the advantage of combining many techniques to have better results, each one (or hybrid techniques) is employed for a certain function in the system. Also, applying the anomaly detection process in a single phase gives the advantage of applying it on any system with any types of attacks, where they are only detected as anomalies and this is enough to raise an alarm. In the future, different classifiers will be investigated in order to find ones with better results, especially for the detection of behavioral attacks of R2L and U2R.

References

1. Teller, T.: The Biggest Cybersecurity Threats of 2013, Forbes magazine, May 2012
2. 2013 Cisco Annual Security Report, Cisco Systems
3. Worldwide Infrastructure Security Report, 2012 vol. VIII, ARBOR Networks
4. Peddabachigari, S., Abraham, A., Grosan, C., Thomas, J.: Modeling intrusion detection system using hybrid intelligent systems. J. Netw. Comput. Appl. **30**(1), 114–132 (2007)
5. Farid, D., Harbi, N., Rahman, M.Z.: Combining naive bayes and decision tree for adaptive intrusion detection. arXiv, preprint arXiv:1005.4496 (2010)
6. Xiang, C., Yong, P.C., Meng, L.S.: Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees. Pattern Recogn. Lett. **29**(7), 918–924 (2008)
7. Omar, S., Ngadi, A., Jebur, H.H.: An adaptive intrusion detection model based on machine learning techniques. Int. J. Comput. Appl. **70** (2013)
8. Mukkamala, S., Janoski, G., Sung, A.: Intrusion detection using neural networks and support vector machines. In: Proceedings of the 2002 International Joint Conference on Neural Networks, IJCNN'02, IEEE, vol. 2, pp. 1702–1707 (2002)
9. Aleksandar, L., Vipin, K., Jaideep, S.: Intrusion detection: a survey. In: Kumar, V. et al. (eds.) Managing Cyber Threats Issues, Approaches, and Challenges, vol. 5, pp. 19–78 (2005)
10. Murali, A., Roa, M.: A survey on intrusion detection approaches. First International Conference on Information and Communication Technologies. pp. 233–240 (2005)
11. Garcia-Teodora, P., Diaz-Verdejo, J., Macia-Fernandez, G., Vazquez, E.: Anomaly-based network intrusion detection: techniques, systems and challenges. Comput. Secur. **28**(1–2), 18–28 (2009)
12. Li, W.: Using genetic algorithm for network intrusion detection. Proceedings of the United States Department of Energy Cyber Security Grou, Training Conference vol. 8, pp. 24–27 (2004)
13. Sinclair, C., Pierce, L., Matzner, S.: An application of machine learning to network intrusion detection. In: Proceedings of 15th Annual Computer Security Applications Conference, ACSAC'99, pp. 371–377, IEEE (1999)
14. Jolliffe, I.: Principal Component Analysis. John Wiley & Sons Ltd, New York (2005)
15. Smith, L.I.: A tutorial on principal components analysis. Cornell University, USA vol. 51, pp. 52 (2002)
16. Hofmeyr, S.A., Forrest, S.: Immunity by design: an artificial immune system. Proceedings of Genetic and Evolutionary Computation Conference, pp. 1289–1296 (1999)
17. Aickelin, U., Dasgupta, D.: Artificial immune systems tutorial. In: Burke, E., Kendall, G. (eds.) Search Methodologies Introductory Tutorials in Optimization and Decision Support Techniques. Kluwer, pp. 375–399 (2005)
18. Greensmith, J., Whitbrook, A., Aickelin, U.: Artificial immune systems. Handbook of Metaheuristics, pp. 421–448. Springer, US (2010)
19. Forrest, S.: Self-nonself discrimination in a computer. IEEE Computer Society Symposium on Research in Security and Privacy, pp. 202–212 (1994)
20. Shen, X., Gao, X.Z., Bie, R., Jin, X.: Artificial immune networks: models and applications. International Conference on Computational Intelligence and Security, vol. 1, pp. 394–397 (2006)

21. Galeano, G.C., Veloza-Suan, A., Gonzalez, F.A.: A comparative analysis of artificial immune network models. Proceedings of the Conference on Genetic and Evolutionary Computation, GECCO ’05, pp. 361–368 (2005)
22. Ulutas, B.H., Kulturel-Konak, S.: A review of clonal selection algorithm and its applications. Artif. Intell. Rev. **36**(2), 117–138 (2011)
23. Iqbal, A., Maarof, M.A.: Danger theory and intelligent data processing. World Academy of Science, Engineering and Technology vol. 3 (2005)
24. Aickelin, U., Cayzer, S.: The danger theory and its application to artificial immune systems. Computing Research Repository—CORR 0801.3 (2008)
25. Greensmith, J., Aickelin, U., Cayzer, S.: Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. Proceedings ICARIS-2005, 4th International Conference on Artificial Immune Systems, LNCS 3627, pp. 153–167, Springer (2005)
26. de Castro, L.N., Timmis, J.: Artificial Immune System: A Novel Paradigm to Pattern Recognition. University of Paisley, vol. 2, pp. 67–84 (2002)
27. de Castro, L.N., Von Zuben, F.J.: Artificial Immune Systems: Part I Basic Theory and Applications, pp. 57–58. Springer, Berlin (1999)
28. Burke, E.K., Kendall, G. (eds.): Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. Springer, Berlin (2005)
29. Middlemiss, M.: Positive and Negative Selection in a Multilayer Artificial Immune System. The Information Science Discussion Paper Series 2006/03, University of Otago (2006)
30. Dasgupta, D.: Immunity-based intrusion detection system: a general framework. In: Proceedings of the 22nd NISSC vol. 1, pp. 147–160 (1999)
31. Liang, G., Li, T., Ni, J., Jiang, Y., Yang, J., Gong, X.: An immunity-based dynamic multilayer intrusion detection system. In Computational Intelligence and Bioinformatics, pp. 641–650. Springer, Berlin (2006)
32. Aziz, A.S.A., Hassanien, A.E., Azar, A.T., Hanafi, S.E.O.: Machine learning techniques for anomalies detection and classification. Advances in Security of Information and Communication Networks, pp. 219–229. Springer, Berlin (2013)
33. Aziz, A.S.A., Hassanien, A.E., Hanafy, S.E.O., Tolba M.F.: Multi-layer hybrid machine learning techniques for anomalies detection and classification approach (2013)
34. A. Aziz, A.S., Salama, M.A., Hassanien, A.E., Hanafy, S.E.O.: Artificial Immune System Inspired Intrusion Detection System Using Genetic Algorithm. Special Issue: Advances in Network Systems Guest Editors: Andrzej Chojnacki vol. 36, pp. 347–357 (2012)
35. Aziz, A.S.A., Azar, A.T., Hassanien, A.E., Hanafi, S.E.O.: Continuous features discretizaion for anomaly intrusion detectors generation. In: WSC17 2012 Online Conference on Soft Computing in Industrial Applications (2012)
36. Khoshgoftaar, T.M., Gao, K., Ibrahim, N.H.: Evaluating indirect and direct classification techniques for network intrusion detection. Intell. Data Anal. **9**(3), 309–326 (2005)
37. Kotsiantis, S.B.: Supervised machine learning: a review of classification techniques. Informatica **31**, 249–268 (2007)
38. Krugel, C., Toth, T.: Using decision trees to improve signature-based intrusion detection. In: Recent Advances in Intrusion Detection, pp. 173–191. Springer, Berlin (2003)
39. Mitchell, T.M.: Machine Learning. McGraw Hill, Burr Ridge (1997)
40. NSL-KDD Intrusion Detection data set, http://iscx.ca/NSL-KDD/ March 2009
41. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani A.A.: A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications (2009)
42. KDD Cup’99 Intrusion Detection data set, Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html Oct 2007

Chapter 10
An Improved Key Management Scheme with High Security in Wireless Sensor Networks

D. Satish kumar, N. Nagarajan and Ahmad Taher Azar

Abstract Security becomes extremely important, when wireless sensor networks are deployed in a hostile environment. In order to provide security, wireless communication should be authenticated and encrypted. Key management is the main problem in wireless sensor networks when concentrated on security. Any key management scheme proposed should have authenticity, integrity, confidentiality, flexibility and scalability. The key management scheme should be scalable to increase in sensor nodes substantially and also its dynamic nature. Asymmetric key management strategies are not suitable for wireless sensor networks as it operate on limited battery life. In the proposed system, key management is provided for privacy and simultaneously validated for security measures. System performance improves in the improved key management scheme by positioning the new node and forming the head for multi-cluster to replace the failed relay nodes. The private key, the multicluster key, the primary key, and the structure key are used to encrypt every message passed within the improved key management scheme. The improved key management scheme acquires results on 4–5 % improved security level with lesser execution time and communication energy consumption. A variety of numerical parameters are computed using ns2 simulator on existing key management schemes. The improved key management scheme is highly realistic because it is intended to incorporate routing layer and security protocol without sacrificing energy.

D. Satish kumar
Nehru Institute of Technology, Coimbatore, India
e-mail: satishcoimbatore@yahoo.co.in

N. Nagarajan
Coimbatore Institute of Engineering and Technology, Coimbatore, India
e-mail: swekalnag@gmail.com

A. T. Azar (✉)
Faculty of Computers and Information, Benha University, Benha, Egypt
e-mail: ahmad.azar@fci.bu.edu.eg

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_10, © Springer-Verlag Berlin Heidelberg 2014

10.1 Introduction

More recently, a pattern shift occurred from traditional macro sensors to the micro-sensors used in Wireless Sensor Relay Networks. A Wireless relay network is comprised of wireless sensor modules, called nodes [1–3]. Each relay node is made up of a few key components such as a micro-sensor to decide the preferred event; a low-cost application-specific microprocessor; memory to store information; a battery; and a transceiver for communication between the node and the rest of the network.

Due to the nature of wireless communication, data is available in the air for any third party to acquire. The security feature along with the ad-hoc nature, irregular connectivity, and resource limitations of relay network result in a number of design challenges [4–6]. For example, the accessibility of data to third parties causes numerous disasters in many military or homeland security applications. Therefore, it is critical to provide privacy and authentication while preventing data information from being compromised. Traditionally, security is provided through public key-based protocols. However, these network topology controls engage great memory bandwidth and composite mechanism.

The incomplete resources of wireless relay network create a category of security schemes unsuitable for implementation. Thus, security provides the unique features and resource limitations of relay network. Currently, very incomplete work has been done on relay network security. The original work on securing relay network has an end-to-end transmission, which requires time synchronization among sensors. A significant improvement for achieving broadcast validation of any messages sent from the base-station (BS).

One of the common drawbacks of those sensor network security schemes is that they do not combine security with energy-efficient hierarchical routing architectures. The wireless sensors only want to account data to the nearby sensors. It caused much overhead if construct secure links between any two relay nodes. Classically, to reduce routing overhead a relay network is able to self-organize itself a multi-cluster architecture after sensor deployment. A multi-cluster includes a group of neighboring nodes where one of the group nodes is selected as Cluster Head (CH). The multi-cluster use parameters such as sensor energy level, mobility, position to form multiple clusters and determine CH. Data is combined by the CH that removes duplicated or redundant information. The aggregations are also be realize by having nodes closer to the CH process the data coming from nodes further than away through eavesdrop.

NTA process set channel ID for mobile stations, and time taken to accept a new mobile station are mainly focused [7]. The mobile stations, which are in the coverage area of base station, are given initial preference, and those outside coverage area is allocating channel ID through relay stations. Existing work overlooks the idea that security scheme should be effortlessly incorporated with the special characteristics of relay network architecture, especially routing protocols. In particular, most of the existing relay network security strategies focus only on key management and security algorithms. For example, all existing keys pre distribution schemes try to establish pair wise keys between each pair of nodes. However, most sensors do not necessitate setting up a direct protected channel with sensors multiple hops. Since, relay network use hop-to-hop communication techniques to achieve long distance transmission.

Most of the traditional sensor network security schemes presently center on end-to-end security issues and ignore relay network topology control details. They do not believe the low energy routing structural design and merely presume the entire network uses tree or flat-based topology. It is necessary and beneficial to consider cluster-based communication architecture to reduce key management overhead in a secure relay network.

Proposed INTK incorporate the cluster-based routing architecture and key management in the relay node for enhanced reliability and security. INTK scheme achieves security in relay network topology control with routing procedure. Performance results show that the proposed multi clustering based intensity keying/re-keying scheme significantly saves energy. It is a dynamic, distributed protocol where security provides independent of central control. An additional significant feature of INTK scheme is that it has a robust broadcast, and it recovers even the multiple key losses.

The chapter is organized as follows. Section 10.2 provide the related work. Section 10.3 describes the Incorporated Network Topological control and Key management with a brief algorithm. Section 10.4 demonstrates the ns2 simulator environment. Section 10.5 details the performance with resultant table and graph. Section 10.6 provides conclusions about proposed work.

10.2 Related Work

Wireless Sensor network (WSN) routing topology inference is an incomplete path measurement set in a collection cycle due to packet loss in real-world environments [8–12]. It does not handle large-scale of WSN consisting of thousands of nodes [13]. The current WSN link loss and delay inference schemes are not extended to deal with realistic WSN under dynamic routing. Multi-channel interface network coding that is based on the combination of a new concept of coded-overhearing and coding aware channel assignment but fails in arbitrary network topologies and new routing algorithms matched to the proposed network-coding scheme [14]. It does not take into account traffic patterns, directions and distributed channel assignments.

The problem of deploying the minimum sensors on grid points construct a connected wireless sensor network able to fully cover critical square grids, termed CRITICALSQUARE-GRID COVERAGE [15]. Polynomial-time distributed algorithm for maximizing the lifetime of the network does not require knowledge of the location's nodes or directional information, which is difficult to obtain in sensor networks. It employs disk sensing and communication models [16].

Distributed energy optimization method for objective tracking application on the sensor deployment considers only non-practical environment and are not a more energy-efficient communication framework [17]. Routing strategy tries to account for link stability and for minimum drain rate energy consumption in order to verify the correctness of the bio objective optimization named Link-stAbility and Energy aware Routing protocols (LAER) but fails in providing security [18].

Localized Power Efficient Data Aggregation Protocols (L-PEDAP) are based on topologies, such as LMST and RNG, that approximate minimum spanning tree and competently computed using only position or distance information of one-hop neighbors [19]. The actual routing tree is constructing over these topologies. Each topology and parent selection strategy are compared in the study of [19] and it was concluded that the best among them is the shortest path strategy over LMST structure.

Yuea et al. [20] proposed an energy efficient and balanced cluster-based data aggregation algorithm (EEBCDA). The results of simulation show that EEBCDA can remarkably enhance energy efficiency, balance energy dissipation and prolong network lifetime.

Ant colony algorithms called DAACA for data aggregation consists of three phases namely initialization, packet's transmissions and operations on pheromones. In the transmission phase, each node estimates the outstanding energy and the quantity of pheromones of neighbor nodes to calculate the probabilities for vigorously selecting the next hop. After certain rounds of transmissions, the pheromone's adjustments are performed, which take the compensation of both global and local merits for evaporating or depositing pheromones [21].

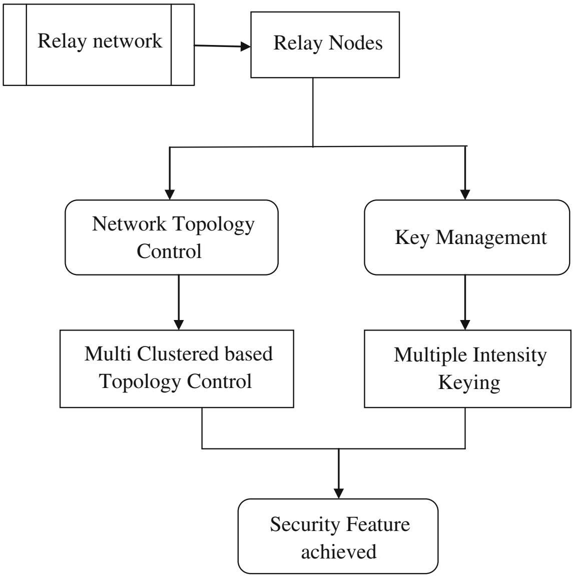
Chao and Hsiao [22] proposed a structure-free and energy-balanced data aggregation protocol, SFEB. SFEB features both efficient data gathering and balanced energy consumption, which results from its two-phase aggregation process and the dynamic aggregator selection mechanism. The simulation and real system implementation results verify the superiority of the proposed mechanism.

Wireless sensor networks are usually deployed in remote and hostile environments to transmit sensitive information, sensor nodes are prone to node compromise attacks and security issues such as data privacy and reliability but data aggregation does not take place in dynamic environments and does not perform a source coding based secure data aggregation [23]. Hua and Yum [24] adopt a model to integrate data aggregation with the underlying routing scheme and present a smoothing approximation function for the optimization problem. The distributed algorithm can converge to the optimal value efficiently under all network configurations.

Sicari et al. [25] presented an approach for dynamic secure end-to-end data aggregation with privacy function, named DyDAP. It has been designed starting from a UML model that encompasses the most important building blocks of a privacy-aware WSN, including aggregation policies. The results showed that DyDAP avoids network congestion and therefore improves WSN estimation accuracy while, at the same time, guaranteeing anonymity and data integrity.

Transmission power between nodes cast a QoS metrics as multi objective problem and operates with any Medium Access control (MAC) protocol. It employs an acknowledgment (ACK) mechanism [26]. Wireless Body Sensor Network constraints, which show the useless of WSN security mechanisms, necessitate an original solution to obtain into explanation the performance of the key biometrics so that their

Fig. 10.1 Processing Flow of INTK Scheme


Relay network
Relay Nodes
Network Topology Control
Key Management
Multi Clustered based Topology Control
Multiple Intensity Keying
Security Feature achieved

use gives a very important impact for the level of security [27]. The implementation intervenes finally to test the efficiency of mechanism in the wireless body sensor networks.

Secure neighbor discovery protocol, SEDINE, for static multihop wireless networks fails in establishing false routes by possibly launching a wormhole attack. It does not consider attacks that prevent two neighboring nodes from becoming neighbors [28]. To overcome all the above issues developed a scheme to incorporate the network topology control and key management in wireless relay network.

10.3 INTK Scheme

Proposed Incorporated Network Topological control and Key management (INTK) is designed to be robust and secure. These changes in network topology control simplify the routing of messages within the relay network. In Fig. 10.1 below, intend to describe security features in the relay network.

The relay network refers to a broad class of network topology commonly used in wireless sensor networks, where the source and destination are interconnected by means of some interrelated nodes. In such a relay network, the source and destination cannot communicate to each other directly because the distance between the source and destination is greater than the transmission range of both of them, hence the need for intermediate nodes to relay in INTK Scheme. Proposed System in the relay network as shown in Fig. 10.1 incorporates the network topology control and the managing the key in a single system to further improve the security feature.

As relay nodes fail from lack of energy, system performance improves in the INTK by positioning the new node and forming the head for multi-cluster to replace the failed relay nodes. Node loss and the exploitation of extra nodes result in a constantly changing network topology control. The private key, the multi-cluster key, the primary key, and the structure key are used to encrypt every message passed within the INTK scheme. The following expressions depict the key generation principle in INTK scheme,

$$I_p = func_{uni\,base}\,(PANC(y)) \tag{10.1}$$

The above equation generates the private key. All keys within the INTK scheme are computed through uni base hash functions and a Pseudo Arbitrary Number Creator (PANC). PANC is used to generate a number for the desired key length. A uni base hash function is then applied to this number in order to generate the key. Expression for generating the multi-cluster key is shown in Eq. (10.2),

$$I_{MC} = func_{uni\,base}\,(PANC(y)) \tag{10.2}$$

Expression for generating the refreshed multi-cluster key is shown in Eq. (10.3)

$$I_{refreshed\,MC} = func_{uni\,base}\,\left(PANC\left(I_{present}I_{MC}\right)\right) \tag{10.3}$$

Expression for generating structure keys is,

$$I_{Structure} = func_{uni\,base}\,(PANC(y)) \tag{10.4}$$

Expression for generating refreshed Structure keys is,

$$I_{refreshed\,Structure} = func_{uni\,base}\,\left(PANC\left(I_{present}I_{Structure}\right)\right) \tag{10.5}$$

In the case of refreshing a present key, the present key is used in place of the generated number, and the hash function is applied to the present key to generate a new key in INTK scheme. The personal keys are generated prior to deployment and are stored within the memory. Multiple keys are important in the INTK Scheme because they make a compromise exceptionally difficult and provide two levels of validation. Not only must a compromised node have knowledge of three different keys (i.e. private key, multi-cluster key and structure key), but also know exactly when to use them. Also, because of different keys and message sizes, it is extremely difficult to decipher the different portions of the message.

INTK scheme uses two keys to provide privacy and validation at every step in the network. All routing information of any message passed within is encrypted with the structure key while the data portion is encrypted by the structure key and the multi-cluster key (MC), or the private key of the relay node. Therefore, if a node is lacking of the structure key, no information is sent or received. This provides first-level validation of the relay node. The data portion of all messages within is encrypted

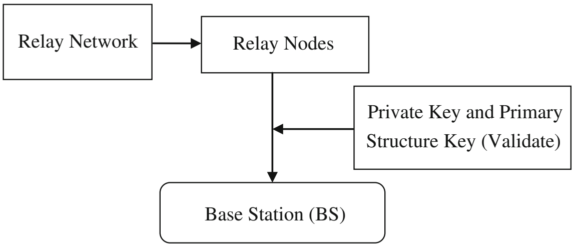


Fig. 10.2 Validation process using structure key

with different kinds of keys. Correspondingly, a relay node needs to have information of network topology control and understand network functionality in order to use the correct key for decrypting the information portion. It finally, provides second level validation of the relay node.

10.3.1 Initial INTK Setup Process

The initial structure setup of INTK consists of three phases namely the validation phase, multi cluster organization phase, and network topology route control phase. Each of these structure setup phases builds upon the preceding phase and completed before the following phase commenced. A detailed description of each phase is described in the following sections.

10.3.1.1 Validation Phase

Every relay node participates in the INTK Scheme must be validated. A node is validated by having the latest structure key. In order to get the latest structure key, a node sends a request to the base Station (BS) in the relay network, encrypted with that node's private key and the primary key. The BS knows that the node is authentic because the relay node has the personal key associated with its node ID. The BS replies to the node with the latest structure key, encrypted by the primary key and the private key of the requesting node. A diagram depicting an overview of the validation phase is shown in Fig. 10.2.

The node receives and decrypts the system key, and attempts to join a multi-cluster. During the primary structure setup phase, some nodes are selected as Cluster Head (CH) based on the load-balanced. When initially requesting the latest structure key, CH validates themselves in the same fashion as relay nodes through their private key and the primary key. A CH needs to request a multi-cluster key which used it to securely organize a cluster among neighboring sensors. Cluster Head (CH) received both the latest structure key and a multi-cluster key from the BS in reply to their validation request.

After this initial validation, and for the rest of its lifetime, a relay node continuously receives and decrypts the latest structure key. All relay nodes in an INTK scheme are continuously and periodically validated and achieved through the periodic refreshing of the structure key.

$$func_{uni\,base}\left(PANC\left(I_{present}I_{Structure}\right)\right) = I_{refreshed\,Structure} \tag{10.6}$$

The structure key is broadcast three times in order to reduce the effects of wireless sensor errors and the resulting chances of a relay node failing to validate. Any message whose topology routing is not encrypted with the most recent structure key is ignored and not ACKed in INTK scheme. Therefore, a relay node will be totally ignored by the rest of the INTK scheme if the node does not have the latest structure key. It guarantees that a node tampered by the enemy does not have the latest system key when it attempts to rejoin the relay network. The relay node cannot attempt to rejoin the INTK scheme, because each private key used once to acquire the latest structure key.

10.3.1.2 Multi-Cluster Organization Phase

INTK scheme multi-cluster organization phase sets up the network topology control in the course of creating multiple clusters, which incorporate the packets with enhanced security. Once a Cluster Head (CH) is selected and validated it broadcasts information, encrypted with the latest structure key. Information contains the cluster ID number, and the multi-cluster key. Relay Nodes pay attention to this information and record their Received Signal Strength (RSS). The strongest recorded RSS is linked with the adjacent CH, and the relay node sends a multi-cluster joined message to this CH encrypted with the multi-cluster key. The multi-cluster key is received through the cluster information.

As multi-cluster joining requests are received, the CH adds those relay nodes to its multi-cluster member registry. The CH keeps a counter who is reset whenever a relay node joins its cluster in INTK scheme. When the counter expires, the CH sends a multi-cluster organization report to the BS, encrypted with the structure key. The multi-cluster organization report is complete with the multi-cluster ID, the CH ID, the present multi-cluster key, and the multi-cluster member registry.

The multi-cluster member registry is a list of all relay nodes within a given cluster. The BS keeps track of network topology control throughout the multi-cluster member registry chart in each CH. A change in the network topology control of a multi-cluster; an innovative multi-cluster organization account is send to the BS. This knowledge is used in the event of CH concession in order to re-organize the multi-cluster.

10.3.1.3 Network Topology Route Control Phase

The phase of network topology route control is responsible for setting up the communication routes for inter multi-cluster and intra multi-cluster routing. After multi-clusters are organized in INTK scheme, the CH sends its principal multi-cluster organization description and locates a topology route to the BS. In precise, if the BS is not one of its neighbors, the CH which transmits a Route Request (RREQ) message. A neighbor is defined to be a relay node that's RSS is above a definite threshold, and every hop of route must take place between neighbors.

All routing messages are transmitted to neighbor CH in a multi cluster way, and CH remains with the series number of each message. A CH receives a request message; it checks to see if the request destination is one of its neighbors in the relay network. If the current recipient is NOT a neighbor of the requested destination, then it forwards the RREQ to all of its neighbors through a broadcast encrypted with the structure key. It appends its own relay node ID to the topology route contained within the request before forwarding the message. The receiving relay node is the destination of the RREQ, and then it creates a Route Reply (RREP) message containing the whole topology route from source to destination.

Only the first received RREQ is replied to an all following RREQ messages with the same series number are ignored. In the event that the request is planned for one of the neighbors of the current recipient, the modified RREQ is promoted only to the destination. The topology route control process in INTK scheme is used for both CH to BS routing, and for node to CH routing within a multi-cluster. Similarly, request and reply messages are encrypted with the structure key allowing the relay nodes and CH to secure on network topology routing information. It allows them to fill in their own routing chart without sending additional RREQ messages.

10.3.2 INTK Structure Operation Algorithm

There are three basic steps such as validation phase, multi cluster organization phase, and network topology route control phase. The procedure of INTK algorithm is as follows.

```
Input: Get sample no of relay nodes to be optimized in WSN
Step 1: Initialize, the value of i= transmit structure key, multi-cluster, $I_p, I_{MC}, I_{refreshed\ MC}, I_{refreshed\ Structure}$, BS, CH, Cluster ID number, and MC key.
    // Validation loop
    Step 2: For (i<=3)
    Step 3: Received packet validated through latest Structure Key.
    Step 4: Periodic refreshment of structure Key

        $I_{refreshed\ Structure} = func_{uni\ base}\left(PANC\left(I_{present} I_{Structure}\right)\right)$
```

```
Step 5: If (latest structure key)
Step 6: Relay node validated.
Step 7: Else, terminates the loop.
//Multi-Cluster loop
Step 8: Next generation of relay node selects the CH.
Step 9: Encrypts $I_{Structure}$, adds to MC member registry.
Step 10: Tracks network topology, BS re-organize the MC $I_{refreshed\,MC}$
for security
//Topology Route Control loop
Step 11: Establish Topology Route
Step 12: RREQ send to neighbors with $I_{Structure}$ key
Step 13: CH sends the $I_{refreshed\,MC}$ with series number
Step 14: Locate the topology route to BS, RREP sent secure (i.e.) privacy
information.
Output: Security enhanced in relay node with minimal energy consumption in relay network.
```

The above algorithm during the initial structure setup phase, INTK scheme achieves authentication through each relay node using its private key and primary key. Once the initial system setup phase has completed, INTK validates the entire structure by periodically refreshing the structure key. In INTK, the structure key is broadcast three times in rapid succession to the relay network at the beginning stage. A CH private key must be used in order to get the structure key for the first time. The global validation is achieved by periodically refreshing the structure key.

A CH is compromised and detected; a removal message is broadcast to the system. The BS generates a re-organization message $I_{refreshed\,MC}$, in response to this removal message, and sends it to the corresponding relay nodes as shown. Similar to the above topology control procedure, INTK achieves privacy through the use of keys and encryption scheme. In situations when INTK uses two different keys to encrypt a message, the relay node needs to have knowledge of both types of keys and the order to use them, which enhances the privacy and node identity.

10.3.3 INTK Security Feature

INTK is innovative in its use of multiple keys for encrypt the message. It makes the node compromise and key compromise extremely difficult. To intercept a message, not only must the right keys be known, but it must also be known in which order to apply them to a given message. The comprehensive encryption and dual key's scheme converse the system responds to compromised nodes in INTK Scheme.

The INTK scheme in rely network utilizes multiple keys to achieve security, validation, and privacy. Due to the limitations of sensor nodes, all keys within INTK are symmetric. The symmetric keys are simpler, smaller, and computationally less rigorous than asymmetric keys. INTK uses three main keys namely the structure key, the private key and the multi-cluster key. The system key is used for global validation purposes and is periodically refreshed. The private key is used for initial node validation during the structure setup phase. The multi cluster key is used for security and used to encrypt the information portions of all messages exchanged within the cluster on the relay network.

Encryption in INTK Scheme is achieved through uni-base hash functions, which have security features such as computational simplicity, low memory and resource overhead. The choice of encryption in INTK scheme is dependent on the application and the network environment. The uni-base transformation hash function implemented in INTK utilizes a numerical value that allows it to change base. Then, a decimal and octal form of key value is transformed into a hexadecimal key to form hash value.

10.4 Simulation Environment

Simulations are used to analyze and evaluate the performance of the INTK scheme. It uses the network simulator named NS2 to simulate the method. A comparative study between the behaviors of the relay network is examined. The well known NS2 simulation tool is used. It is an isolated event network simulator for networking research. NS2 provides a complete development environment for performance evaluation of communication networks and distributed systems. It provides a substantial support to simulate the group of protocols.

To verify the incorporated network topology and key management algorithm, the results are compared with existing Network Topology Acquisition (NTA) processes for non transparent mode relay networks. RWM use and standard of the total number of mail sent or received per node as calculated of the communication requirements, and measure resiliency by counting the number of times must run the protocol in order to detect a single node replication.

The wireless relay nodes were arranged randomly in the field 500 m × 500 m in the sensor fields. The time for transmitting such a packet is considered, and relay nodes were also arranged. The relay nodes perform the simulation with 600 simulation seconds, fixed a pause time of 30 simulation seconds and a minimum moving speed of 1 m/s of each node.

In the Random Way Point (RWM) model, each relay node shifted to an erratically chosen location with a arbitrarily selected speed between a predefined smallest amount and highest speed. It assumes the normal unit disc bidirectional communication replica and adjusts the message range, so that each relay node will have roughly 60 neighbors on average. The purpose of the study investigates the behavior of communication energy, security level, and execution time.

Table 10.1 Node density versus communication energy

Node density	Communication energy (J)	
(nodes/10 m^2)	NTA process	INTK scheme
5	5.9	4.1
10	10.1	8.7
15	11.3	9.5
20	20.5	16.2
25	25.3	21.8
30	28.1	24.5
35	31.2	25.8

10.5 Results and Discussion

Incorporated Network Topological control and Key management (INTK) scheme is compared with the existing Network Topology Acquisition (NTA) processes for non transparent mode relay networks in measuring the communication energy consumption, security level and execution time. Communication energy consumption is defined as the amount of energy consumed to transfer the information from source relay node to destination relay node in the wireless relay network. It is measured in terms of joules (J).

$$\text{Communication Energy Consumption} = \text{Ts}^2$$

where, 'T' = Total number of information and 's' represents the speed of transferring effect of data in relay node.

The security is defined as the amount of security given for the fulfillment of an obligation (i.e.) the information encrypted and decrypted using INTK scheme in the wireless relay network. It is measured in terms of percentage (%).

Execution time is when a series is running. That is, when start a series running, it is the runtime for that series. The execution time is defined as the time taken to transfer the data from the source relay node to destination relay node in the relay network. It is measured in terms of milliseconds (ms).

$$Execution\,Time = RREQtime - RREPtime \tag{10.7}$$

Table 10.1 describes the communication energy consumption based on the node density and it is illustrated in Fig. 10.3. The data portion of all messages within is encrypted with different multi-cluster keys. Correspondingly, a relay node needs to have information of network topology control and understand network functionality in order to use the correct key for decrypting the information portion by reducing the communication energy consumption. The INTK scheme energy used to communicate are 5–10 % lesser when compared with the NTA process.

The security level of NTA process and INTK schemes are examined, and the output obtained in terms of percentage (%) as shown in Table 10.2 and Fig. 10.4.

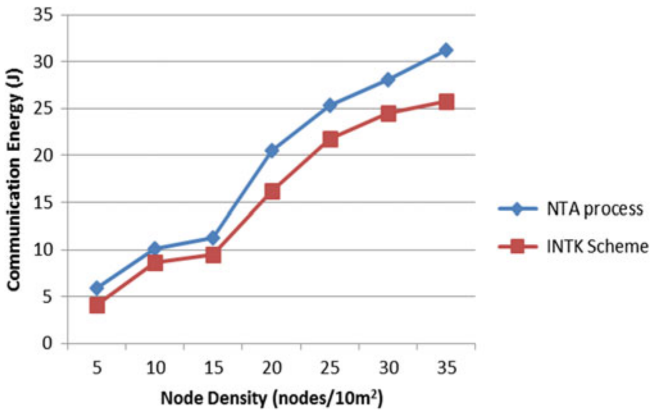


Fig. 10.3 Node density versus communication energy

Table 10.2 Technique versus security

Technique	Security level (%)
INTK scheme	88.8
NTA process	84.3

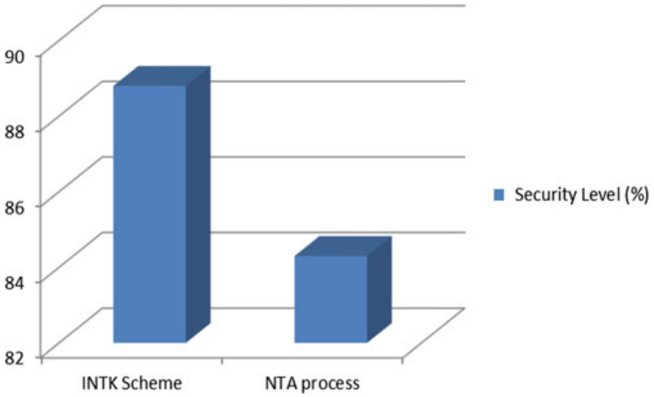


Fig. 10.4 Technique versus security

INTK Scheme achieved security level of 88 % while NTA process achieved 84.3 % security level. The results demonstrated that the multi cluster key used for security and is used to encrypt the information portions of all messages exchanged within the cluster on the relay network. The encrypted message is decrypted on other sides, which will definitely improve the security of INTK scheme to 4.5 % when compared with the NTA process.

Table 10.3 Node density versus communication energy

No. of nodes	Execution time (ms)	
	NTA process	INTK scheme
10	376	360
20	377	362
30	397	383
40	481	460
50	482	469
60	483	471

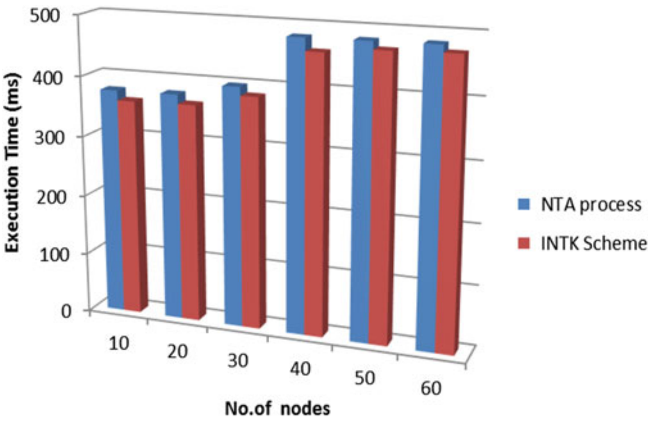


Fig. 10.5 No. of nodes versus execution time

Table 10.3 describes the execution time based on the average nodes involved in the processing and is illustrated graphically in Fig. 10.5. The INTK scheme is lesser than NTA scheme in execution time by 2–5 % because Cluster Head (CH) is selected and validated. The broadcast's information, encrypted with the latest structure key to reduce execution time in the proposed system when compared with the existing NTA process. Therefore, a relay node will be totally ignored by the rest of the INTK scheme if the node does not have the latest structure key.

Finally, relay nodes are tiny sensors; the security protocols to have low energy consumption for communication. Even a single relay node misses the latest structure key, it no longer function in the structure since its topology routing header cannot be decrypted. The relay node is ignored and eventually removed by the structure, if not validated, and unable to reenter the INTK relay network scheme. Relay nodes that miss the structure key due to an enemy trying to infiltrate the relay network by physically compromising the relay node will be kept out of the NTK relay network scheme in the same manner.

10.6 Conclusion

Present solutions to the security issue in the relay network were developed with validation and privacy in mind using the Incorporated Network Topological control and Key management scheme. Validation for security measure is offered in relay nodes and simultaneously, Key management is provided for privacy. INTK scheme encompasses the incorporation of security and routing, active security, robust re-keying, low complexity and the multiple intensities of encrypt features in relay networks. It is far from optimal because network topology routing and securities are closely associated. Multi cluster based topology control through an intensity keying consumes lesser communication energy due to its multi cluster key executive. INTK is highly realistic because it is intended to incorporate routing layer and security protocol without sacrificing energy. A variety of numerical parameters are computed using ns2 simulator on INTK Scheme acquires 4.5 % improved security level with lesser execution time and minimal communication energy consumption. Relay network provides the effective routing and security solution.

References

1. Calinescu, G., Tongngam, S.: Relay Nodes in Wireless Sensor Networks. Wirel. Algorithms Syst. Appl. Lect. Notes Comput. Sci. **5258**, 286–297 (2008). doi:10.1007/978-3-540-88582-5_28
2. Doss, R., Schott, W.: Cooperative relaying in wireless sensor networks. In: Misra, S.C., Woungang, I., Misra, S. (eds.) Guide to Wireless Sensor Networks Computer Communications and Networks, pp. 159–181. Springer, London (2009). doi:10.1007/978-1-84882-218-4_6
3. Liu, B.H., Lin, Y.X., Wang, W.S., Lien, C.Y.: A modified method for constructing minimum size homogeneous wireless sensor networks with relay nodes to fully cover critical square grids. Genet. Evol. Comput. Adv. Intell. Syst. Comput. **238**, 213–220 (2014). doi:10.1007/978-3-319-01796-9_22
4. Komninos, N., Vergados, D.D., Douligeris, C.: Security for ad hoc networks. In: Stavroulakis, P., Stamp, M. (eds.) Handbook of Information and Communication Security, pp. 421–432. Springer, New York (2010). doi:10.1007/978-3-642-04117-4_22
5. Lou, W., Fang, Y.: A survey of wireless security in mobile ad hoc networks: challenges and available solutions. Ad Hoc Wireless Networking, Network theory and applications, vol. 14, pp. 319–364 (2004)
6. Pervaiz, M.O., Cardei, M., Wu, J.: Routing security in ad hoc wireless. In: Huang, S.C.H., MacCallum, D., Du, D.Z. (eds.) Networks Network Security, pp. 117–142 (2010). doi:10.1007/978-0-387-73821-5_6
7. Kumar, D.S., Nagarajan, N.: Improved network topology acquisition processes in IEEE 802.16j nontransparent mode relay networks. J. Discrete Math. Sci. Crypt. **15**(1), 57–71 (2013)
8. Dargie, W., Poellabauer, C.: Fundamentals of Wireless Sensor Networks: Theory and Practice. Wiley, UK (2010)
9. Ismail, M., Sanavullah, M.Y.: Security topology in wireless sensor networks with routing optimization. In: Fourth International Conference on Wireless Communication and Sensor Networks (WCSN 2008), Allahabad, India, 27–29 Dec 2008, pp. 7–15. doi:10.1109/WCSN.2008.4772673
10. Jing, G., Jia, L., Xie, L., Hu, Q., Liu, S.: Fluctuation control for many-to-one routing in wireless sensor networks. J. China Univ. Posts Telecommun. **19**(6), 35–44 (2012)

11. Üster, H., Lin, H.: Integrated topology control and routing in wireless sensor networks for prolonged network lifetime. Ad Hoc Netw. **9**(5), 835–851 (2011)
12. Xu, D., Gao, J.: Comparison study to hierarchical routing protocols in wireless sensor networks. Procedia Environ. Sci. **10**, Part A, 595–600 (2011)
13. Liang, Y., Liu, R.: Routing topology inference for wireless sensor networks. ACM SIGCOMM Comput. Commun. Rev. **43**(2), 21–28 (2013)
14. Kwon, S.C., Hendessi, F., Fekri, F., Stuber, G.L.: A novel collaboration scheme for multi-channel/interface network coding. IEEE Trans. Wirel. Commun. **10**(1), 188–198 (2011)
15. Ke, W.C., Liu, B.H., Tsai, M.J.: Efficient algorithm for constructing minimum size wireless sensor networks to fully cover critical square grids. IEEE Trans. Wirel. Commun. **10**(4), 1154–1164 (2011)
16. Kasbekar, G.S., Bejerano, Y., Sarkar, S.: Lifetime and coverage guarantees through distributed coordinate-free sensor activation. IEEE/ACM Trans. Networking **19**(2), 470–483 (2011)
17. Wang, X., Ma, J., Wang, S., Bi, D.: Distributed energy optimization for target tracking in wireless sensor networks. IEEE Trans. Mob. Comput. **9**(1), 73–86 (2010)
18. De Rango, F., Guerriero, F., Fazio, P.: Link-stability and energy aware routing protocol in distributed wireless networks. IEEE Trans. Parallel Distrib. Syst. **23**(4), 713–726 (2012)
19. Tan, H.O., Korpeoglu, I., Stojmenovi, I.: Computing localized power efficient data aggregation trees for sensor networks. IEEE Trans. Parallel Distrib. Syst. **22**(3), 489–500 (2011)
20. Yuea, J., Zhang, W., Xiao, W., Tang, D., Tang, J.: Energy efficient and balanced cluster-based data aggregation algorithm for wireless sensor networks. Procedia Eng. **29**(2012): 2009–2015 (2012). http://dx.doi.org/10.1016/j.proeng.2012.01.253
21. Lin, C., Wu, G., Xia, F., Li, M., Yao, L., Pei, Z.: Energy efficient ant colony algorithms for data aggregation in wireless sensor networks. J. Comput. Syst. Sci. (2012)
22. Chao, C.M., Hsiao, T.Y.: Design of structure-free and energy-balanced data aggregation in wireless sensor networks. J. Netw. Comput. Appl. **37**, 229–239 (2014). http://dx.doi.org/10.1016/j.jnca.2013.02.013
23. Ozdemir, S., Xiao, Y.: Secure data aggregation in wireless sensor networks: a comprehensive overview. Comput. Netw. **53**(12), 2022–2037 (2009)
24. Hua, C., Yum, T.P.: Optimal routing and data aggregation for maximizing lifetime of wireless sensor network. IEEE/ACM Trans. Networking **16**(4), 892–903 (2008)
25. Sicari, S., Grieco, L.A., Boggia, G., Coen-Porisini, A.: DyDAP: a dynamic data aggregation scheme for privacy aware wireless sensor networks. J. Syst. Softw. **85**(1), 152–166 (2012)
26. Djenouri, D., Balasingham, I.: Traffic-differentiation-based modular QoS localized routing for wireless sensor networks. IEEE Trans. Mob. Comput. **10**(6), 797–809 (2011)
27. Mesmoudi, S., Feham, M.: BSK-WBSN: biometric symmetric keys to secure wireless body sensors networks. Int. J. Netw. Secur. Appl. **3**(5), 155–166 (2011). doi:10.5121/ijnsa.2011.3512155
28. Hariharan, S., Shroff, N.S., Bagchi, S.: Secure neighbor discovery through overhearing in static multihop wireless networks. Comput. Netw. **55**(6), 1229–1241 (2011)

Chapter 11
Key Pre-distribution Techniques for WSN Security Services

Mohamed Mostafa M. Fouad and Aboul Ella Hassanien

Abstract In recent years, thanks to technology advances in low-power wirelessly networked systems and, we have witnessed the emergence of Wireless Sensor Networks (WSNs) and embedded computing technologies in many fields of our life; which range from military to medical applications and from industry to home appliances. Although most of researchers focus on designing protocols that maximizes both the processing capabilities and energy reserves, many of these protocols pay little attention to securing this WSNs. Nowadays, security goal is vital for ensuring the performance and the acceptance of the wireless sensor networks in many recent applications. This goal is still a challenge on account of the constraint resources of these wireless sensor nodes. This chapter givesan overviewof the desired security services required for the WSNs, their threats model and finally the chapter presents in details different pairwise key distribution security techniques for distribution WSNs.

11.1 Introduction

Wireless sensor networks consist of a large number of low-cost and low-power sensor devices. There are countless number of applications that canbenefit from WSN technology; such as object tracking, surveillance, habitat monitoring, healthcare monitoring, and chemical or biological attacks detection. Usually, sensor nodes

M. M. M. Fouad (✉)
Arab Academy for Science, Technology, and Maritime Transport, Scientific Research Group in Egypt (SRGE), Cairo, Egypt
e-mail: mohamed_mostafa@aast.edu
URL: http://www.egyptscience.net

A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Faculty of Computers and Information, Cairo University, Cairo, Egypt
e-mail: aboitcairo@gmail.com
URL: http://www.egyptscience.net

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_11, © Springer-Verlag Berlin Heidelberg 2014

are deployed randomly, or at pre-determined locations using a preset scheme, in a designated area. They automatically collaborate together to formulate a network through wireless communications. According to their resource limitations that inherited from their low bandwidth, restricted processing capability, and usually nonrenewable batteries, the wireless sensor networks face many unique security challenges. Security solutions available for traditional networks, such as Public-key cryptography, create additional challenge when employed with sensor nodes; since they are not optimized for energy usage. Therefore it is demanding to design a proper security mechanism that attends to these limitations.

11.2 Security Services for the WSNs

Although communication as a service may have standards, processes, procedures, and tools, but this all should be guided and nurturedwith security as a goal. The International Telecommunication Union (ITU) has a recommendation for the X.800 as an international standard for defining the desired security services and mechanism for the Open Systems Interconnection (OSI) [1, 2]. This recommendedstandard was considered as a road map for many researchers and practitioners who aim to develop secure systems.

The X.800 standard scores five objectives as essential security services: *Authentication, Data Confidentiality, Data Integrity, Access Control, and Non-repudiation*. However, *Access Control* and *Non-repudiation* are not part of the services issues any more since one WSN deployment generally falls under one administrative domain [3]. Although *Availability* has not originally been considered as one of the security services in X.800, it pertains to be a desired security services for WSNs.

Actually, another set of secondary security services should be involved within any new security algorithm [4] such as: *Data Freshness, Self-Organization, Time Synchronization and Secure Localization*. While Fig. 11.1 elaborates all these security services, the following subsections provide brief their definitions.

11.2.1 Authentication

Authentication is the process of identifying an individual. In sensor networks it is essential for each sensor node and base stations. They must have the ability to verify that the data received was really sent by a trusted sender or by an attacker. The common types of attacks aim either to modify the data packet or to inject additional packets (also known as forging packets) in the network. Data authentication can be achieved through a message authentication code (MAC) that computed from the shared secret keys among sender and receiver nodes [5]. Although there are many designed protocol applying data authentication mechanisms, it is still an extremely security challenging due to the nodes' computational abilities and their large scale deployment.

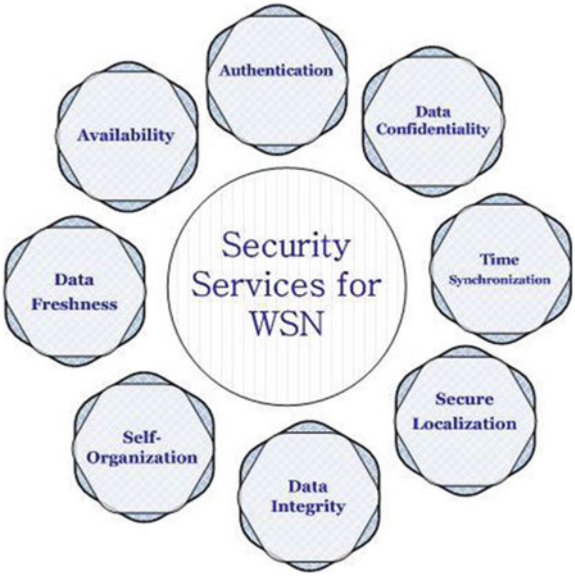


Fig. 11.1 Security services required for wireless sensor networks

11.2.2 Data Confidentiality

In today world, information has value therefore data confidentiality become a must. Confidentiality, as a concept, is a set of rules that protect packets that travel through the network from disclosureto an unauthorized passive attacker. This is the most important issue in network security since a sensor node should not reveal its readings to any of its neighbors unless they are authorized to access these data [6]. A very key component of protecting data confidentiality would be data encryption. If authentication is taken care of correctly, then confidentiality is a relatively simple process.

11.2.3 Data Integrity

Data integrity refers to the completeness, accuracy and consistency of data. In sensor networks data integrity is needed to ensure the reliability of the data and guarantee that the message or packet being delivered has not been tampered with, altered, or changed. Since the sensor nodes are scattered in hazard environments, Data integrity can also be threatened by a number of conditionalthreats, such as heat, dust, and electrical surges. For example, a message could be corrupted due to radio propagation impairment. Nevertheless, there is always a possibility that a malicious node has modified the content of the message. Also, data integrity is linked to authentication since it lies under the threat of unauthorized manipulation of data [7].

11.2.4 Availability

The availability of sensors service is essential to the success of the purpose of the WSN. This requirement ensures that the desired services provided by a WSN, by a part of it, or even by a single sensor node, should be available always even in the presence of an internal or external attacks. The most common attacks against the network availability are the denial of service attacks (DoS), which can be launched at any layer of WSN through radio jamming. The failure of the base station or cluster leader's availability will eventually threaten the entire sensor network [4]. Wireless sensors availability becomes crucial for medical and military applications [8].

11.2.5 Data Freshness

In highly dynamic event monitoring applications, it is not enough to guarantee confidentiality and authentication of a message travels inside a wireless sensor network, but it is important to measure the freshness of this message. Therefore, the data confidentiality and data integrity services should be followed by a data freshness mechanism (no old packet has been replayed). Also, the freshness of messages is important especially when the sensor nodes use security shared keys for their communications. With no data freshness, an adversary can use any replayed old messages to propagate false data within the wireless network. To solve this problem nodes maintain incoming and outgoing time-specific freshness counters to maintain data freshness. The data freshness techniques are classified based on the message ordering into weak and strong freshness [5]. Weak freshness provides only partial message ordering but gives no information associated with the delay and latency of the message. Strong freshness conversely, is useful for time synchronization within the network; it gives complete request response pair and allows the delay estimation.

11.2.6 Self-Organization

Due to the randomly deployments of thousands of sensor nodes in a geographic area, Self-organization and Self-healing are of the important characteristics of a wireless sensor network. That after the deployment, sensors are responsible for autonomously organize themselves to form a network of their own. The main challenge is to determine the best organization that would maximizes the life of the whole network as well as the quality of data transmission.

On the other hand, Self-healing services refer to the ability of sensor nodes to autonomously adapt themselves to changes resulted from different situations (such as nodes' failures, mobility, node inclusion, or response to a detected event). Self-organization and self-healing operations may occur in different ways along the network lifetime for the purpose of extending that lifetime.

11.2.7 Time Synchronization

Time synchronization is important for any distributed system. In particular, wireless sensor networks make extensive use of synchronized time in many contexts; such as nodes' scheduling, routing synchronization, data aggregations, or time-based data gathering. Also, and due to the energy constrains attached with sensor nodes it is essential to conserve the lifetime of the WSN through turning off some nodes (Sleep Mode) and based on predetermined periodic intervals turn them on (Wake up) in order to exchange information. Any security protocol for WSN should consider this time-synchronization required feature. Also the protocol should consider the packet delay between any two peers of nodes [9].

11.2.8 Secure Localization

Secure localization is an important concern for network designers. That usually sensor nodes are deployed in hostile environments thus they need to be secured against common attacks (tampering, substitution, etc.). As another aspect identifying the node's positions within the targeted area is important in order to discover node fault.

Determine nodes location is important for a number of applications for example; environment monitoring [10], healthcare monitoring [8], etc. Another still open research topic is how to secure against the ability of an adversary to provide false location information by replaying false signals [11].

11.3 WSN's Threat Model

Karlof et al. [12] had classified the WSNs threat modelinto three categories; based on attacker access location, the attacker level, or based on the specs of the used device. Figure 11.2 which sketched from [13] plots these types of threats.

- **Attacks based on location; External/Internal attacks**: In external attacks (outsider), the adversary uses external nodes that do not belong to the WSN. It has no access to most cryptographic materials in sensor networks. Whereas the internal attacks (insider) occur when legitimate nodes of a WSN behave in unintended or unauthorized ways. It may have partial key materials and thus it gainsthe trust of other sensor nodes.

  Internal attacks are much harder to be detected. Usually, the majority of external attacks aimto consume the network resources by means of performing passive eavesdropping on data transmissions, injecting false data, jamming the communication signals, or by extending these attacks to perform denial of service (DoS) attacks [5]. Internal attacks may be mounted from either compromised sensor

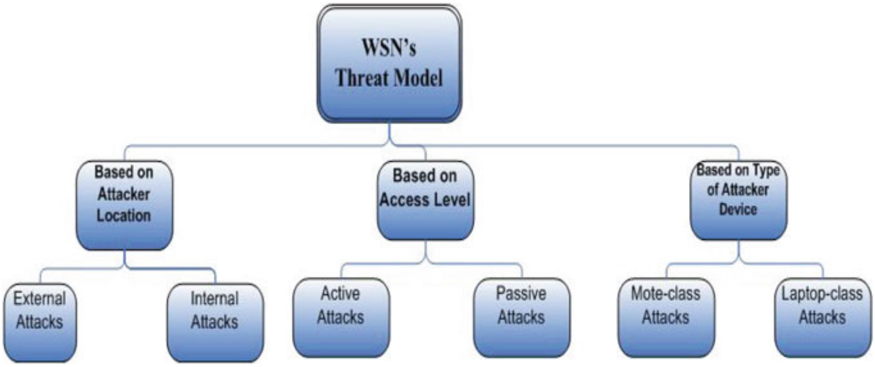


Fig. 11.2 Wireless sensor networks' threat model

nodes running malicious code. Also it can be an adversary who havestolen the cryptography key material, nodes' code, and data from legitimate nodes, and then uses these data against the network.

- **Attacks based on access level; Passive/Active attacks**: Passive attacks are in the nature monitoring and listening of the communication channel by unauthorized attackers. It is usually attacks against the privacy. Even if the packets were encrypted, passive attacks can be used to analyze traffic within the network.
  The active attacks involve modifications of the data stream by creating a false stream in a WSN or attempts to gain unauthorized access to the WSN; such as spoofed, altered and replayed routing information, etc.
- **Attacks based on type of attacker device; Mote-class/Laptop-class attacks**: In mote-class (sensor-class) attacks, an adversary attacks a WSN by using small physical devices with capabilities similar to sensornodes. Therefore it is capability is limited to monitoring communication channel between limited numbers of nodes. In contrast, in laptop-class attacks, an adversary can use more powerful devices (with more battery power, faster CPU, sensitive antenna, powerful radio bandwidth, etc) like laptop, PDA, etc. A single laptop-class attacker may eavesdrop the entire network. It can do more damage such as jamming radio frequencies in surrounding areas.

11.4 Security Attacks on Sensor Networks

Wireless sensor networks address unique security challenges, especially for security-based critical applications. Altogether, they are vulnerable to various types of security attacks due to their deployment strategies. Table 11.1 issues wide variety of common attacks targeting the WSNs and the current suggested countermeasures.

Along with previous defined attacks there are other three types of attacks that usually are launched againstspecific WSNs' protocols; such as topology maintenance

Table 11.1 WSN's security attacks and their countermeasures [14, 15]

Attacks	Description	Countermeasures
Spoofed, altered, or replayed routing information	Create routing loop, attract or repel network traffic from selected nodes, extend or shorten source routes, generate false error messages, causing network partitioning, increasing end-to-end latency, and etc	Egress filtering, authentication, monitoring
Selective forwarding	Either in-path or beneath path by deliberate jamming, allows to control which information is forwarded. A malicious node acts as a black hole; as it is simply dropping certain messages instead of forwarding every message	Egress filtering, authentication, monitoring
Sinkhole attacks	The adversary places a malicious node where it can attract most of the traffic—for example a cluster head, in order to prepare a future selective forwarding	Redundancy checking
Sybil attacks	A single node presents multiple identities, allows us to reduce the effectiveness of fault-tolerant schemes such as distributed storage and multi-path	Authentication, monitoring, redundancy
Wormhole attacks	Tunneling of messages over alternative low-latency links to confuse the routing protocol, thereby creating sinkholes, etc	Authentication, probing
Hello floods	An attacker sends or replays a *HELLO* message with stronger transmission power and pretending that this message is coming from the base station. All nodes will be responding to *HELLO* floods and wasting their energies	Authentication, packet leashes by using geographic and temporal information

protocols [16, 17]: *sleep deprivation attacks*, *snooze attacks*, and *network substitution attacks*. The following points discuss these attacks:

- **Sleep Deprivation Attack**

The capability of network protocol to turn off some sensor nodes (enter a low power sleep mode) is very useful for maximizing network lifetime. The natural mechanism of a typical topology maintenance algorithm is to substitute active nodes with other sleeping nodes. The sleep deprivation attacker, which introduced by the author of [18], tries to induce a node in a specific area to stay awake. This attack reduces the victim node's lifetime through increasing its energy expenditure. In addition it increased energy consumption due to congestion and contention especially in densely deployed networks [19].

- **Snooze Attack**

The snooze attack has a number of scenarios; first scenario is for the cluster-based networks; in which an attacker put itself in sleep mode in order to save more energy, so its chances of becoming a cluster head in the next become higher. In this case attacker will receive the data from its cluster members and not sending any data to the base station [20].

In an alternative scenario, the snooze attacker enforces nodes to remain in their sleeping state. The whole network can be turned off by this type of attacks. in other situation the attacker can reduce the sensing coverage in a selected area. For example, an adversary can selectively turn off nodes that are monitoring an intruder's path through an area in which a sensor field has been deployed for surveillance. Karlof and Wagner [12] had described in their paper the affects of the snooze attack against some protocols such as: GAF [21], SPAN [22], and CEC [23].

- **Network Substitution Attack**

The adversary has the ability to deploy a set of malicious nodes to control the entire network or a portion of it. The attacker substitute legitimate nodes through enforcing them to go into a sleep mode. The maintenance procedure will use malicious nodes in order to maintain the network connectivity. As soon as the protocol has activated any of these nodes, this node gives the control of that portion of the network to the adversary. Once the adversary has the control on that portion of the network, it performs other attacks such as traffic analysis and selective forwarding or complete packet dropping. This type of attacks cannot be easily detected because the adversary still maintain the network connectivity and keep it operating normally.

11.5 Key Distribution Techniques for Distributed WSNs

Wireless Sensor Networks is defined as a large collection of sensor nodes that deployed in open and unattended environments without any physical protection. Therefore the WSN architecture is organized in distributed structure; which means no predefined fixed hierarchy prior the deployment.

These networks are prone to different types of malicious attacks. Providing secure, authenticated, and encrypted communication channels between sensor nodes become important issue. This issue stills a challenge on account of the resources limitations of these sensor nodes. The open problem is how to set up the secret keys between communicating nodes. This problem is known as the key agreement problem [24]. Any new keyscheme will be suggestedas a solution for such problem has to satisfy certain requirements [25]:

- Providing the capability of node-to-node secure communication without any control from the base station.
- The future added nodes can easily perform a secure communication channel with the previously deployed nodes.
- Prohibiting any unauthorized node from establishing any communication with any of the network's nodes. Also the scheme has a duty to be robust against DoS attacks.

- The scheme has to work even without prior knowledgeof which nodes will come into communication range of each other after deployment.
- The scheme has to minimize both the node's computational overhead and the storage requirements.

Although many solution schemes had been proposed, some of them are not applicable due to the constraint resources attached with a sensor node. The sensor node is typically powered by limited lifetime batteries and it has limited computational capabilities. Asymmetric cryptosystems (public key based digital signatures) approaches such as Diffie-Hellman key agreement [26] or RSA [27], which are typically used for securing communication channels in traditional networks, are infeasible on the basis of constraint resources of these sensor nodes as it was mentioned in [28].

On the other hand the key pre-distribution schemes proved that they are recommended key management schemes to be used by WSN. They provide three approaches; (i) probabilistic, (ii) deterministic, or (iii) hybrid, for solving both the pairwise and the group-wise key distribution problem in distributed wireless sensor networks [29].

11.6 Pairwise Key Distributed Schemes

Pairwise key is a fundamental service required for WSN for securing communication channels between sensor nodes. These schemes based on a simple idea that allows each sensor node to randomly select a subset of keys from a pool of keys prior their deployment. There is a chance that each node will share a common key(s) in their subsets with other nodes within a network. The selected key is used for establishing a pairwise key. Then the pairwise key is used either to encrypt messages or to authenticate messages sender and receiver [25, 30, 31].

One of the simplest solutions is to pre-install a single secret master key in all nodes. This master key will be used within the message encryption or messages authentication processes. The advantage of such solution is the minimal memory storage requirementas only one key is stored in the memory. However the capture of one node will compromise the whole network. One of the recommendations is to store this master key in a tamper resistant hardware [32, 33]; however, the employ of such hardware will increases the complexity of the sensor nodes in terms of bothcost and energy consumption.

Another simplest solution is applying afully pairwisescheme, in which every node shares a unique key with every other node within the network. i.e., for a network of $\boldsymbol{n}$ nodes, each node stores $\boldsymbol{n-1}$ keys. As a result the total number of keys used by every node in the network is $\boldsymbol{n(n-1)/2}$. The resilience of this scheme is perfect because a compromised node only reveals $\boldsymbol{n-1}$ link keys (from the total of $\boldsymbol{n(n-1)/2}$ keys). It will not reveal information about other current communications in other parts of the network. However, the amount of storage requirement by each node increases linearly with the size of the network. Thus this scheme is impractical for sensors with

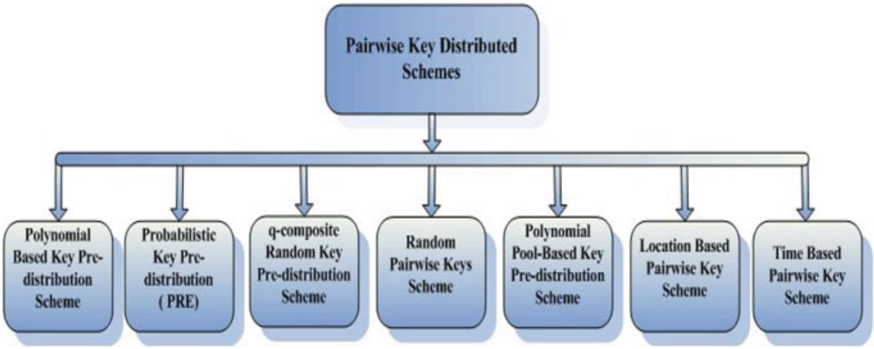

Pairwise Key Distributed Schemes
Polynomial Based Key Pre-distribution Scheme
Probabilistic Key Pre-distribution (PRE)
q-composite Random Key Pre-distribution Scheme
Random Pairwise Keys Scheme
Polynomial Pool-Based Key Pre-distribution Scheme
Location Based Pairwise Key Scheme
Time Based Pairwise Key Scheme

Fig. 11.3 Pairwise key schemes

an extremely limited amount of memory. Furthermore this solution influences the scalability feature of a network; in which adding new nodes to a pre-existing sensor network will be difficult, because the existing nodes do not have the new nodes' keys.

While Fig. 11.3 exemplifies the well-known schemes proposed for establishing pairwise keys between sensors, the following subsections briefly describe their algorithms.

11.6.1 Polynomial Based Key Pre-distribution Scheme

Blundo et al. [34] had proposed a key pre-distribution scheme particularly for group key pre-distribution. For establishing the pairwise key the setup server randomly generates a bivariate ***t***-degree polynomial ***f***(***x, y***) (see Eq. 11.1, where ***i***, and ***j*** in the equation are nodes' IDs) over a finite field ***Fq***, where ***q*** is a prime number larger enough to accommodate a cryptographic key under the assumption of the property of ***f***(***x, y***) = ***f***(***y, x***). In another word, the setup server is computing for each node ***i*** polynomial share of ***f***(***x, y***), then node ***i*** storesa ***t***-degree polynomial of ***f***(***i, y***). The storage space needed per each node for saving its assigned polynomial keys is (***t*+*1*** **log** ***q***) [35]. For establishing a key between two nodes, both nodes need to evaluate the polynomial at the ID of the other node as it is illustrated in Fig. 11.4.

$$f(x, y) = \sum_{i,j=0}^{t} a_{ij}\, x^i\, y^j \qquad (11.1)$$

The scheme offers a non-communication overhead. It is also proved that it is unconditionally secured scheme for up to ***t*** compromised nodes. The main disadvantage of this scheme is the linearly increasing of the memory storage space needed by each node ***i*** for storing a ***t***-degree polynomial ***f***(***i, y***) with the increasing of the network's size [35].

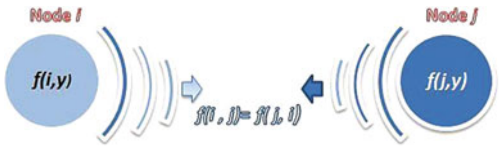


Fig. 11.4 Polynomial key evaluation process to find a common security shared key [41]

Blundokey management scheme is used as the initial trust relationship for secure communications between the Wireless Body Area Networks (BAN) devices; a new E-healthcare enabling technology [8], before they are actually deployed as it was introduced in [36].

11.6.2 Probabilistic Key Pre-distribution (PRE)

Nearly all of the key management protocols for WSNs are probabilistic and distributed schemes. The connectivity of probabilistic key distribution scheme can be modeled using random graph theory [37, 38]. A random graph $\boldsymbol{G(n; c)}$ is a graph of $\boldsymbol{n}$ nodes and the probability that a link (or an edge) exists between any two nodes (or vertices) is $\boldsymbol{c}$. When $\boldsymbol{c = 1}$, the graph is fully connected (there exists an edge between all pairs of vertices). When $c = 0$, there is no edge exist between any two nodes. The scheme exploits this property by setting c larger than a certain value lies between 0, and 1, so that the entire network is almost connected. Eschenauer and Gligor [39] have presented the probabilistic key pre-distribution scheme, in where each sensor node receives a random subset of keys ($\boldsymbol{s} \approx$ key ring) from a large pool of keys $\boldsymbol{k}$ prior deployment. These key rings are selected based on a predefined probabilistic value $\boldsymbol{p}$ of keys sharing amongnodes (also known as Overlapping Probability). To agree on a key for communication, two nodes find one common key within their subsets and use that key as their shared security key.

Eschenauer and Gligor [39] formulated Eq. 11.2 to calculate the node degree $\boldsymbol{d}$ (the expected number of secure links a node can establish during the sharedkey discovery phase) in terms of the network size $\boldsymbol{n}$:

$$d = \left(\frac{n-1}{n}\right)(\ln(n) - \ln(\ln(c))) \tag{11.2}$$

Let $\boldsymbol{s}$ be expected number of keys within a node's key ring. For the value of $\boldsymbol{d}$ required for a network to be connected, the calculated probability for required key sharing between two nodes ($\boldsymbol{p}$) can be calculated by Eq. 11.3:

$$p = \frac{d}{s} \tag{11.3}$$

An operator can adjust the key distribution parameters (i.e., size of key pool and size of keys stored at each node) that satisfy the value of required $\boldsymbol{p}$.

11.6.3 Q-Composite Random Key Pre-distribution Scheme

Chan et al. [25] improve the security of the previous scheme by proposing a q-composite random key pre-distribution scheme. Their scheme allows two sensor nodes to establish a secure link in between, if they share at least q-common pre-distribution keys ($q \geq 1$) in their subsets. As the increase in the amount of keys overlap between two sensor nodes, the network resilience against node capturing attacks is increased.

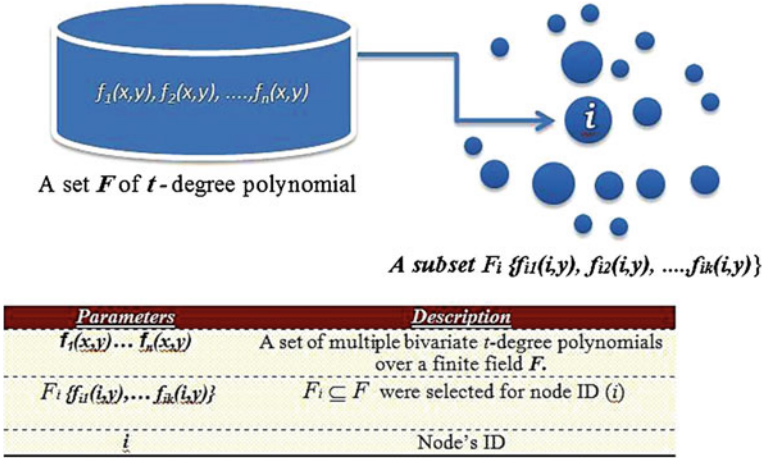
However, still this scheme has a number of limitations. Since, each node in the network selects a certain number of keys from the key pool, and the necessity to reduce the size of the key pool in order to satisfy a given probability $\boldsymbol{p}$ (two nodes sharing sufficient keys) for establishing a secure link. This allows the attacker to gain a larger sample of keys by breaking fewer nodes, thus he can gain entry into the network. Moreover memory overhead becomes a concern especially with large scale networks.

11.6.4 Random Pairwise Keys Scheme

However, the pairwise key establishment problem is still not fully solved. For the basic probabilistic and the q-composite key pre-distribution schemes, as the number of compromised nodes increases, the fraction of affected pairwise keys increases quickly. As a result, a small number of compromised nodes may affect a large fraction of pairwise keys. Therefore, Chan et al. [25] also had proposed a random pairwise keys scheme to overcome the node capture attacks. The setup server in his scheme randomly selects a pair of nodes and assigns a unique random key between them. One of the advantages of this scheme is that none of the directly shared keys between non-compromised nodes is revealed no matter how many sensor nodes are compromised. Another advantage is the node-to-node authentication mechanism that enables nodes to verify the identity of nodes with whom it is communicating. The disadvantage of this scheme is the limitation for maximizingthe network size due to the maximum node's storage overhead and with the adjusted probability of sharing keys between nodes.

11.6.5 Polynomial Pool-Based Key Pre-distribution Scheme

The polynomial pool-based key pre-distribution scheme proposed by Donggang Liu, and Peng Ning [35] can be considered a combination of both ideas in [40]: Polynomial Based Key Pre-distribution scheme and Random Pairwise Keys scheme, with the idea of Probabilistic Key Pre-distribution (PRE) scheme [39]. It issimply based on the


$f_1(x,y), f_2(x,y), \ldots, f_n(x,y)$
A set F of t-degree polynomial
i
A subset F_i $\{f_{i1}(i,y), f_{i2}(i,y), \ldots, f_{ik}(i,y)\}$

Parameters	Description
$f_1(x,y) \ldots f_n(x,y)$	A set of multiple bivariate t-degree polynomials over a finite field F.
F_i $\{f_{i1}(i,y), \ldots f_{ik}(i,y)\}$	$F_i \subseteq F$ were selected for node ID (i)
i	Node's ID

Fig. 11.5 The key pre-distribution setup phase

fact that not all pairs of sensor nodes have to establish a pairwise key in between. The Polynomial pool-based key pre-distribution schemegoes through three phases:

- ***Phase 1: Key pre-distribution setup***, in where a setup server generates a pool of multiple bivariate t-degree symmetric polynomials (i.e., $f(x, y) = f(y, x)$) over a finite field Fq. The setup server also uses nodes' IDs for assigns a randomly selected subset of these polynomial shares to each sensor node as demonstrated in Fig. 11.5 (that was driven from [41]).
- ***Phase 2: Direct Key Establishment***, any two nodes can establish a pairwise key for their communication channel if both of them have shares on the same bivariate polynomial in their subsets. The process of polynomial share discovery may be predetermined before deployment or a real-time on the fly discoveryas illustrated in Fig. 11.6 (that was redrawn from [to each sensor node as demonstrated in Fig. 11.5 (that was driven from [41]).
- ***Phase 3: Path Key Establishment***, if there are no common polynomials shares between two sensor nodes, both nodes try to connect with an intermediated node(s) adjacentto them (neighbor nodes) trough phase 2. These intermediated nodes act as a communication path (route) between the previous two nodes as illustrated in Fig. 11.7 (that was redrawn from [to each sensor node as demonstrated in Fig. 11.5 (that was driven from [41]).

The Polynomial pool-basedpre-distribution scheme provides a great security performance reached to less than 60 % of compromised nodes when compared to other previously defined security scheme. For instance, the Polynomial pool-based pre-distribution scheme offers better resilience to node capture attack over the Random pairwise key scheme as the polynomial share is used no more than t times, which makes it difficult for an attacker to capture specific nodes (t sensor nodes) in order to compromise the key derived from their particular polynomial [35, 42].

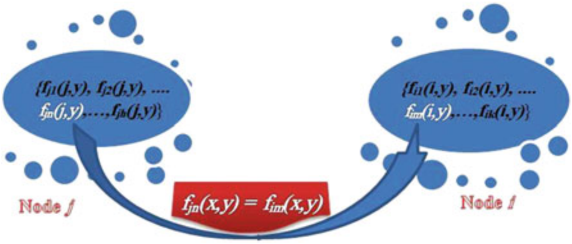

$\{f_{j1}(j,y), f_{j2}(j,y), ...$
$f_{jn}(j,y),\ldots,f_{jb}(j,y)\}$
$\{f_{i1}(i,y), f_{i2}(i,y), ...$
$f_{im}(i,y),\ldots,f_{ik}(i,y)\}$
$f_{jn}(x,y) = f_{im}(x,y)$
Node j
Node i

Fig. 11.6 The direct key establishment phase

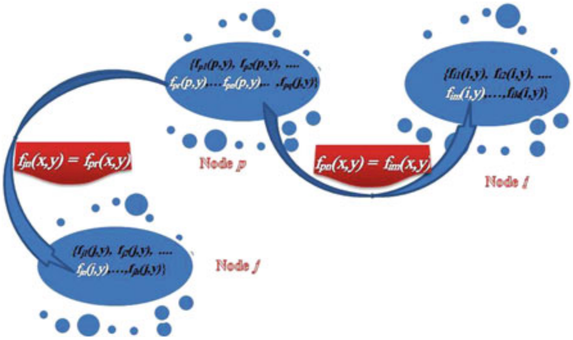

$\{f_{p1}(p,y), f_{p2}(p,y), ...$
$f_{pr}(p,y)....f_{pn}(p,y)...\ .f_{pq}(j,y)\}$
$\{f_{i1}(i,y), f_{i2}(i,y), ...$
$f_{im}(i,y),\ldots,f_{ik}(i,y)\}$
$f_{jn}(x,y) = f_{pr}(x,y)$
Node p
$f_{pn}(x,y) = f_{im}(x,y)$
Node i
$\{f_{j1}(j,y), f_{j2}(j,y), ...$
$f_{jn}(j,y),\ldots,f_{jb}(j,y)\}$
Node j

Fig. 11.7 The path key establishment phase

Another advantage guaranteed by Polynomial pool-basedpre-distribution scheme over the Random pairwise key scheme, that new sensor nodes can be added dynamically without the need to consult the previously deployed sensor nodes (the setup server in the Random pairwise key scheme has to predetermine additional space for any added nodes may be in future, this put in an extra storage overhead even if these nodes never been deployed) so it provides efficient network scalability [35, 42].

The Polynomial pool-based pre-distribution scheme is more efficient and flexible than the probabilistic scheme and q-composite scheme in terms of communication overhead owing by the small number of keys carried by each sensor node and the small size of the pool of keys. On the other hand the Polynomial pool-based pre-distribution scheme has an expensive computational overhead since it has to evaluate a $\boldsymbol{t}$-degree polynomial [35, 42].

11.6.6 Location Based Pairwise Key Scheme

Usually, position based routing (Geographic routing) is susceptible to a number of attacks that inject fake routing information such as Sybil attacks, Sinkhole attacks, and selective forwarding attacks. Zhao et al. [43] proposed a merging scheme between

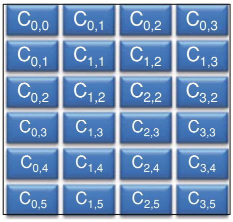

$C_{0,0}$ $C_{0,1}$ $C_{0,2}$ $C_{0,3}$
$C_{0,1}$ $C_{1,1}$ $C_{1,2}$ $C_{1,3}$
$C_{0,2}$ $C_{1,2}$ $C_{2,2}$ $C_{3,2}$
$C_{0,3}$ $C_{1,3}$ $C_{2,3}$ $C_{3,3}$
$C_{0,4}$ $C_{1,4}$ $C_{2,4}$ $C_{3,4}$
$C_{0,5}$ $C_{1,5}$ $C_{2,5}$ $C_{3,5}$

Fig. 11.8 Adjacent equal-sized cells [43]

the location pairwise keys bootstrap security scheme [44] and the Geographical and Energy Aware Routing Protocol (GEAR) [45]. The proposed SGEAR (Secure Geographic and Energy Aware Routing) Starts by partitioning the area of interest into small equal-sized adjacent zones called cells (Fig. 11.8). Each cell signified as $\boldsymbol{C_{r,\,c}}$ (where $\boldsymbol{r}$ represent its raw position, andcrepresent its column position) is associated with a unique random bivariate polynomial $\boldsymbol{f_{r,\,c}(x,\,y)}$. That, for each sensor located in a cell, the setup server distributes to it a set of polynomial shares for its home cell and for its four adjacent cells.

Since the sensor authentication key is related to its location, and its assigned polynomial share that related to its ID, it's difficult to a Sybil attacker to obtain a polynomial share of the location even if it declare multiple identities. Also, the SGEAR security routing is reliable against selective forwarding attacks; that it is provide the ability to forwarding packets through several paths from source to destination.

11.6.7 Time Based Pairwise Key Scheme

Jang et al. [46] had mentioned a security beach that could happened during the initial key establishment phase (their experiments where applied on the Localized Encryption and Authentication Protocol (LEAP) [47]). That most of the key pre-distribution techniques assume that the required time to discover the neighboring nodes $\boldsymbol{T_{est}}$ and the time interval for an attacker to compromise a sensor node $\boldsymbol{T_{min}}$, is larger than $\boldsymbol{T_{est}}$. But in fact, due to technical or environmental issues, it may takes longer than expected $\boldsymbol{T_{est}}$ which makes the sensor node vulnerable to adversary attacks. Jang introduces a time-based key pre-distribution security protocol. The proposed protocol uses a multiple initialkey $\boldsymbol{K_I}$ over a probabilistic time intervals. Within his article [46] the process starts by the key setup phase that generates a pool of initial keys $\boldsymbol{K_I}$. Also, the lifetime of the sensor network is being divided into equal time slots $\boldsymbol{T_I}$. Each time slot is randomly assigned with an initial key from the pool. In further step each sensor node selects randomly a set of initial keys which attached with one of the deployment time slots and master keys $\boldsymbol{m}$. These master keys are calculated from the chosen initial key and sensor ID.

Figure 11.9 illustrates the main concept behind the proposed time-based protocol. The figure shows the lifetime of a sensor network after its division into groups $\boldsymbol{N_n}$.

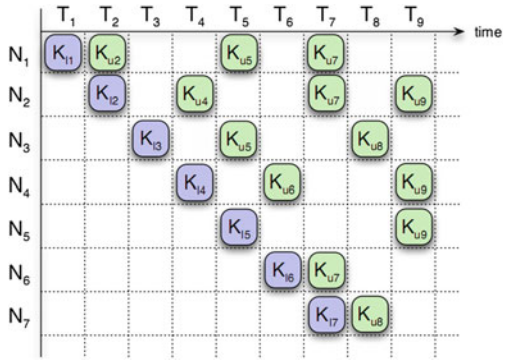


Fig. 11.9 Probabilistic time intervals for initial key $\boldsymbol{K_I}$ and master keys $\boldsymbol{K_{uI}}$ for each node group [46]

Each group within a time slot has an initial key $\boldsymbol{K_I}$ and a set of randomly calculated master keys $\boldsymbol{m}$. For the first time slot after node's deployment, each sensor nodes can set pairwise keys with other neighboring nodes using its $\boldsymbol{K_I}$. For the other time slots, sensor nodes are able to establish a secure link with other nodes through one of the m master keys.

Although the proposed protocol does not show the better performance compared to other protocols such as LEAP, it reduces the possibility of compromising a sensor network through the discloser of the initial key of a captured node.

The proposed protocol is being tested in [48] against the cloning attack, which is injecting a clone adversary attack of a compromised node. That the capture of a node can recover the primary key of specific period, so it can be easily cloned. Therefore the proposed protocol requires more enhancements in order to produce a suitable solution against this type of attacks.

11.7 Conclusion and Future Works

In countless applications, sensor nodes are often deployed in hostile and uncontrolled environments. Accordingly there is a crucial need for security mechanism to ensure a number of desired security services for operating these nodes, such as data confidentiality, the node authentications, network availability, etc.

Providing such fundamental security service for these networks is a challenge owing to the resource constraints attached with theses nodes. For instance, it is not practical to apply asymmetric cryptograph algorithms. The suitable key management protocols for WSN are based on symmetric key algorithms. Recently a number of key pre-distribution techniques have been suggested for establishing a symmetric pairwise key between sensor nodes. This chapter magneton state-of-the-art of pairwise key pre-distribution techniques proposed for wireless sensor networks.

As a future direction, the recently proposed key pre-distribution techniques have a number of drawbacks, which makes the wireless network vulnerable to a variety of adversary attacks. Furthermore, more studies are needed to determineboth the size of key pool, and the suitable probability for neighbor nodes to share common keys, in addition toreducing the threat of compromised nodes.

References

1. http://www.itu.int (2014). Accessed 16 Jan 2014
2. I.-T.R. X.800: Security Architecture for Open Systems Interconnection for CCITT Applications, Geneva (1991)
3. Ozgur, A., Lévêque, O., Tse, D.: Spatial degrees of freedom of large distributed MIMO systems and wireless ad hoc networks. IEEE J. Sel. Areas Commun. **31**(2), 202–214 (2013)
4. Yassine, M., Ezzati, A.: A review of security attacks and intrusion detection schemes in wireless sensor network. Int. J. Wireless Mob. Netw. **5**(6), 79 (2013)
5. Aashima, S., Sachdeva, R.: Review on security issues and attacks in wireless sensor networks. Int. J. **10**(1), 1–24 (2013)
6. Singh, S.K., Singh, M.P., Singh, D.K: A survey on network security and attack defense mechanism for wireless sensor networks. Int. J. Comput. Trends Technol. **1**(1) (2011) (ISSN: 2231–2803)
7. Gopalakrishnan, S.: A survey of wireless network security. Int. J. Comput. Sci. Mobile Comput. **3**(1), 53–68 (2014)
8. Fouad, M.M.M., El-Bendary, N., Ramadan, R.A., Hassanien, A.E.: Wireless sensor networks: a medical perspective. In: Wireless Sensor Networks: From Theory to Applications Book, CRC Press, USA (ISBN 9781466518100)
9. Manisha, P., Singh, Y.: Security issues and sybil attack in wireless sensor networks. Int. J. P2P Netw. Trends Technol. **3**(1), 7–13 (2013)
10. El-Bendary, N., Fouad, M.M.M., Ramadan, R.A., Banerjee, S., Hassanien, A.E.: Smart environmental monitoring using wireless sensor networks. In: Wireless Sensor Networks: From Theory to Applications Book, CRC Press, USA (2013) (ISBN 9781466518100)
11. Yongji, R.: BRS-based robust secure localization algorithm for wireless sensor networks. Int. J. Distrib. Sens. Netw. **2013**, 46–55 (2013)
12. Karlof, C., Wagner, D.: Secure routing in sensor networks: attacks and countermeasures. Elsevier's Ad Hoc Netw. J. Spec. Issue Sens. Netw. **1**(2–3), 293–315 (2003)
13. Mohammadi, S., Hossein, J.: A Comparison of Link Layer Attacks on Wireless Sensor Networks. arXiv preprint arxiv.org/abs/1103.5589, (2011)
14. Zomaya, A.Y.: Algorithms and Protocols for Wireless Sensor Networks. Wiley, New York (2009) (ISBN 978-0-471-79813-2)
15. Sen, J.: A survey on wireless sensor network security. Int. J. Commun. Netw. Inf. Secur. **1**(2), 55–78 (2009)
16. Gabrielli, A., Mancini, L.V., Setia, S., Jajodia, S.: Securing topology maintenance protocols for sensor networks. In: 1st Conference on IEEE Security and Privacy for Emerging Areas in Communications Networks (SecureComm), pp. 101–112. ACM Press, New York (Sept 2005)
17. Gabrielli, A., Conti, M., Di Pietro, R., Mancini, L.: Sec-TMP: a secure topology maintenance protocol for event delivery enforcement in WSN. In: Proceedings of 5th Conference on Security and Privacy in Communication Networks (SecureComm 2009), pp. 265–284. Athens, Greece (Sept 2009)
18. Stajano, F.: Security for Ubiquitous Computing. Wiley, New York (2002)
19. Matthew, P., Sencun, Z., Vijaykrishnan, N., McDaniel, P., Kandemir, M., Brooks, R.: The sleep deprivation attack in sensor networks: analysis and methods of defense. Int. J. Distrib. Sens. Netw. **2**(3), 267–287 (2006)

20. Meenakshi, T., Gaur, M.S., Laxmi, V., Sharma, P.: Detection and countermeasure of node misbehaviour in clustered wireless sensor network. In: Proceedings of the ISRN Sensor Networks (2013)
21. Xu, Y., Heidemann, J., Estrin, D.: Geography-informed energy conservation for Ad-hoc routing. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 70–84. Atlanta (2001)
22. Chen, B., Jamieson, K., Balakrishnan, H., Morris, R.: Span: an energy-efficient coordination algorithm for topology maintenance in ad hoc wireless networks. Wireless Netw. **8**(5), 481–494 (2002)
23. Xu, Y., Heidemann, J., Estrin, D.: Energy conservation by adaptive clustering for ad-hoc networks. In: Poster Session of the 3rd ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc'02). Lausanne, Switzerland (June 2002)
24. Fouad, M.M.M., Dawood, A.R., Mostafa, M.-S.M.: Study of the effects of pairwise key predistribution scheme on the performance of a topology control protocol. In: Proceedings of the 2nd International Workshop on Mobility in Wireless Sensor Networks (MobiSensor'2011) that held in conjunction with the 7th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS '2011), Barcelona, Spain, (June 2011)
25. Chan, H., Perrig, A., Song, D.: Random key predistribution schemes for sensor networks. In: Proceedings of IEEE Symposium on Security and Privacy, IEEE Computer Society, pp. 197–213. Washington, DC, (2003)
26. Diffie, W., Hellman, M.E.: New directions in cryptography. IEEE Trans. Inf. Theory **22**, 644–654 (1976)
27. Rivest, R.L., Shamir, A., Adleman, L.M.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
28. Perrig, A., Szewczyk, R., Wen, V., Cullar, D., Tygar, J.D.: Spins: security protocols for sensor networks. In: Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom), pp. 189–199. Rome, Italy, (July 2001)
29. Camtepe, S.A., Yener, B.: Key distribution mechanisms for wireless sensor networks: a survey. Rensselaer Polytechnic Institute, pp. 05–07. Troy, New York, Technical Report (2005)
30. Liu, D., Ning, P., Li, R.: Establishing pairwise keys in distributed sensor networks. ACM Trans. Inf. Syst. Secur. **8**(1), 41–77 (2005)
31. Fouad, M.M.M., Mostafa, M.-S.M., Dawood, A.R.: SOPK: Second opportunity pairwise key scheme for topology control protocols. In: Proceedings of IEEE 3rd International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 632–638. (2012)
32. Anderson, R., Kuhn, M.: Tamper resistance—a cautionary note. In: Proceedings of the Second Usenix Workshop on Electronic Commerce, pp. 1–11. (November 1996)
33. Kalpana, S., Ghose, M.K., Kumar, D., Singh, R.P.K., Pandey, V.K.: A comparative study of various security approaches used in wireless sensor networks. Int. J. Adv. Sci. Technol. **17**, 31–44 (2010)
34. Blundo, C., Santis, A.D., Herzberg, A., Kutten, S., Vaccaro, U., Yung, M.: Perfectly-secure key distribution for dynamic conferences. Advances in Cryptology. Lecture Notes in Computer Science, pp. 471–486. Cambridge University Press, Cambridge (1992)
35. Liu, D., Ning, P.: Security for wireless sensor networks. Advances in Information Security, vol. 28, pp. 63–76. Springer (Dec 2006) (ISBN 0387327231)
36. Ming, L., Yu, S., Guttman, J.D., Lou, W., Ren K.: Secure ad hoc trust initialization and key management in wireless body area networks. ACM Trans. Sens. Netw. (TOSN) **9**(2), 18 (2013)
37. Spencer, J.: The strange logic of random graphs. Algorithms and Combinatorics 22, Springer-Verlag, (2000). ISBN: 3-540-41654-4
38. Seema, V.: Analysis of a new random key pre-distribution scheme based on random graph theory and kryptograph. In: Mobile Communication and Power Engineering, pp. 349–352. Springer, Berlin (2013)
39. Eschenauer, L., Gilgor, V.D.: A key management scheme for distributed sensor networks. In: Proceedings of 9th ACM Conference on Computer and Communications Security, pp. 41–47 (Nov 2002)

40. Wu, S.-L., Tseng, Y.-C.: Wireless Ad Hoc Networking. Wireless Networks and Mobile Communications Series, Auerbach Publications, USA (2007)
41. Liu, D., Ning, P.: Establishing Pairwise Keys in Distributed Sensor Networks. Lecture Notes in Computer Science. http://www.cs.iastate.edu/~cs610jw/CS610_LiuNing.ppt (2014). Accessed 15 Jan 2014
42. Shuang-Hua, Y.: WSN security. In: Wireless Sensor Networks, pp. 187–215. Springer, London (2014)
43. Zhao, H., Li, Y., Shen, J., Zhang, M., Zheng, R., Wu Q.: A New Secure Geographical Routing Protocol Based on Location Pairwise Keys in Wireless Sensor Networks (2013)
44. Liu D., Ning, P.: Location-based Pairwise Key Establishments for Static Sensor Networks. In: Proceedings of the 1st ACM Workshop on Security of Ad Hoc and Sensor Networks, (2003)
45. Yu, Y., Govindan, R., Estrin D.: Geographical and Energy Aware Routing: a Recursive Data Dissemination Protocol for Wireless Sensor Networks. Technical report ucla/csd-tr-01-0023, UCLA Computer Science Department, (2001)
46. Jang, J., Kwon, T., Song, J.: A time-based key management protocol for wireless sensor networks. Information Security Practice and Experience, pp. 314–328. Springer, Berlin (2007)
47. Zhu, S., Setia, S., Jajodia, S.: Leap+: efficient security mechanisms for large-scale distributed sensor networks. ACM Trans. Sen. Netw. **2**(4), 500–528 (2002)
48. Nait Hamoud, O., Kenaza, T., Nouali-Taboudjmat N.: The cloning attack vulnerability in WSN key management schemes. In: SENSORCOMM 2013, The 7th International Conference on Sensor Technologies and Applications, pp. 151–156. (2013)

Part III
Biometrics Technology and Applications

Chapter 12
Fusion of Multiple Biometric Traits: Fingerprint, Palmprint and Iris

N. L Manasa, A Govardhan and Ch Satyanarayana

Abstract Biometric recognition protocols involving single sources of information for human authentication which are commonly termed as unimodal systems though show satisfying performance, still suffered from problems relating to non-universality, permanence, collectability, convenience and susceptibility to circumvention. This paper emphasizes the priority of biometric information fusion by analyzing two kinds of fusion: Fusion of multiple representations of single biometric trait and Fusion of multiple biometric traits. As biometric traits possess large variance between persons and small variance between samples of the same person, it is important to capture this information using multiple representations at both global-level and local-level and perform fusion at feature-level. As a feature set is a straightforward representation of raw biometric data, it is theoretically presumed to incorporate richer information. Hence, we propose to use a fusion method that maximally correlates information captured from both the features and eliminates the redundant information giving a more compact representation. Fusion of multiple biometric traits is realized using fingerprint, palmprint and iris modalities. We explore this kind of fusion using two architectures: Parallel architecture and Hierarchical-cascade architecture. Multi-biometric recognition systems designed with hierarchical architecture not only are robust, fast and highly secure but also mitigate problems like missing and noisy data associated with parallel and serial architectures respectively, not to be forgotten that parallel architectures are preferred in high security-demanding defense/military applications as they evidently provide more precision for the reason that they combine more modalities and evidences about the user for recognition. Parallel framework

N. L. Manasa (✉) · Ch. Satyanarayana
Jawaharlal Nehru Technological University Kakinada, Kakinada, Andhra Pradesh, India
e-mail: nadipally.m@gmail.com

Ch. Satyanarayana
e-mail: chsatyanarayana@yahoo.com

A. Govardhan
Jawaharlal Nehru Technological University Hyderabad, Hyderabad, Andhra Pradesh, India
e-mail: govardhan_cse@jntuh.ac.in

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_12, © Springer-Verlag Berlin Heidelberg 2014

proposed in this work takes advantage of score-level fusion. Score-level fusion is widely put to use as it offers best trade-off between ease and efficiency. We propose two score-level fusion techniques which rely on Equal Error Rates of individual modalities. Since error rate is a percentage of misclassified samples, we attempt to minimize the overlapped area between genuine and imposter curves by choosing to maximize the stability of the modality with superior performance. The proposed rule addresses the fusion problem from error rate minimization point of view so as to increase the decisive efficiency of the fusion system. To take the advantage of feature-level fusion, serial/cascade architecture and hierarchical architectures, we also propose a two-stage cascading frame-work based on fusion of fingerprint and palmprint feature sets in the first stage and iris features to eliminate the ambiguity of false matches in the next stage. The proposed frame work takes advantage of both unimodal and multimodal architectures. Proportionate experimental results reported on both real and virtual databases in this work demonstrate the superior performance of a multimodal recognition system over a unimodal system but however infers that the design of a multimodal biometric system predominantly depends on the application criteria and so is difficult to anticipate the best fusion strategy. The review of biometric based recognition systems indicate that a number of factors including the accuracy, cost, and speed of the system may play vital role in assessing its performance. But today with the cost of biometric sensors constantly diminishing and high speed processors and parallel programming techniques widely available to affordable research, accuracy performance has become predominant focus of biometric system design. The main aim of the present work is to improve the accuracy of a multimodal biometric recognition system by reducing the error rates.

12.1 Introduction

Human authentication happens in a three step process: identify-authenticate-authorize, which is done countless times every minute around us by humans and computers. Human authentication has three important alternatives to choose from: using something we know (such as passwords and pass-phrases), something we have (such as access tokens, smart cards, and so on) or something we are (biometrics). Each authentication method has its proposals and counter-arguments. Each authentication method can be chosen keeping in mind the application, the environment and the users. Biometrics is a special field of human authentication which today finds boundless applications in myriad fields.

Contemporaneity of the research community has evidently endowed the world of security with appreciable number of biometrics. But the question remains: apart from being secure, can a biometric trait also be comprehensive. That is to say, easily accessed and conditions to setup such a biometric security system be favorable. So intuitively, a cost effective, accelerated, portable, accurate, hazzle-free acquisition and testing based biometric is the requisite of the day. But de facto, in regard to a secure and non-counterfietable biometric, none of the available biometric traits are non-counterfietable. With the advent of technology and information invasion, today security has its predicament. We analyze different biometric traits that are used in commercial application packages today and enlist their pros and cons, hence enlightening the importance of multi-modal biometric recognition systems over unimodal systems.

Hand Geometry systems measure demographic aspects like thickness and width of the palm, length and width of fingers, and so on. Because of its adaptability, ease of measurement and storage, hand Geometry based systems are highly acceptable than finger-print based systems. But because of its biometric properties it is only suitable for verification mode and fails in identification mode. Also the false acceptance rates of a biometric system based solely on hand geometry features might be high. It always has to be associated with some other biometric trait for perfect authentication.

Retinal scanning based systems use infrared illumination while acquisition and compare images of the blood vessels in the back of the eye, the choroidal vasculature [1]. Apparently, retina recognition not only is futile for people suffering from serious eye illnesses but also raises privacy issues in case of misuse of acquired data.

DNA 99.7 % of human DNA is shared. 0.3 % (1 million nucleotides) is variable and so is unique [2]. These variable regions, called Short Tandem Repeats (or STRs), can be examined to distinguish one person from another. DNA samples can be isolated from a sample such as saliva, blood, hair, tissue or semen. But it suffers from the following complications: (i) DNA matching is not done in real-time, a physical sample must be taken unlike other biometric systems which use an image or a recording, (ii) invasion of civil liberties, (iii) storage of DNA and (iv) extraction and process time [3]. DNA based biometric system cannot be easily simulated but is invasive and arduous to setup.

Complex Eye Movements is a very recent biometric trait which was brought to light by komogortsov et al. [4] and kasprowski et al. [5]. They carried out considerable research on CEM and established commendable results. In [4] Complex Eye Movements are combined with Oculomotor Plant Characteristics where a mathematical model for eye and its associated muscle movement is established when eyes respond to a stimuli. This Biometric is still in its infancy but seems to be a non-counterfietable Biometric. But in regard to its ease-to-acquire and easy-to-use, justification is still void.

Hand Vascular Pattern makes use of infrared light to produce an image of a person's vein pattern in their face, wrist or hand, as veins are stable through one's life. The vein pattern recorded by any device like video camera is used as a personal code which is acutely laborious to duplicate. The fact that the use of this biometric needs no physical contact with the sensor and that it provides notable convenience and no performance degradation even with scars or hand contamination makes this physiological biometric a reliable one.

Face From facile edge-based algorithms to advanced pattern recognition methods, a wide range of techniques have been proposed for face recognition. Numerous existing face recognition techniques succeed with frontal faces of similar sizes [6] and even with distorted facial images [7]. While in reality, this presumption may not hold good as human face is dynamic in nature hence has a high degree of variability in its appearance, making face detection an intricate problem in computer vision. Factors such as changing hairstyles, beard, moustache and aging only make righteous face recognition more difficult. Bruce Schneier, in his book Beyond Fear calculated the math and stated that if a face detection system is 99.9 % accurate, still it would generate 10,000 false alarms for every single real terrorist in 10 million civilians.

Voice verification identifies myriad characteristics of a human voice like frequency, nasal tone, cadence, inflection to recognize the speaker. Voice recognition systems take advantage that they do not require expensive input devices and can even accomplish the recognition task in the background while the person speaks without explicitly forcing the users to spend time to do the same. But like all other biometrics, voice systems have their fair share of shortcomings, for instance, record and play attacks in fixed-text models, also some people might skillfully duplicate/imitate others' voices. Voice of an individual may also change with age, illness, mental state, etc.

Gait can be defined as the coordinated, cyclic combination of movements that result in human locomotion [8]. Examples of gaits include jogging, running, walking, jumping, sitting down, picking up an object, climbing stairs, etc. According to [8] the performance of gait recognition systems is below what is required for use in biometrics as this biometric recognition system is confounded by the following factors viz, terrain, injury, footwear, any kind of training to the human body, passage of time.

Body Odor It is a contact-less biometric that confirms a person's identity by analyzing the olfactory properties of the human body scent [9]. Cambridge university has developed electronic sensors to gather the human odor, usually from the non-intrusive areas, such as the back of the hand. Each human smell is made up of chemicals known as volatiles. Each chemical of the human odor is extracted by the biometric recognition system and converted into a unique data string. But privacy of the individual will be compromised while using this biometric as body odor carries an amount of sensitive personal information [3].

Signature This biometric is in use for several centuries and so is largely accepted as a biometric. There are two subtypes of signature verification systems, namely, static signature verification systems and dynamic signature verification systems. In the later subtype, speed, velocity, pressure, angle of the pen and the number of times the pen is lifted from the pad, etc will be measured while in static subtype only the image of the signature is used. The dynamic signature verification is more secure and reliable than static signatures [10]. Shortcomings of signature biometrics include inconsistence i.e., signatures lack permanence: which may change under the influence of illness, emotions, age, etc. These systems render performance only in verification mode and not in identification mode.

Fingerprint verification though has the advantage that no two individuals possess the same fingerprints, not even identical twins, suffers from few disadvantages.

Dry, wet, damaged, dirty or diseased skin may affect the quality of the fingerprint. With fingerprints, the attacking technology is as easeful as the defending technology, this fact has been proven by many successful security attacks. Tsutomu Matsumoto of the Yokohama National University victoriously counterfeited numerous fingerprint based biometric systems into accepting fake fingers made of gelatin gaining an 80-percent success rate. It is also difficult to acquire fingerprint features for some classes of people like manual laborers, elderly people, etc. Inspite of these shortcomings, till date fingerprint recognition algorithms are researched upon for perfection because of their applicability and entailment. For instance, almost always in forensic scenarios latent prints are the only trace for fraud identification.

Palmprint As palmprint comprises wider area than any other biometric traits, it can be used even in fallacious conditions like burns, boils, cuts, dirt and oil stains on palms. Also when fading of palm texture occurs due to lot of physical work with hands.

Iris is a thin annular structure around the pupil of the human eye. Its complex pattern is constructed of many idiosyncratic features such as fibres, freckles, furrows, arching ligaments, ridges, serpentine vasculature, rings, rifts and corona. All these establish a distinctive signature for human authentication. Patterns in human iris have abundance of invariance. Iris patterns emerge during the eighth month of the fetal term and remain stable throughout the life time of an individual. On the lines of precision, [11] report successful authentication across millions of cases without a single failed test. Given its non-invasive nature and affordable hardware solutions [12], iris based authentication systems have become an indispensible tool for many high-security applications. Apparently, both iris and retina recognition would not work for visually-challenged people and people suffering from serious eye illnesses.

Research demonstrates that fingerprint, palmprint and iris are the most widely used biometrics of the moment. Fingerprint and palmprint are accepted for their ease-to-setup and use characteristics and iris is preferred for its uniqueness and less-intrusiveness characteristic. So, the former is generally put to practice in low-to-medium security places and the latter in high security-demanding scenarios. Besides, features for representing these three biometric traits can be extracted even from very low-resolution images collected from touch-less, portable, less-expensive acquisition devices.

Upon working with biometric systems involving single sources of information for human authentication which are commonly termed as Unimodal systems, we realized that these systems suffered from problems relating to non-universality, permanence, collectability, convenience and susceptibility to circumvention. According to Roger Clarke [13] every biometric should possess the following characteristics, Fig. 12.1, shows the 12 characteristics that a biometric should possess. Ironically no known biometric completely satisfies all these criteria.

A part of these problems can be mitigated by using Multimodal biometric systems, which are systems that consolidate evidence from multiple biometric sources. Apart

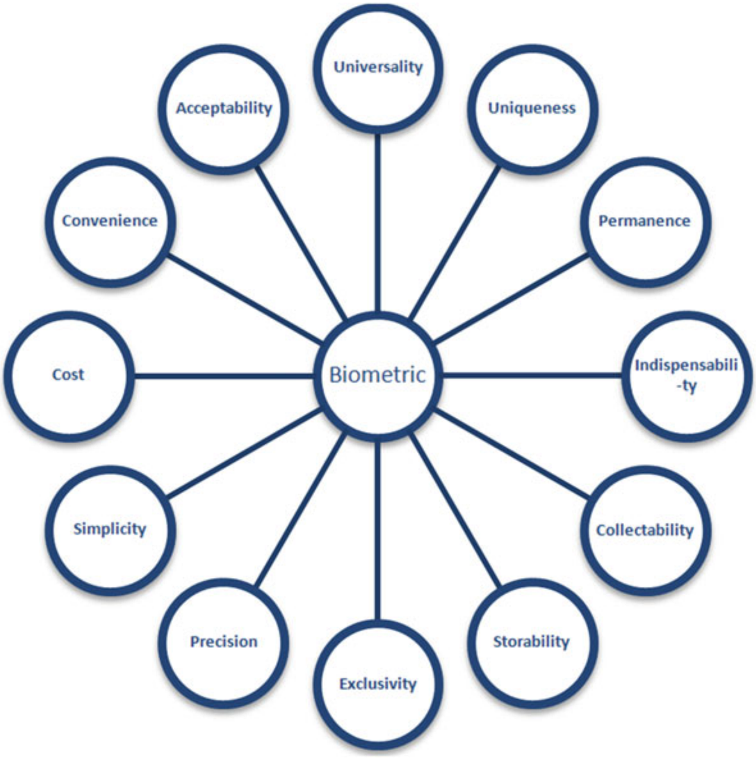


Fig. 12.1 Essential characteristics of a biometric

from leveraging on these factors, multimodal systems possess lot many advantages over unimodal systems. They are proven to significantly improve the accuracy and precision of the system hence reducing the Failure-to-Enroll rates as they use information from more than one biometric trait. Multimodal biometric systems impede spoof attacks and major circumvention as it is comparably difficult to defraud multiple biometric traits. They also ensure speed when dealing with large databases when designed appropriately. Though matching of more than one biometric trait increases the response time, they can be designed hierarchically by relying on simple modality in the first stage and complex time-consuming modality in the final stage thus ensuring robustness. More advantages of a multimodal recognition systems can be found in Sect. 1.3.4.

Many multimodal authentication systems have been proposed in literature each involving different modalities, different fusion mechanisms, different architectures and different levels of fusion. There are pros and cons to every fusion mechanism which proves that the design of a multimodal biometric system predominantly depends on the application criteria.

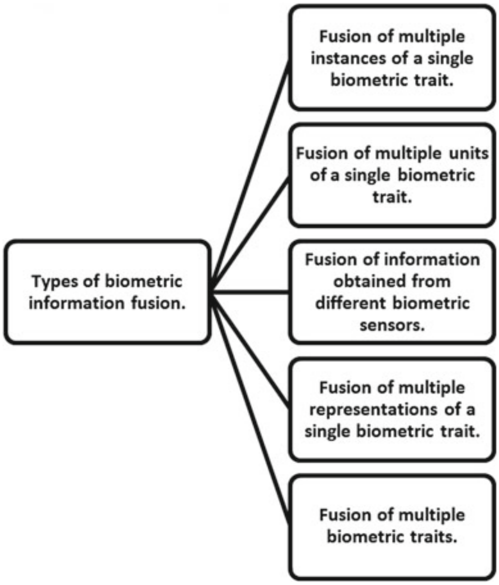

Types of biometric information fusion.
Fusion of multiple instances of a single biometric trait.
Fusion of multiple units of a single biometric trait.
Fusion of information obtained from different biometric sensors.
Fusion of multiple representations of a single biometric trait.
Fusion of multiple biometric traits.

Fig. 12.2 Types of biometric information fusion

12.1.1 Kinds of Information Fusion

Information fusion in the world of biometrics materializes in five different ways, see Fig. 12.2 viz.,

1. Fusion of multiple instances of a single biometric trait, e.g., multiple instances of face taken over time. This fusion ensures permanence but aggravates cost by increasing the storage space.
2. Fusion of multiple units of a single biometric trait, e.g., fusion of left palm and right palm to recognize an individual. This fusion satisfies uniqueness and precision characteristics to an extent.
3. Fusion of information obtained from different biometric sensors which capture the same biometric trait, e.g., multi-spectral palmprint images captured at different wavelengths of the electromagnetic spectrum. Fusion of such distinguishing information gratifies uniqueness and precision but aggravates cost and hinders convenience and collectibility.
4. Fusion of multiple representations of a single biometric trait, e.g., using global texture features and local texture features to represent iris texture. This fusion ensures uniqueness and precision as different representations capture varied patterns of information from the same image. Also is cost-effective, facilitates collectivity, and convenience as single capturing device is used. Since a single image is sufficient, storage space is reduced.
5. Fusion of multiple biometric traits, e.g., fusion of fingerprint, palmprint and iris features of an individual. This kind of fusion satisfies the most important

characteristic: uniqueness, because if one biometric trait is lost, authentication can still be assured using the other trait. Also ensures universality, precision and exclusivity but aggravates cost, inconvenience, simplicity is lost and increases storage space.

As stated above, every fusion mechanism has advantages and drawbacks. But as the last two fusion techniques catered to the needs of modern day recognition applications, considerable research underwent in exploring them [14]. Hence, we study the performance evaluation of these two kinds of biometric fusion mechanisms involving three different biometric modalities viz., Fingerprints, Palmprints and Iris. Fusion of multiple representations of a single biometric trait is studied in Sect. 12.4 while Fusion of multiple biometric traits in regard to two different architectures is studied in Sect. 12.5.

12.2 Motivation and Challenges

12.2.1 Fusion of Multiple Representations of Single Biometric Trait

A good biometric trait should prove large variance between persons and small variance between samples of the same person. Hence, it is important to capture features at both global-level and local-level. Such appropriate feature that can be extracted even from low-resolution biometric images is Texture. Texture is one of the clearly observable, permanent, distinguishable features on the biometric image. This peculiarity inspires us to pursue a reliably working authentication system based on texture description. Texture can be clearly described by both, features at the local level and at the global level. Global features possess the following characteristics: (a) insensitive to noise; (b) insensitive to shift changes; (c) easy to compute; and (d) have high convergence within the group and good dispersion between groups [50]. Global features can lower the effect raised by local noises thus supplementing each other. Motivation for the proposed approach is as follows,

- Although biometric samples are diverse among individuals, it is very difficult to distinguish solely based on global texture features as some of these patterns are so similar at a coarse level.
- Most of the widely popular coding-based methods like palmcode [51] neglect the multi-scale characteristic of palm lines [52] and construct authentication systems based on structural similarity.
- In a peg-free and unconstrained acquisition environment, translation and rotation variations are inevitable. Image description at a local level handles such interferences reasonably [53].

Also, as patterns in biometric images have abundance of invariance, the inter-class and intra-class variability of these features makes it difficult for just one set of features to capture this variability. The global features are insensitive to affine transformations, noise, and captures large between-class variance and small within-class variance while local features capture significant within-class variance. A chance of recognition rate being compromised in the case of very large databases is high with different images having similar global features. Hence it is important to use local texture features in combination with global texture features for accurate recognition.

Direct fusion of individual feature sets obtained from respective biometric traits is more appropriate as fusion at this level directly integrates information from diverse biometric traits which render complementary data.

As feature set is a straightforward representation of raw biometric data, it is theoretically presumed to incorporate richer information. So, integration at the feature-level apparently provides better authentication than integration at score-level and decision-level. All these factors inspired us to propose a hybrid fusion approach which performs a feature-level fusion of local and global texture features. The proposed method maximizes the correlation between the feature sets at descriptor or image level.

A feature descriptor fusion method based on Canonical Correlation Analysis is used to combine two features, a global feature set and a local feature descriptor. Canonical Correlation Analysis (CCA) [22] is one of the important statistical multi-data processing methods which deals with the mutual relationships between two random vectors.

The proposed method relies on coarse ROI localization and extracts both the feature descriptors. Canonical correlation analysis is used to combine the features at the descriptor level which ensures that the information captured from both the features are maximally correlated and eliminate the redundant information giving a more compact representation. This approach is thoroughly described in Sect. 12.4.

12.2.2 Fusion of Multiple Biometric Traits

The premise for our work on multimodal biometric recognition is,

- for better performance as no individual biometric trait possesses basic requirements like permanence, universality, distinctiveness [17]. So no monomodal biometric authentication system is free of errors.
- ensures collectability which might be affected by trouble in data acquisition sensors, noise in acquired data, etc.,
- deters intra-class variations that might occur due to large distinction between enrollment template and test template.
- enhances population coverage and acceptability thus lowering failure-to-enroll rate.
- impedes spoof attacks and major circumvention.
- also to complement the strengths and diminish the weaknesses of individual biometric traits.

In the present work, we aim to establish the importance of multi- modal biometric recognition systems involving fingerprint, palmprint and iris biometric traits. The rationale for preferring a combination of fingerprint, palmprint and iris features in this work is:

1. All these three biometric traits can be acquired from contact-less environment hence facilitating a mobile biometric recognition system ensuring portability. With the growing concern towards hygiene, a contact-free biometric recognition system is the need of the day.
2. Fingerprint and palmprint image acquisition not only is easy but also inexpensive while iris image acquisition requires expensive hardware so these traits complement each other.
3. Iris is one among the very few highly secure biometric traits while others being retina, DNA, vein pattern and facial thermogram but they suffer from problems like collectability and permanence.
4. The proposed system relies on texture information which can be extracted even from low-resolution images acquired from a web-camera in a real-time environment with dynamic backgrounds.
5. Fallacious conditions associated with fingerprints and palmprints like burns, boils, cuts, dirt and oil stains, fading of finger/palm texture due to lot of physical work with hands; can be compensated by integrating the authentication system with iris texture features.

Research demonstrates that fingerprint, palmprint and iris are the most widely used biometrics of the moment. Fingerprint and palmprint are accepted for their ease-to-setup and use characteristics and iris is preferred for its uniqueness and less-intrusiveness characteristic. So, the former is generally put to practice in low-to-medium security places and the latter in high security-demanding scenarios. Besides, features for representing these three biometric traits can be extracted even from very low-resolution images collected from touch-less, portable, less-expensive acquisition devices. Hence a multi-modal biometric recognition system based on fingerprint, palmprint and iris biometric traits is appropriate; which to our knowledge, none has explored till date. Also recently, Indian governments initiative for unique identification scheme AADHAAR has adopted a similar scheme for recognition of 1.2 billion individuals. The proposed method is thoroughly elucidated in Sect. 12.5. Segmentation problems associated to a real and dynamic environment, projective distortions that emerge due to the absence of contact plane (scale, rotation, occlusion and translation variations), variations in illumination conditions and non-uniform background; all these problems are well handled by the proposed recognition system.

12.3 Contributions

Significant contributions of the present work are,

- Review of biometric based recognition systems indicate that a number of factors including the accuracy, cost, and speed of the system may play vital role in assessing its performance. But today with the cost of biometric sensors constantly diminishing and high speed processors and parallel programming techniques widely available to affordable research, accuracy performance has become predominant focus of biometric system design [54]. The main aim of this thesis is to improve the accuracy of a biometric recognition system by reducing the Error rates, that is, False Accepts and False Rejects.
- We propose a two-stage hierarchical cascading framework based on fusion of fingerprint and palmprint feature sets in the first stage and iris features to eliminate the ambiguity of false matches in the next stage. The proposed frame work takes advantage of both unimodal and multimodal architectures.
- We propose a new deterministic rule to deal with fusion score optimization problem. The proposed rule in parallel architecture addresses the fusion problem from error rate minimization point of view so as to increase the decisive efficiency of the score-level fusion system. It is a linear, data-driven approach to optimize the cumulative fusion scores with respect to individual error rates and is inspired from one of the measures of central tendency which is harmonic mean as it gives less significance to high-value outliers providing a truer picture of the average. This deterministic approach yields better performance in comparison with the widely used Matcher weighting fusion technique.
- As an inception in literature, we propose a complete contact-free multi-modal recognition algorithm which works significantly on low-resolution images involving fusion of Fingerprints, Palmprints and Iris biometric traits. Reasons for preferring a combination of these three traits over the rest has been clearly elucidated in Sect. 1.3.4. A proprietary real multimodal database is established comprising of fingerprint, palmprint and iris images from 50 subjects. The reason for collecting this dataset is to validate the proposed method on contact-free, low-resolution images, refer Table 12.4.

Rest of the chapter is organized as follows. Section 12.4 explores the fusion of multiple representations of a single biometric trait which is the method proposed to extract individual fingerprint, palmprint and iris features. Section 12.4.1 describes the global feature extraction using DTCW while Sect. 12.4.2 describes the local feature extraction using LBP. Section 12.4.3 explains the feature-level fusion of global and local features of each biometric trait using Canonical Correlation Analysis (CCA). Section 12.4.4 describes the matching criterion adopted in this work. Section 12.5 explores the fusion of multiple biometric traits using two different architectures. Section 12.5.1 illustrates the proposed parallel architecture based multi-modal framework and Sect. 12.5.2 demonstrates the proposed hierarchical-cascade architecture based multi-modal framework. Section 12.6 describes the databases used in this work. Section 12.7 concludes the chapter.

12.4 Fusion of Multiple Representations of Individual Fingerprint, Palmprint and Iris

As patterns in biometric images have abundance of invariance, the inter-class and intra-class variability of these features makes it difficult for just one set of features to capture this variability. Hence multiple representations of single biometric trait are important for accurate recognition.

Sun et al. [44, 45] propose a two-stage cascaded classifier method in which the first stage is a traditional Daugman like classifier and the second stage is a global classifier which are areas enclosed by zero crossing boundaries. Later Sun et al. tries to improve the cascaded classifier using LBP to extract local texture features [46]. This fusion method significantly improves the recognition performance of Daugmans approach [46]. Similarly, Vatsa et al. [48], Park and Lee [21], Zhang et al. [47] also show an improvement in performance by using more than one feature to capture the distinct iris patterns. They fuse the features at the image level where as the proposed method maximizes the correlation between the feature sets at feature level. Experiments depict that the combination of the two features yield better performance than either alone [21]. Kumar and Zhang [49] demonstrated that the way to improve performance is intra-modal combination of texture-based, line-based and appearance-based features in the palm. They used various score-level and decision-level fusion techniques and claimed the superior performance of product of sum rule which still had higher error rates. You et al. [50] designed a hierarchical palmprint recognition approach and thus inferred that local features perform better than hierarchical approach when false acceptance rate is more than 5 %.

These factors inspire us to propose a hybrid feature extraction and fusion approach for biometric recognition based on texture information available in the image. We test our approach on publicly available CASIA [24] databases of three widely used biometric traits which are Fingerprint-Palmprint-Iris.[1]

The Global features are insensitive to affine transformations, noise, and captures large between-class variance and small within-class variance while Local features capture significant within-class variance. Scale, shift and rotation (Affine) invariance, good directional sensitivity properties of Dual-tree Complex Wavelets [19] makes it a choice to capture texture features at global level. Dual-Tree Complex Wavelets capture the global information ensuring scale-invariance and shift-invariance which helps in discriminating between locally similar regions. All these properties along with its good directional selectivity in 2D ensure favorable recognition of similar patterns [55]. DTCW features compensate the error in localizing the region-of-interest as they are invariant to the rotation and the inexact localization.

A chance of recognition rate being compromised in the case of very large databases is high with different FPI images having similar global features. Hence it is important to use local texture features in combination with global texture features for accurate recognition. Experiments depict that the combination of the two features yield better

[1] Referred as FPI in rest of the chapter.

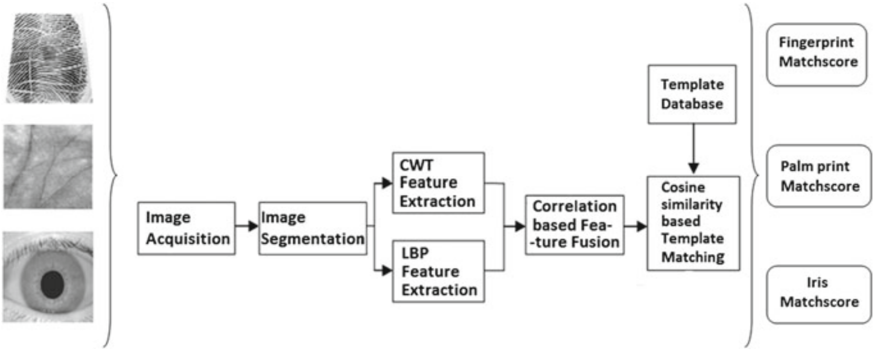


Fig. 12.3 Block diagram for match score extraction of individual biometric trait

performance than either alone [21]. Local Binary Pattern [18] on the other hand being gray-scale and rotation invariant, captures local fine textures effectively. These local features are sensitive to position and orientation of the FPI images.

Canonical Correlation Analysis [22] is used to combine the features at the descriptor level which ensures that the information captured from both the features are maximally correlated and eliminates the redundant information giving a more compact representation. The proposed method relies on coarse segmentation and extracts both the feature descriptors. It maximizes the correlation between the feature sets at feature level. See Fig. 12.3.

12.4.1 Global Feature Extraction

Discrete Wavelet Transforms based methods have been successfully applied to a variety of problems like denoising, edge detection, registration, fusion etc., Discrete wavelet transforms have 4 basic problems such as oscillation, shift variance, aliasing and lack of directionality. By using complex valued basis functions instead of real basis functions these four problems can be minimized. This change is inspired by the Fourier transform basis functions. Complex wavelet transform (CWT) [19] is represented in the form of complex valued scaling functions and complex valued wavelet functions.

$$\psi_c(t) = \psi_r(t) + j\psi_i(t) \tag{12.1}$$

$\psi_r(t)$ are real and even and $j\psi_i(t)$ are imaginary and odd. $\psi_r(t)$ and $\psi_i(t)$ form a Hilbert transform pair (90° out of phase each others) and $\psi_c(t)$ is the analytics signal. Complex scaling function is also defined in similar ways. Projecting the signal onto $2^{j/2}\psi_c(2^j t - n)$, we obtain complex wavelet transforms as follows,

$$d_c(j,n) = d_r(j,n) + jd_i(j,n) \tag{12.2}$$

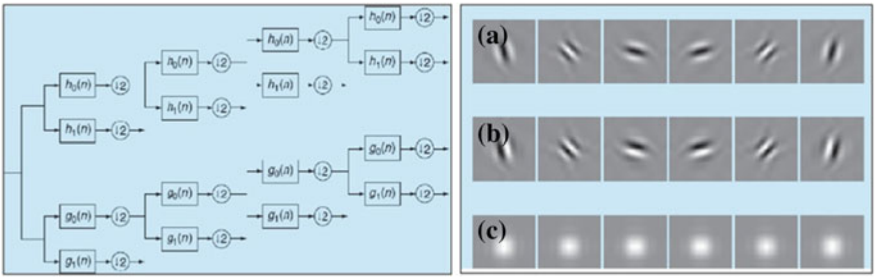


Fig. 12.4 (*left*) DTCW filter bank. (*right*) Typical wavelets associated with the 2-D dual-tree CWT. **a** illustrates the real part of each complex wavelet; **b** illustrates the imaginary part; **c** illustrates the magnitude. Courtesy [19]

In the Complex Wavelet domain, analysis depends on two factors, frequency content (which is controlled by scale factor j) and different time (which is controlled by time shift n).

The Dual tree CWT employs two real Discrete Wavelet Transforms (DWT); the first DWT gives the real part of the transform while the second DWT gives the imaginary part. The analysis Filter Bank (FB) used to implement the dual-tree CWT is illustrated in Fig. 12.4.

To design an overall transform we use two sets of filters. Each set of filter represents real wavelet transform. This overall tranform is approximately analytic. Let $h0(n)$, $h1(n)$ denote the low-pass/high-pass filter pair for the upper FB, and let $g0(n)$, $g1(n)$ denote the low-pass/high-pass filter pair for the lower FB. Denote the two real wavelets affiliated to each of the two real wavelet transforms as $\psi_h(t)$ and $\psi_g(t)$. Filters are designed so that the complex wavelet definition in Eq. (12.1) is approximately estimated. Equivalently, they are designed so that $\psi_g(t)$ is approximately the Hilbert transform of $\psi_h(t)$ [denoted as $\psi_g(t) \approx H\psi_h(t)$]. If the two real DWTs are characterized by the square matrices F_h and F_g, then the dual-tree CWT can be represented by,

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_g \\ \mathbf{F}_h \end{bmatrix} \tag{12.3}$$

If the vector x represents a real signal, then $w_h = F_h x$ represents the real part and $w_g = F_g x$ represents the imaginary part of the dual-tree CWT. The complex coefficients are given by $w_h + j w_g$.

In dual-tree CWT, consider the 2-D wavelet $\psi(x, y) = \psi(x)\psi(y)$ associated with the row-column implementation of the wavelet transform, where $\psi(x)$ is a complex wavelet given by $\psi(x) = \psi_h(x) + j\psi_g(x)$. Obtain $\psi(x, y)$ for the expression,

$$\begin{aligned}\psi(x, y) &= [\psi_h(x) + j\psi_g(x)][\psi_h(y) + j\psi_g(y)] \\ &= \psi_h(x)\psi_h(y) - \psi_g(x)\psi_g(y) \\ &\quad + j[\psi_g(x)\psi_y(x) + \psi_h(x)\psi_g(y)]\end{aligned} \tag{12.4}$$

Take the real part of this complex wavelet, then obtain the sum of two divisible wavelets

$$RealPart\psi(x, y) = \psi_h(x)\psi_h(y) - \psi_g(x)\psi_g(y) \tag{12.5}$$

We perform a 4-level decomposition and compute the wavelet energy (square sum of wavelet coefficients around 5×5 window) of the real-part at each level and use this as feature vector. These parameters were empirically determined to attain highest accuracies for recognition experiments presented in this chapter.

12.4.2 Local Feature Extraction

LBP captures the local level texture variations. Local binary patterns introduced by Ojala et al. [18] use local texture descriptor. In its simplest form, an LBP description of a pixel is created by thresholding the values of the 3×3 neighborhood of the pixel against the central pixel and explicating the result as a binary number. The Local binary pattern (LBP) operator was originally designed as a texture descriptor. The LBP operator attributes a label to every pixel of an image by thresholding the 3×3 neighborhood of each and every pixel value with the center pixel value and assigns binary value (0,1) based on the following equation,

$$I(x, y) = \begin{cases} 0 & \text{if } N(x, y) < I(x, y) \\ 1 & \text{else } N(x, y) \geq I(x, y) \end{cases}$$

where $I(x, y)$ is the center pixel value and $N(x, y)$ is neighborhood pixel value. After thresholding, central pixel value is represented by a binary number (or decimal number) called label. Histogram of these labels is used as texture descriptor.

LBP operator is extended to scale invariance and rotation invariance texture operator for images. LBP operator deals with textures at different scales using neighborhoods of different sizes. Local neighborhood can be defined using circular neighborhood. Circular neighborhood is a set of evenly spaced sampling points on a circle, whose center is the pixel to be labeled. Radius of circle controls the spatial resolution of operator and number of sampling points controls angular space quantization. Interpolation is used when a sampling point does not fall in the mediate of a pixel. Notation (P; R) will be used for pixel neighborhoods which contemplates P sampling points on a circle of radius of R.

Figures 12.5, 12.6 and 12.7 show the extracted LBP features of fingerprint, palmprint and iris respectively. To achieve the gray level invariance we subtract the center pixel value with all circular neighborhood pixel values and assume that this difference is independent of center pixel value.

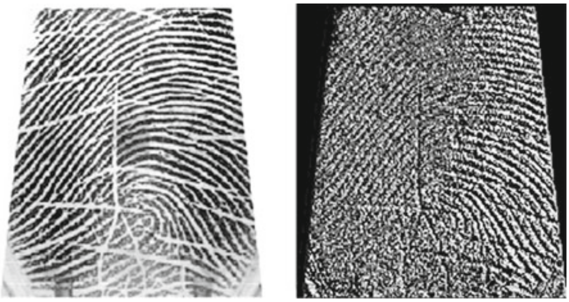


Fig. 12.5 (*left*) Original fingerprint image. (*right*) Computed LBP features on the original image

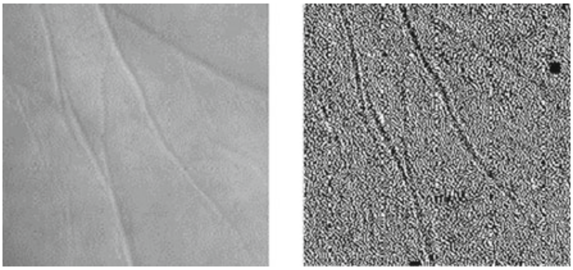


Fig. 12.6 (*left*) Extracted 128×128 palm ROI. (*right*) LBP features on the ROI

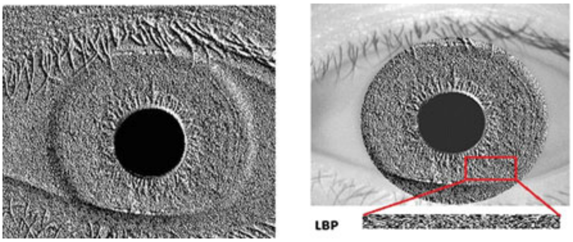


Fig. 12.7 (*left*) LBP features on the whole eye image (*right*) Computed LBP features on segmented Iris region

12.4.3 Feature Fusion

Canonical correlation analysis can be defined as the complication of finding two sets of basis vectors, one for **x** and one for **y**, in a way that the correlations between the projections of the variables onto these basis vectors are mutually maximized. Linear combinations $x = \mathbf{x}^T \hat{\mathbf{w}}_x$ and $y = \mathbf{y}^T \hat{\mathbf{w}}_y$ of the two variables is maximized as follows,

$$\rho = \frac{E[xy]}{\sqrt{E[x^2]E[y^2]}} = \frac{E[\hat{\mathbf{w}}_{\mathbf{x}}^{\mathbf{T}}\mathbf{x}\mathbf{y}^{\mathbf{T}}\hat{\mathbf{w}}_{\mathbf{y}}]}{\sqrt{E[\hat{\mathbf{w}}_{\mathbf{x}}^{\mathbf{T}}\mathbf{x}\mathbf{x}^{\mathbf{T}}\hat{\mathbf{w}}_{\mathbf{x}}]E[\hat{\mathbf{w}}_{\mathbf{y}}^{\mathbf{T}}\mathbf{y}\mathbf{y}^{\mathbf{T}}\hat{\mathbf{w}}_{\mathbf{y}}]}} = \frac{\mathbf{w}_x^T \mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x^T \mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y^T \mathbf{C}_{yy}\mathbf{w}_y}} \tag{12.6}$$

The maximum of ρ with respect to $\mathbf{w_x}$ and $\mathbf{w_y}$ is the maximum canonical correlation.

$$\begin{cases} E[x_i x_j] = E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{x}^T \mathbf{w}_{xj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xx} \mathbf{w}_{xj} = 0 \\ E[y_i y_j] = E[\mathbf{w}_{yi}^T \mathbf{y}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{yi}^T \mathbf{C}_{yy} \mathbf{w}_{yj} = 0 \\ E[x_i y_j] = E[\mathbf{w}_{xi}^T \mathbf{x}\mathbf{y}^T \mathbf{w}_{yj}] = \mathbf{w}_{xi}^T \mathbf{C}_{xy} \mathbf{w}_{yj} = 0 \end{cases}$$

The projections onto $\mathbf{w}_x$ and $\mathbf{w}_y$, i.e. x and y, are called canonical variates.

$$a^2 + b^2 = c^2 \tag{12.7}$$

The covariance matrix between two random variables $\mathbf{x}$ and $\mathbf{y}$ with zero mean is defined as.

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} = E\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \right] \tag{12.8}$$

where $\mathbf{C}$ is a block matrix where $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$ are the within-sets covariance matrices of $\mathbf{x}$ and $\mathbf{y}$ respectively and $\mathbf{C}_{xy} = \mathbf{C}_{yx}^T$ is the between-sets covariance matrix. The canonical correlations between $\mathbf{x}$ and $\mathbf{y}$ can be found by solving the eigenvalue equations

$$\begin{cases} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho^2 \hat{\mathbf{w}}_x \\ \mathbf{C}_{yy}^{-1} \mathbf{C}_{yx} \mathbf{C}_{xx}^{-1} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \rho^2 \hat{\mathbf{w}}_y \end{cases} \tag{12.9}$$

where the eigenvalues ρ^2 are the squared canonical correlations and the eigenvectors $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are the normalized canonical correlation basis vectors. The number of non-zero solutions to these equations are limited to the smallest dimensionality of $\mathbf{x}$ and $\mathbf{y}$.

Just one of the eigenvalue equations needs to be solved since the solutions are related by

$$\begin{cases} \mathbf{C}_{xy} \hat{\mathbf{w}}_y = \rho \lambda_x \mathbf{C}_{xx} \hat{\mathbf{w}}_x \\ \mathbf{C}_{yx} \hat{\mathbf{w}}_x = \rho \lambda_y \mathbf{C}_{yy} \hat{\mathbf{w}}_y, \end{cases} \tag{12.10}$$

where

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{\hat{\mathbf{w}}_y^T \mathbf{C}_{yy} \hat{\mathbf{w}}_y}{\hat{\mathbf{w}}_x^T \mathbf{C}_{xx} \hat{\mathbf{w}}_x}}. \tag{12.11}$$

As discussed below, we apply method proposed by [22] to combine the output of DTCW and LBP and maximize the information present in these two feature vectors.

Let the two feature extractors be trained by L training images. Let $A = [a_1, a_2, ..., a_L]$ and $B = [b_1, b_2, ..., b_L]$ be the corresponding outputs of the two extractors, and n1 and n2 be the dimensions of the two outputs, where $n1, n2 \leq L1$.

The covariance matrices for $\mathbf{A}$ and $\mathbf{B}$ are given as $\mathbf{C}_{aa}$ and $\mathbf{C}_{bb}$ respectively. $\mathbf{C}_{ab}$ is the between-set covariance matrix. $\hat{\mathbf{w}}_x$ and $\hat{\mathbf{w}}_y$ are canonical basis vectors of feature vectors $\mathbf{A}$ and $\mathbf{B}$. a_i and b_i are two feature vectors of image i. Fusion of these two feature vectors is defined as,

Table 12.1 Statistics of individual Fingerprint, Palmprint and Iris recognition systems analyzed on CASIA [24] databases

Sl. No.	Biometric trait	Accuracy (%)	EER (%)
1	Fingerprint	99.1	1.0
2	Palmprint	97.2	3.2
3	Iris	98.2	1.8

$$\mathbf{F}_i = \begin{bmatrix} \hat{\mathbf{w}}_x^T a_i \\ \hat{\mathbf{w}}_y^T b_i \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{w}}_x & 0 \\ 0 & \hat{\mathbf{w}}_y \end{bmatrix}^T \begin{bmatrix} a_i \\ b_i \end{bmatrix}. \tag{12.12}$$

12.4.4 Feature Matching

For the matching purpose we use cosine similarity measure. Cosine similarity measure is defined as cosine angle between test image fused feature vector and training images fused feature vector,

$$\arg max_{j\in[1,2\cdots L]} \left(\frac{\mathbf{F}_i^T \mathbf{F}_j}{\|\mathbf{F}_i\| \cdot \|\mathbf{F}_j\|} \right) \tag{12.13}$$

The maximum value according to Eq. (12.13) is estimated as an authentic FPI match. The accuracies and error rates of individual biometric traits used in this work are shown in Table 12.1. Figure 12.8 depicts the ROC curves drawn against the False Acceptance Rate and False Rejection Rate of fingerprint, palmprint and iris recognition systems.

12.5 Fusion of Multiple Biometric Traits: Fingerprint, Palmprint, Iris

Although fusion of multiple representations of a single biometric trait has preference over the rest, it still suffered from the drawbacks possessed by a unimodal recognition system [14] where as fusion of information from multiple modalities is gaining utmost attention for the reasons stated in Sect. 1.3.4. The proposed fusion scheme is inspired by the idea that a combination of uncorrelated modalities (like, fingerprint and palmprint, fingerprint and iris, etc.,) is expected to result in improvement in performance than a combination of correlated modalities (different impressions of the same finger, different representations of same image, etc.) [17].

Different architectures can be used to integrate information from multiple biometric traits. Architecture of a multi-biometric recognition system is the way in which individual biometric traits are processed to converge onto a single identity. These ways are generally progressed in **serial or cascade** mode, **parallel** mode and **hierarchical** mode. In the serial/cascade mode, different biometric traits are processed one after the other and so are inter-dependent while in parallel mode, all

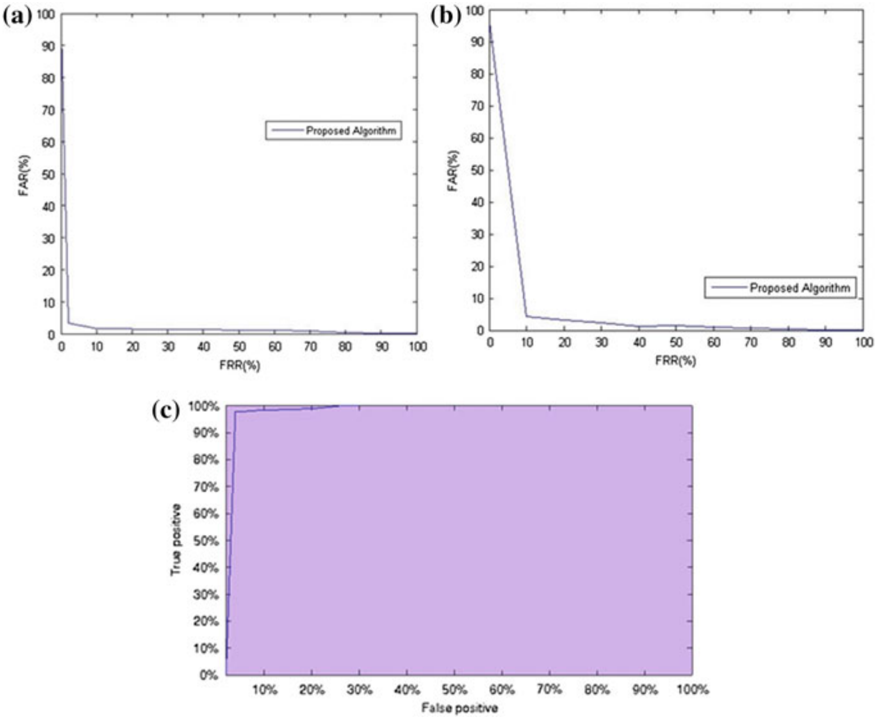


Fig. 12.8 ROC for fusion of multiple representations of **a** fingerprint **b** palmprint **c** iris biometric traits on CASIA [24] databases

biometric traits are independently processed and the results are combined to make a final authentication decision. Both these architectures have points in favor of and against. Cascade architecture is preferred specially in applications involving searching large scale databases in contrast to parallel architecture which is relatively heavy. Here, if a confident authentication decision can be made using first modality, then the user might not be asked to provide another modality which saves time. Also, in case a user is crippled of one biometric modality, he/she can choose to authenticate using another biometric modality. Example, in a multimodal cascade architecture that uses palmprint and iris modalities; if a user looses his hands in an accident after enrollment, he can still choose to authenticate using iris modality. Cascade architectures are usually preferred in less security-demanding civil applications where processing time is of the essence. In [15] Hong and Jain proposed a cascaded multibiometric system using face and fingerprint modalities at two stages.

On the other hand, parallel architectures evidently provide higher accuracies as they combine more modalities and evidences about the user for recognition. They are preferred in high security-demanding defense/military applications. In [16] Ross and Jain proposed a parallel architecture using face, fingerprint and hand geometry. Parallel architectures have the advantage that, if one modality is degraded

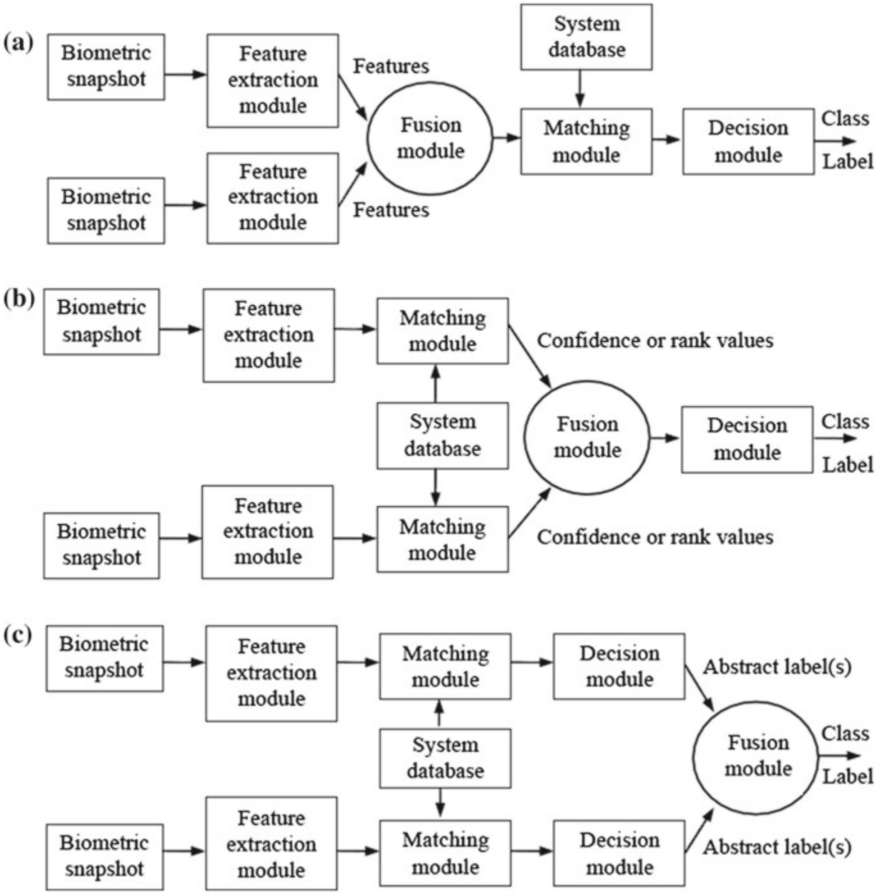


Fig. 12.9 Different levels of fusion in a parallel fusion mode: **a** fusion at the feature extraction level; **b** fusion at matching score (*confidence or rank*) level; **c** fusion at decision (*abstract label*) level. Courtesy [17]

by introduction of noise due to the capturing device, other modality compensates it as authentication decision is finally made by fusing multiple modalities. This fusion can generally be performed at four levels, illustration of these levels of fusion can be found from Fig. 12.9 viz.,

1. Fusion at sensor level,
2. Fusion at feature extraction level,
3. Fusion at decision level and rank level,
4. Fusion at matching score level.

Sensor level fusion is the fusion which takes place in the initial stages of the recognition system, that is, just after the image acquisition. Acquired images are combined in some way to create a new input template for further processing. This is possible only if these images are instances of a same biometric trait and are acquired using a single sensor or more than one compatible sensors. Examples include registration of different images of a single object, which is the task of spatially aligning a pair of images of the same biometric trait acquired from different compatible sources, viewpoints and at different times. Mosaicking is also another example of sensor level fusion which is explored by [26, 27] on fingerprint images. Fusion at this level is generally not preferred for biometric authentication systems as acquiring data only from compatible devices and knowing correspondence points for mosaicking in prior are difficult in terms of time and cost.

Feature level fusion is fusion of feature vectors that are obtained (i) by applying multiple algorithms on the same biometric image (e.g, using Wavelets and Local Binary Pattern algorithms to acquire different kinds of information from same image), (ii) from multiple images of same biometric trait obtained over time, (iii) multiple units of same biometric trait (e.g, different face images obtained at different view points, images of five fingers of a person's hand) (iv) by using multiple acquisition sensors to obtain the same biometric trait (e.g, using both infra-red camera and visible-light camera to obtain images of a palm) (v) from multiple biometric traits (e.g, fusion of features from palmprint and fingerprint). Although fusion of individual feature sets corresponding to multiple biometric traits has advantages, it is difficult to achieve because of the following reasons,

- procedures to extract the feature sets might differ giving rise to incompatible feature vectors.
- the correlation among the feature spaces of biometric modalities might be unidentified (e.g., feature vectors of minutiae points and eigen coefficients extracted from fingerprints).
- curse-of-dimensionality problems emanate from unusual increment in length of the feature vectors during fusion which might lead to escalated matching time.

Because of these problems, not much research has gone into exploring this profitable level of fusion. Few attempts in literature [28–30] to combine palmprint and hand geometry features; face and hand geometry features; fingerprint and palmprint features respectively gained only limited success. Sharma and Kaur [35] uses multiclass support vector machines for classification after combining the feature vectors from face, fingerprint and palmprint extracted using principal component analysis. As a feature set is a straightforward representation of the raw biometric data, it is theoretically presumed to incorporate richer information. So, integration at the feature-level apparently provides better authentication than integration at score-level or decision-level.

Decision level fusion takes place when multiple biometric systems process multiple biometric traits and are asked to arrive at an authentication decision independently. Decision in such systems is generally made based on majority. Methods like majority voting [31], AND/OR rules [32], etc., were proposed in literature to consolidate the decisions given by individual systems. This level of fusion is also termed as abstract fusion as here the decision is made depending on the abstract or incomprehensible information rendered by individual systems.

In contrary to this, if a rank is given to the output obtained from each independent unimodal authentication system and these ranks are consolidated using some rules like borda count, logistic regression [33] to make a final decision to identify an individual, such level of fusion is termed as **Rank level fusion**.

Score-level fusion is combination of different match scores given from individual biometric traits considered in that multimodal framework. Matching score from every biometric trait is the representation of its proximity to the identity of an individual. There are two types of score-level fusion.

Score-level fusion can be posed as classification problem in the context of biometric verification. Individual scores obtained from each biometric traits are combined to form a feature vector. This feature vector is classified as "accept" or "reject" by the classifier. Here no prior processing is required before invoking the classifier. Work involving classifiers like decision trees [16] to combine scores from face, fingerprint, hand geometry modalities; Abbas et al. [34] uses one-class support vector machine classifier and compares it with the standard two-class SVM classifier for classification of face and fingerprint score vectors; were profitably applied to various multimodal systems.

Score-level fusion can also be posed as combination problem where match scores from individual modalities are combined using some rules to arrive at a final scalar score for identification decision. This kind of fusion is quite worthy but requires normalization of individual match scores before fusion for homogeneity. Shariatmadar and Faez [36] presents Finger-Knuckle-Print recognition algorithm based on multi-instance fusion at the matching score level. Before fusing, a novel normalization strategy is applied on each score and a fused score is generated for the final decision by summing the normalized scores. Yuanyuan et al. [37] uses five score-level fusion rules to combine three different shape representations of gait. Yu et al. [38] proposes a method to authenticate individuals based on score-level fusion of multi-scale and multi-resolution palm print images using simple SUM and MAX rules. Eskandari and önsen [39] focuses on combining the strengths of face and iris modalities by employing particle swarm optimization (PSO) to select facial features and uses Weighted-Sum Rule for fusion of individual scores. In this work, we propose an incomparable score-level fusion scheme based on Equal Error Rates of individual fingerprint, palmprint and iris modalities. No prior work has been done in literature by combining these three modalities. Reasons for choosing a combination of fingerprint, palmprint and iris has been well elucidated in Sect. 12.2.2.

To take advantage of both serial and parallel architectures, it is possible to design a hierarchical/tree-like architecture. This architecture not only is robust, fast and highly secure but also mitigates problems like missing and noisy data associated with parallel and serial architectures respectively. To the authors' knowledge, not much study has gone into exploring this profitable architecture.

In this work, we propose two multi-biometric fusion frameworks based on parallel and hierarchical architectures. To take advantage of every biometric trait considered in this work viz., fingerprint, palmprint and iris; we initially combine different features extracted at global level and local level from individual biometric traits and

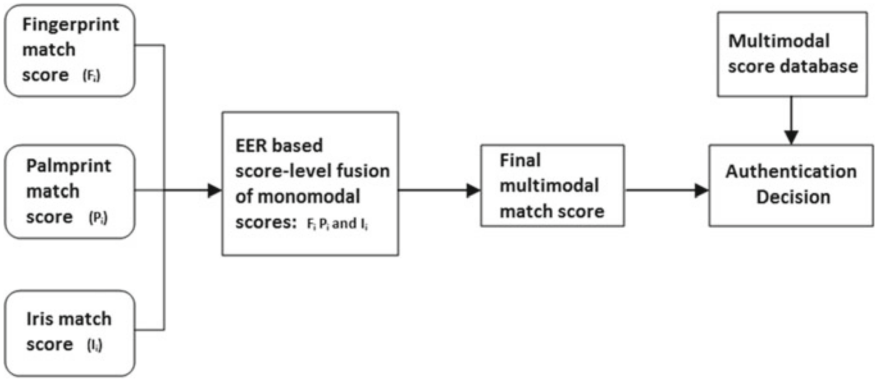

Fingerprint match score (F_i)
Palmprint match score (P_i)
Iris match score (I_i)
EER based score-level fusion of monomodal scores: F_i P_i and I_i
Final multimodal match score
Multimodal score database
Authentication Decision

Fig. 12.10 Block diagram for the proposed parallel multimodal score-level fusion recognition system

eventually propose two multi-modal recognition architectures employing different fusion techniques by,

1. Combining the match scores of each biometric using two score-level fusion rules—*parallel architecture*, Sect. 12.5.1.
2. Adopting a two stage cascade framework using feature-level fusion—*hierarchical architecture*, Sect. 12.5.2.

12.5.1 Parallel Architecture Based Fusion

The block diagram for the proposed method is shown in Fig. 12.10. The match scores obtained from fusion of local and global representations of individual biometric traits viz., fingerprint, palmprint and iris of an individual are combined using two score-level fusion techniques which rely on Equal Error Rates of individual traits, viz.,

i. Simple Matcher Weighting fusion technique.
ii. Proposed score-level fusion rule.

Reasons for proposing fusion techniques based on individual error rates are, (a) these index measures are simple and direct in terms of interpretation as compared to the ROC; (b) EER at once characterizes both Genuine Accepts and False Accepts; Genuine Rejects and False Rejects; and (c) the EER is based on a projected optimal operating point (of total error rate, TER) where the FAR curve meets the FRR curve [40].

The relevance of making a decision based on a simple score fusion is worthy as each unimodal recognition scheme proposed in this work (see Sect. 12.4) has its own leverage. Determining the most productive fusion technique for score-level fusion depends on functional issues like availability of data, performance requirements and validity of inferred assumptions.

12.5.1.1 Matcher Weighting Fusion Technique

Individual matching scores of fingerprint, F_i; palmprint, P_i and iris, I_i are combined using Matcher Weighting fusion technique [25]. This technique takes advantage of the Equal Error Rate (EER) values obtained from recognition using individual biometric traits. Here, individual match scores are multiplied with a weighting factor which is proportional to the inverse of its EER. Then these weighted scores are simply summed to obtain a final match score. As depicted in Eq. (12.15), the match score corresponding to the biometric trait with less EER attains higher weightage thus assuring higher precision.

$$M_i = \sum_{m=1}^{M} w^m x^m = \sum_{m=1}^{M} \frac{1}{e^m} x^m \tag{12.14}$$

where M_i is the final matching score of the proposed multimodal recognition system, w^m and e^m are respectively the weighting factor and EER of the mth biometric, x^m and M is the number of biometric traits considered.

According to this technique,

$$M_i = \frac{1}{EER_{F.P}} F_i + \frac{1}{EER_{P.P}} P_i + \frac{1}{EER_{Iris}} I_i \tag{12.15}$$

$$M_i = \frac{1}{1.0} F_i + \frac{1}{3.2} P_i + \frac{1}{1.8} I_i \tag{12.16}$$

where F_i, P_i and I_i are the individual match scores corresponding to fingerprint, palmprint and iris respectively.

As envisaged, the above linear score-level fusion approach has demonstrated a notable performance gain over unimodal systems. Also decrease in error rates has been reported on database 1, refer Sect. 12.6 for the description of databases used. As is evident from Fig. 12.11, the proposed system performs at an accuracy of 99.3 % and an EER of 0.75 %.

12.5.1.2 Proposed Score-Level Fusion Rule

The proposed rule addresses the fusion problem from error rate minimization point of view so as to increase the decisive efficiency of the fusion system. It is a linear, data-driven approach to optimize the cumulative fusion scores with respect to individual error rates and is inspired from one of the measures of central tendency which is harmonic mean as it gives less significance to high-value outliers providing a truer picture of the average.

To give further emphasis to the modality with low error rates, we propose a new weighting approach. Error rate is a percentage of misclassified samples. We attempt to minimize the overlapped area between genuine and imposter curves by choosing to

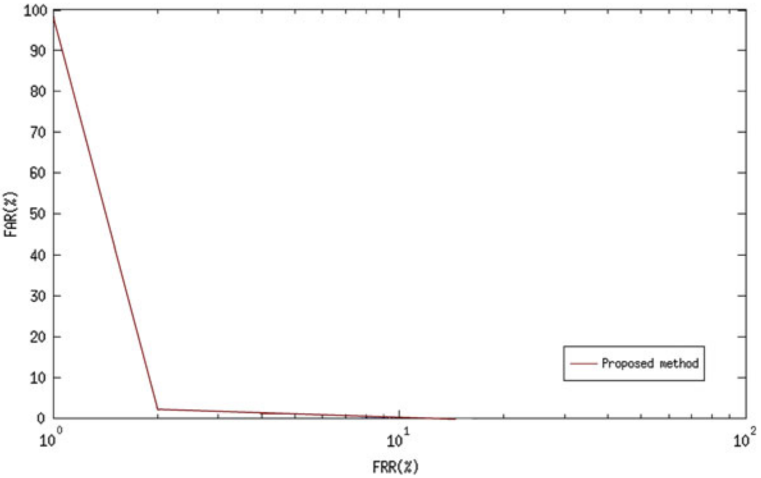


Fig. 12.11 ROC for the conventional score-level fusion technique: Matcher weighing

maximize the stability of the modality with superior performance. Hence minimizing the chances of false authentication using the following combination,

$$M_i = \frac{\frac{1}{EER_{F.P}} F_i}{\frac{1}{EER_{P.P}} P_i + \frac{1}{EER_{Iris}} I_i} \tag{12.17}$$

$$M_i = \frac{\frac{1}{1.0} F_i}{\frac{1}{3.2} P_i + \frac{1}{1.8} I_i} \tag{12.18}$$

The proposed deterministic rule to deal with fusion optimization problem is evaluated on database1. Our empirical evaluations show promising potential in terms of performance accuracy and error rates in comparison with the conventional matcher weighting technique. We report an accuracy of 99.48 % and an EER of 0.72 %, Fig. 12.12, Table 12.2.

12.5.2 Hierarchical-Cascading Architecture Based Fusion

Though fusion of individual scores enhances the performance, direct fusion of individual feature sets obtained from respective biometric traits is more appropriate as fusion at this level directly integrates information from diverse biometric traits which render complementary data.

So, to take the advantage of feature-level fusion, serial/cascade architecture and hierarchical architectures, we propose a two-stage cascading framework based on

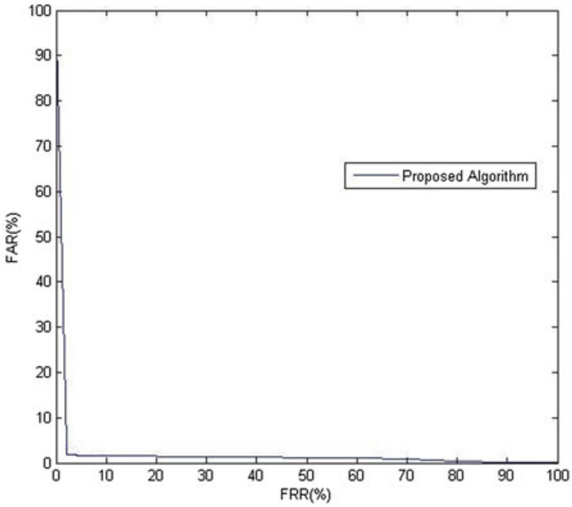


Fig. 12.12 ROC for the proposed score-level fusion technique

Table 12.2 Performance of the two proposed fusion frameworks

Proposed fusion framework	Accuracy (%)	EER (%)
Parallel framework using score-level fusion rule1	99.3	0.75
Parallel framework using score-level fusion rule2	99.48	0.72
Hierarchical-cascade framework	99.5	0.78

fusion of fingerprint and palmprint feature sets in the first stage and iris features to eliminate the ambiguity of false matches in the next stage, refer Fig. 12.13. The proposed frame work takes advantage of both unimodal and multimodal architectures.

In the first stage, canonical correlation based feature vectors, say F_v and P_v, extracted from fingerprint and palmprint images of an individual are combined by simple concatenation. This combined feature vector is compared with the template database comprising of the enrolled concatenated finger and palmprint data, using cosine similarity measure. While P_v is thorough at a length of 100, F_v required a length of 125 to attain maximum recognition capability.

The matching algorithm proceeds as follows. In the first stage, a similarity score is assigned to the comparison. As the proposed system uses cosine similarity measure for comparison, the similarity score falls in the range (-1 to 1). As this value goes higher, the images in comparison are estimated to be more similar. Consequently, genuine or imposter decision is made using a decision threshold value (DT). Generally, if the matching score is higher than the DT, the comparison is termed as genuine and vice versa.

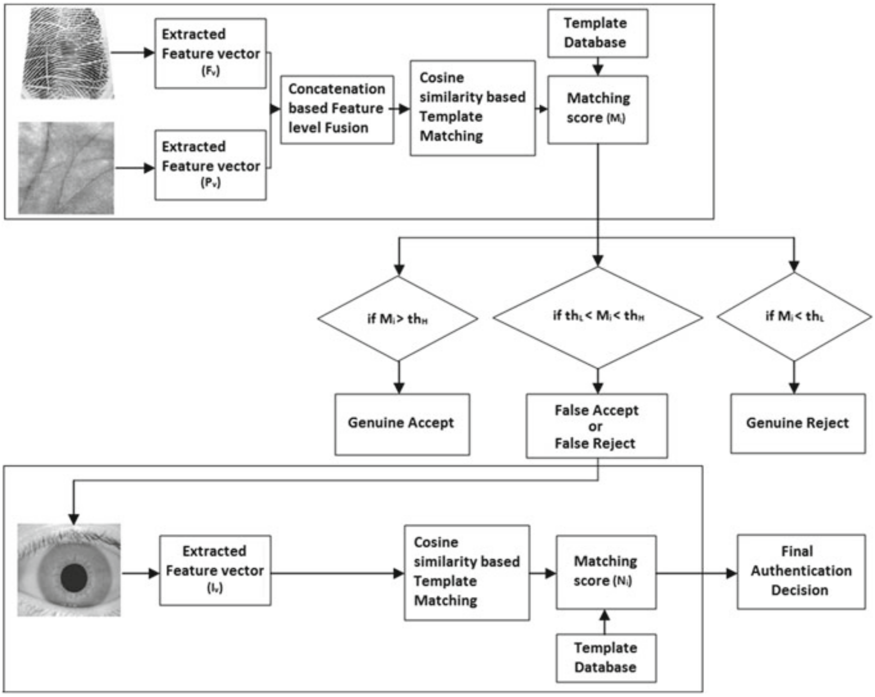

Extracted Feature vector (Fᵥ)
Extracted Feature vector (Pᵥ)
Concatenation based Feature level Fusion
Cosine similarity based Template Matching
Template Database
Matching score (Mᵢ)
if Mᵢ > thH
if thL < Mᵢ < thH
if Mᵢ < thL
Genuine Accept
False Accept or False Reject
Genuine Reject
Extracted Feature vector (Iᵥ)
Cosine similarity based Template Matching
Matching score (Nᵢ)
Template Database
Final Authentication Decision

Fig. 12.13 Block diagram for the proposed multimodal hierarchical-cascading scheme based recognition system

Contrary to this, the proposed method, considers a lower threshold value, say th_L and upper threshold value, say th_H. These values are empirically decided for the available data considered in this work. As shown in Fig. 12.13,

i if the match score is greater than the upper threshold, then the comparison is termed as a genuine match with zero probability of error, $M_i > th_H$.
ii if the match score is lower than the lower threshold, then the comparison is termed as imposter match with zero probability of error, $M_i < th_L$.
iii otherwise, if the score is a value between lower threshold and upper threshold, $th_L < M_i < th_H$, decision can not be made with certainty. Chances of mistaken classification exist. This case of uncertainty arises from the large overlap between genuine and imposter match scores. Such comparison, say MF_i, can be termed as either false accept or false reject. Such comparisons with a match score between th_L and th_H are directed to the next stage to converge onto a single identity.

Figure 12.14 shows the genuine and imposter score distribution functions for cosine similarity matching algorithm in stage 1. These distributions can not be accurately formulated by a known statistical model so are estimated from empirical data. These curves are drawn against the match scores derived from cosine similarity matching and their percentage of frequency. From Fig. 12.14 it is evident that the

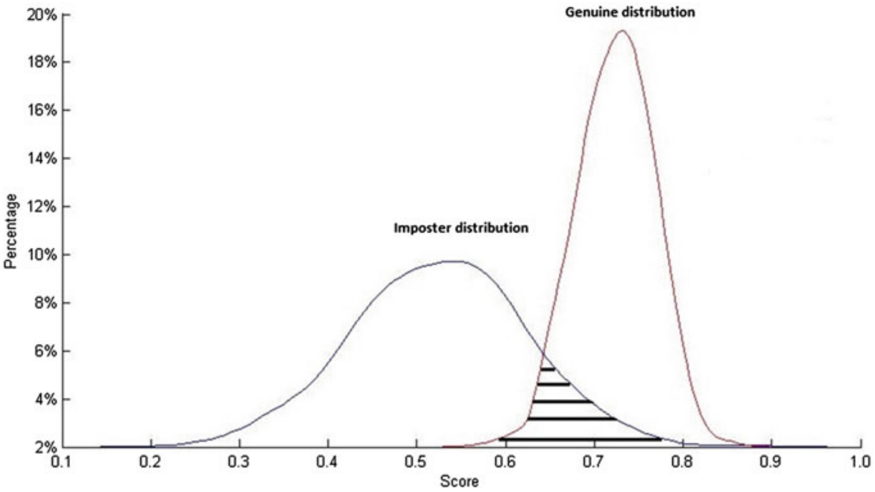


Fig. 12.14 Genuine and Imposter score distributions of authentication based on combined fingerprint and palmprint features in the first stage

overlap between the genuine and imposter score distributions in stage 1 is large, thus ascertaining the existence of false matches. These parameters were empirically determined to attain highest accuracies on database 2 (refer Sect. 12.6 for the description of databases) for recognition experiments presented in this paper.

In the second stage, corresponding iris features of those templates with match score MF_i are extracted and compared with the enrolled iris template database using cosine similarity measure. Score distributions for iris biometric trait considerably lower but do not eliminate the overlap between the genuine and imposter score functions. Thus a satisfying decision threshold, th_{DT} is decided. Consequently, the match scores greater than the decision threshold, $MF_i > th_{DT}$ are classified as genuine and those lower than the decision threshold, $MF_i < th_{DT}$ are classified as imposter.

The concatenated feature vector in the first stage minimizes the curse-of-dimensionality problem and since the procedures to extract fingerprint feature set and palmprint feature set are the same, incompatibility issues for feature-level fusion stated above do not arise. Figure 12.15 shows the false acceptance and false rejection rates of the proposed system. The accuracy of this multimodal recognition framework is estimated as 99.5 % and Equal Error Rate (EER) as 0.78 % which is a considerable performance gain over individual Unimodal systems as reported in Fig. 12.8 and Table 12.1.

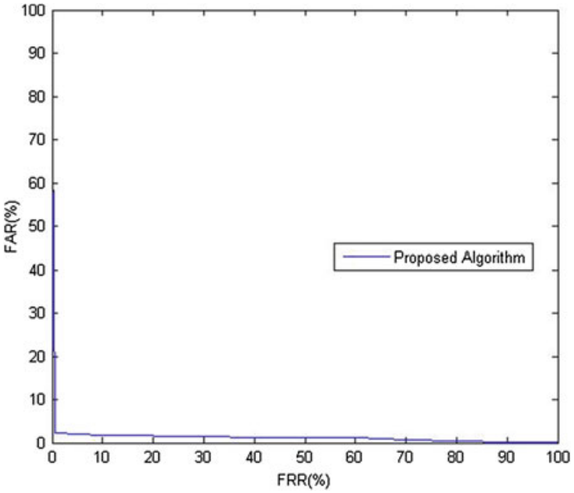


Fig. 12.15 ROC curve for the proposed Hierarchical-cascading scheme based multi-modal biometric recognition system

Table 12.3 Statistics of database 1

Biometric trait	No. of subjects	No. of samples per subject	Total no. of images
Fingerprint	300	20	6,000
Palmprint	300	16	4,800
Iris	300	40	12,000

12.6 Description of Databases

Database 1

It is a virtual multimodal dataset comprising of fingerprint, palmprint and iris images from 300 subjects. Statistics of database 1 can be seen from Table 12.3. This attempt is compelled by the reason that no large publicly available multimodal dataset exists for the considered biometric traits. However, [23] validated the use of virtual datasets in multimodal environments and concluded that the variation in performance between virtual and real datasets was not statistically significant. This is possibly appropriate especially when the considered modalities are uncorrelated and independent. Virtual records are created by pairing a user from one unimodal database with a user from another database [14]. Database 1 is created by integrating CASIA fingerprint V5, CASIA palmprint V1, and CASIA Iris V3 databases [24].

Database 2

It is a real multimodal dataset comprising of fingerprint, palmprint and iris images from 50 subjects. The reason for collecting this dataset is to validate the proposed method on contact-free, low-resolution images. Fingerprint and palmprint images were acquired using a basic 1.5 megapixel web-camera with image size 640 × 480 and both X and Y resolutions 72 pixels/in. Subjects were absolutely not restricted

Table 12.4 Statistics of database 2

Biometric trait	No. of subjects	No. of samples per subject	Total no. of images
Fingerprint	50	10	500
Palmprint	50	8	400
Iris	50	8	400

of hand distance from the camera during the acquisition process thus retaining maximum scale, rotation, projection and translation variance in the images. Similarly, iris images of size 640x480 were collected in single session using a hand-held iris sensor. Statistics of database 2 can be seen from Table 12.4. From the results in Sect. 12.6, it is evident that the proposed multimodal cascade framework employing fingerprint, palmprint and iris biometric traits is suitable for a mobile, touch-less biometric recognition environment.

12.7 Discussion and Future-Work

Present work is largely intended to mitigate problems associated with unimodal biometric recognition systems. To accomplish this, many multimodal authentication systems have been proposed in literature each involving different modalities, different fusion mechanisms, different architectures and different levels of fusion. There are pros and cons to every fusion mechanism which proves that the design of a multimodal biometric system predominantly depends on the application criteria along with factors like location, task to be accomplished (identification or verification), population coverage, security hazards, user circumstances, existing data, etc.

According to our study, as the two fusion techniques considered in this work, viz., Fusion of multiple representations of a single biometric trait and Fusion of multiple biometric traits, catered to the needs of modern day recognition applications we explore these fusion mechanisms with respect to three widely used biometric modalities viz., Fingerprints, Palmprints and Iris. The review of biometric based recognition systems indicate that a number of factors including the accuracy, cost, and speed of the system may play vital role in assessing its performance. But today with the cost of biometric sensors constantly diminishing and high speed processors and parallel programming techniques widely available to affordable research, performance accuracy has become predominant focus of biometric system design [54]. The main aim of this thesis is to improve the accuracy of a biometric recognition system by reducing error rates, that is, false accepts and false rejects.

As an inception in literature, we propose a complete contact-free multi-modal recognition algorithm which works significantly on low-resolution images involving fusion of Fingerprints, Palmprints and Iris biometric traits. Reasons for preferring a combination of these three traits over the rest has been clearly elucidated in Sect. 12.2.2. A proprietary real multimodal database is established comprising of fingerprint, palmprint and iris images from 50 subjects. The reason for collecting this dataset is to validate the proposed method on real, contact-free, low-resolution images, refer Table 12.4.

Fusion of multiple biometric traits is realized using two frameworks, viz., parallel and hierarchical. Both serial and parallel multibiometric architectures have flaws and advantages. Multi-biometric recognition systems designed with hierarchical architecture not only are robust, fast and highly secure but also mitigate problems like missing and noisy data associated with parallel and serial architectures respectively. We explore both parallel and hierarchical multimodal architectures using fingerprint, palmprint and iris modalities. Also, a new score-level fusion rule based on individual error rates has been proposed in this work. The proposed rule addresses the fusion problem from error rate minimization point of view so as to increase the decisive efficiency of the fusion system. We compare it's efficiency with widely used Matcher weighting score-level fusion rule to report surpassed results.

Feature-level fusion though is an understudied problem, our study on it indicates that as feature set is a straightforward representation of the raw biometric data, it is theoretically presumed to incorporate richer information. So, integration at the feature-level apparently provides better authentication than integration at score-level and decision-level. All these factors inspired us to propose a hybrid approach based on fusion of local and global texture features. The proposed method maximizes the information from two feature vectors of the same pattern by combining them at the descriptor level which ensures that the information captured from both the features are maximally correlated and eliminates the redundant information giving a more compact representation. However, it is difficult to anticipate the perfect fusion strategy for a given scenario. Also, the biometric recognition methods proposed in this paper do not support the integration of incompatible feature sets as normalization of either feature-vectors or match-scores is not studied here.

There is adequate scope for improvement relating to human authentication systems and protocols employing biometrics. Main problems include noise, privacy issues and spoofing.

As we are interested in working on contact-free biometric systems which work well with even low-resolution images captured in unconstrained environments, we would like to investigate the use of periocular region in this respect as this biometric trait can be captured from a normal CCD camera in visible light focusing from longer distances in comparison with other ocular biometric traits which require well equipped acquisition setup and different frequencies of electronic spectrum. Unsang et al. [43] proves that features in periocular region are best used as a complementing biometric than as a full-grown individual recognition system. Hence we intend to investigate the use of periocular features in combination with iris features for better performance than [43]. Our study presumes that traits related to ocular region of the human body are data rich. Another biometric that caught plenty of research interest in the recent time is Complex Eye Movements [4, 5] which is rather a behavioral biometric trait than a physiological one. In [4] Complex Eye Movements are combined with Oculomotor Plant Characteristics where a mathematical model for eye and its associated muscle movement is established when eyes respond to a stimuli.

Similarly research on using physiological signals like ECG [41] as a complementary biometric trait in multimodal systems is another aspect of interest. Recently, researchers at MITs Artificial Intelligence lab have shown how signals of heart rate variability can be captured from the minute head movements which can be captured by a smart phone camera by detecting head motion artifact associated with each beat of the heart [42]. More detailed study can go into using these physiological signals as biometric traits to ensure security. Such systems also ensure privacy to an extent unlike systems that capture physical traits of human body like palmprint, face, ear, etc.,.

The enrollment time and the failure to enroll (FTE) rate can be significantly mitigated by designing multimodal biometric systems which ensure user convenience i.e., systems that are designed to capture multiple traits simultaneously, for example, multimodal system involving face, periocular region and iris as biometric traits.

A further prevailing problem that hinders accurate recognition and escalates error rates is noise in the acquired sample induced either because of a noisy background or a degraded sensor. This can be mitigated by designing multimodal systems with serial/cascade architecture in which multiple traits can substitute each other, for example, combining voice and face modalities, because while acquiring the voice sample if background noises dominate then recognition can still be accomplished using face modality. Also, there is scope for improvement in regard to fusion techniques that operate on data before matching as such data is presumed to incorporate richer information. Also, adapting the system to multiple possible variations of a biometric trait by providing it with multi-session data can largely help in reducing false rejections.

References

1. Bolle, R.M., Connell, J.H., Pankanti, S., Ratha, N.K., Senior, A.W.: Guide To Biometrics. Springer, Berlin (2003)
2. Body Odor Recognition, http://biometrics.pbworks.com/w/page/14811351/Authentication20technologies
3. Korotkaya, Z.: Biometric person authentication: Odor, pp. 1–6
4. Komogortsev, O.V., Karpov, A., Price, L.R.: Biometric authentication via oculomotor plant characteristics, In: IARP/IEEE International Conference on Biometrics, March 29–April 1, 2012 New Delhi, India
5. Kasprowski, P., Ober, J.: Eye movement in biometrics. In: Proceedings of Biometric Authentication Workshop, European Conference on Computer Vision in Prague 2004, LNCS 3087, Springer-Verlag 2004
6. Chellappa, R., Wilson, C.L., Sirohey, S.: Human and machine recognition of faces: a survey. Proc. IEEE **83**, 5 (1995)
7. Bruce, P., Ashraful Amin, M., Hong Y.: Performance evaluation and comparison of PCA Based human face recognition methods for distorted images. Int. J. Mach. Learn. Cybern. **2**, 245–259 (2011)
8. Tistarelli, M., Bigun, J., Grosso, E. (eds.): Biometrics school 2003, LNCS 3161, 1942 (2005)

9. Wu, X., Wang, K., Zhang, D.: Line feature extraction and matching in palmprint. In: Proceedings of the Second International Conference on Image and Graphics, pp. 583–590 (2002)
10. Nalwa, V.S.: Automatic on-line signature verification. In: Proceedings of the IEEE Transactions on Biometrics, 85(2), February 1997
11. Daugman, J.: How iris recognition works. In: IEEE Transaction Circuits System for video technology, vol. 14, no. 1, January 2004
12. Iris identification solutions, http://www.neurotechnology.com/verieye-technology.html
13. Human Identification in Information Systems, http://www.anu.edu.au/people/Roger.Clarke/DV/HumanID.html
14. Karthik, N.: Integration of multiple cues in biometric systems. MS thesis, Michigan State University (2005)
15. Hong, L., Jain, A.K.: Integrating Faces and Fingerprints for Personal Identification. IEEE Trans. Pattern Anal. Mach. Intell. **20**(12), 1295–1307 (1998)
16. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recog. Lett. Special Issue on Multimodal Biometrics **24**(13), 2115–2125 (2003)
17. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Trans. Circuits Syst. Video Technol. Special Issue on Image- and Video-Based Biometrics **14**(1), 420 (2004)
18. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Trans. PAMI **24**(7), 971–987 (2002)
19. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.C.: The dual-tree complex wavelet transform. IEEE Signal Process. Mag. **2**(2), 123–151 (2005)
20. Belcher, C., Du, Y.: Region-based SIFT approach to iris recognition. In: Optics and Lasers in Engineering, vol. 47 pp. 139–147, Elsevier (2009)
21. Park, C.H., Lee, J.J.: Extracting and combining multimodal directional iris features. vol. 3832, pp. 389–396. Springer, Heidelberg (2005)
22. Quan-Sen, S., et al.: A theorem on the generalized canonical projective vectors, In: Pattern Recognition, vol. 38, pp. 449–452 (2005)
23. Indovina, M., Uludag, U., Snelick, R., Mink, A., Jain, A.K.: Multimodal biometric authentication methods: a COTS approach. In: Proceedings of workshop on multimodal user authentication, pp. 99106, Santa Barbara, USA, December 2003
24. http://www.idealtest.org/dbDetailForUser.do?id=5
25. P. Ejarque, A. Garde, J Anguita, J. Hernando; On the use of genuine-imposter statistical information for score fusion in multimodal biometrics. In, Ann. Telecommun., (2007).
26. Ross, A., Jain, A.K.: Fingerprint mosaicking. In: Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP), pp. 4064–4067. Florida, U.S.A., May 2002
27. Moon, Y.S., Yeung, H.W., Chan, K.C., Chan, S.O.: Template synthesis and image mosaicking for fingerprint registration: an experimental study. In: Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP), vol. 5, pp. 409–412. Quebec, Canada, May 2004
28. Kumar, A., Wong, D.C.M., Shen, H.C., Jain, A.K.: Personal verification using palmprint and hand geometry biometric. In: Proceedings of Fourth International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA), pp. 668–678, Guildford, U.K., June 2003
29. Ross, A., Govindarajan, R.: Feature level fusion using hand and face biometrics. In: Proceedings of of SPIE Conference on Biometric Technology for Human Identification, vol. 5779, pp. 196–204, Florida, U.S.A., March 2005
30. Krishneswari, K., Arumugam, S.: Multimodal biometrics using feature fusion. J. Comput. Sci. **8**(3), 431–435 (2012)
31. Lam, L., Suen, C.Y.: Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **27**(5), 553–568 (1997)
32. Daugman, J.: Combining Multiple Biometrics. Available at http://www.cl.cam.ac.uk/users/jgd1000/combine/combine.html

33. Radha, N., Kavitha, A.: Rank level fusion using fingerprint and iris biometrics. Indian J. Comput. Sci. Eng. **2**(6), 917–923 (2012)
34. Abbas, N., Bengherabi, M., Boutellaa, E.: Experimental investigation of OC-SVM for multibiometric score fusion. In: 8th International Workshop on systems, signal processing and their applications (WoSSPA), pp. 250–255, 12–15 May 2013
35. Sharma, P., Kaur, M.: Multimodal classification using feature level fusion and SVM. Int. J. Comput. Appl. **76**(1–16), 26–32 (2013)
36. Shariatmadar, Z.S., Faez, K.: Finger-knuckle-print recognition performance improvement via multi-instance fusion at the score level. Optik - Int. J. Light Electron Opt. **125**(3), 908–910 (2014)
37. Yuanyuan. Z., Shuming, J., Zijiang Y., Yanqing Z.: A score level fusion framework for gait-based human recognition: Multimedia Signal Processing (MMSP), In: IEEE 15th International Workshop on, pp. 189–194, Sept. 30 2013–Oct. 2 2013
38. Yu, Y., Tang, Y., Cao, J., Gan, J.: Multispectral palmprint recognition using score-level fusion. In: IEEE International Conference on and IEEE Cyber, Physical and Social Computing, pp. 1450–1453, 20–23, August 2013
39. Eskandari, M., önsen, T.: Score level fusion for face-Iris multimodal biometric system: information sciences and systems 2013. Lect. Notes Electr. Eng. **264**, 199–208 (2013)
40. Kar-Ann, T., Jaihie, K., Sangyoun, L.: Biometric scores fusion based on total error rate minimization. Pattern Recog. **41**(3), 1066–1082 (2008)
41. Labati, R.D., Sassi, R., Scotti, F.: ECG biometric recognition: Permanence analysis of QRS signals for 24 hours continuous authentication, In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 31–36, 18–21 November (2013)
42. http://www.extremetech.com/computing/159309-mit-researchers-measure-your-pulse-detect-heart-abnormalities-with-smartphone-camera
43. Unsang, P., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum: a feasibility study, In: Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Applications and systems, pp. 153–158 (2009)
44. Zhenan, S., Yunhong, W., Tieniu, T., Jiali, C.: Cascading statistical and structural classifiers for iris recognition. In: International Conference on Image Processing (2004)
45. Zhenan, S., Yunhong, W., Tieniu, T., Jiali, C.: Improving iris recognition accuracy via cascaded classifiers. IEEE Trans. Syst. Man Cyber. **35**(3), 435–441 (2005)
46. Sun, Z., Tan, T., Qiu, X.: Graph matching iris image blocks with local binary pattern. In: Springer LNCS 3832: International Conference on Biometrics, January 2006
47. Zhang, P.-F., Li, D.-S., Wang, Q.: A novel iris recognition method based on feature fusion, In: International Conference on Machine Learning and Cybernetics, pp. 3661–3665 (2004)
48. Vatsa, M., Singh, R., Noore, A.: Reducing the false rejection rate of iris recognition using textural and topological features. Int. J. Signal Process, **2**(2), 66–72 (2005)
49. Kumar, A., Zhang, D.: Personal authentication using multiple palmprint representation. Pattern Recog. **38**, 1695–1704 (2005)
50. You, J., et al.: On hierarchical palmprint coding with multiple features for personal identification in large databases. IEEE Trans. Circuits Syst. Video Tech. **14**(2), 234–243 (2004)
51. Kong, W., Zhang, D., Li, W.: Palmprint feature extraction using 2-d gabor filters. Pattern Recog. **36**(10), 2339–2347 (2003)
52. Zuo, W., Lin, Z., Guo, Z., Zhang, D.: The multiscale competitive code via sparse representation for palmprint verification. In: Proceedings of CVPR, pp. 2265–2272 (2010)
53. Pan, X., et al.: Palmprint recognition using fusion of local and global features. In: Proceedings of the International Symposium on Intelligent Signal Processing and Communication Systems, pp. 642–645 (2007)
54. Hong, L., Jain, AK., Pankanti, S.: Can multi-biometrics improve performance. In: Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies (WAIAT-99), Morristown NJ, pp. 59–64, October 1999
55. Chen, G.Y., Xie, W.F.: Pattern recognition with SVM and dual-tree complex. Image Vis. Comput. **25**(6), 960–966 (2007)

Chapter 13
Biometric Recognition Systems Using Multispectral Imaging

Abdallah Meraoumia, Salim Chitroub and Ahmed Bouridane

Abstract Automatic personal identification is playing an important role in secure and reliable applications, such as access control, surveillance systems, information systems, physical buildings and many more applications. In contrast with traditional approaches, based on what a person knows (password) or what a person has (tokens), biometric based identification providing an improved security for their users. Biometrics is the measurement of physiological traits such as palmprints, fingerprints, iris etc., and/or behavioral traits such as gait, signature etc., of an individual person for personal recognition. Hand-based person identification provides a good user acceptance, distinctiveness, universality, relatively easy to capture, low-cost and inexpensive. Palmprint identification is one kind of hand-biometric technology and a relatively new biometrics due to its stable and unique traits. The rich texture information of palmprint offers one of the powerful means in personal identification. Several studies for palmprint-based person identification have focused on the use of palmprint images captured in the visible part of the spectral band. However, recently, the multispectral palmprints have been rendered available and the tendency now in the community is how to exploit these multispectral data to improve the performances of the palmprint-based person identification systems. In this chapter, we try

A. Meraoumia (✉)
Faculté des nouvelles technologies de l'information et de la communication, Laboratoire de Génie Électrique, Université de Ouargla, 30000 Ouargla, Algeria
e-mail: ameraoumia@gmail.com

S. Chitroub
Image Processing Laboratory, Electronics and Computer Science Faculty, USTHB, P.O. box 32, El Alia, Bab Ezzouar, 16111 Algiers, Algeria
e-mail: S_chitroub@hotmail.com

A. Bouridane
School of Computing, Engineering and Information Sciences, Northumbria University, Pandon Building, Newcastle upon Tyne, UK
e-mail: Ahmed.Bouridane@northumbria.ac.uk

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_13, © Springer-Verlag Berlin Heidelberg 2014

to evaluate the usefulness of the multispectral palmprints for improving the palmprint based person identification systems. For that purpose, we propose several systems of exploiting the multispectral palmprints. The results on a medium-size database show good identification performance based on individual modalities as well as after fusing multiple spectral bands.

13.1 Introduction

Accurate automatic personal identification is critical in a variety of applications, such as physical buildings, access control and information systems. Biometrics, which refers to identification based on physical or behavioral characteristics, is being increasingly adopted to provide identification with a high degree of confidence [1]. Thus, while traditional security measures such as PINs and passwords may be forgotten, stolen or cracked, biometrics characteristics cannot be easily duplicated or forged. Currently, a number of biometrics based technologies have been developed and the one of the most popular biometric systems is based on the hand due to its ease of use. Thus, there are several motivations for a hand biometric [2]; firstly, the data acquisition is economical via commercial low-resolution cameras, and its processing is relatively simple. Secondly, hand based access systems are very suitable for several usages. Thirdly, the hand features are more stable over time and are not susceptible to major changes. Oftentimes, some features related to a human hand are relatively invariant and distinctive to an individual. Among these features, palmprint is one biometric that has been systematically used to make identification for last years. Palmprint identification is a biometric technology which recognizes a person based on his/her palm pattern. Thus, the rich texture information of palmprint offers one of the powerful means in personal identification [3].

Palmprint identification is becoming a popular and convincing solution for identifying person's identity since palmprint is proved to be a unique and stable personal physiological characteristic. So far, majority of studies on palmprint identification are mainly based on image captured under visible light. However, multispectral imaging, which give different information from a variety of spectral bands, have been recently used to improve the performance of palmprint identification because each spectral band highlights specific features of the palm, making it possible to collect more information to improve the accuracy and anti-spoofing capability of palmprint systems [4]. These spectral bands provide different and complementary information on the same palmprint. In multispectral imaging technique, an acquisition device, to capture the palmprint images under visible and infrared light resulting into several bands, is used. The idea is to employ the resulting information in these bands to improve the performance of palmprint identification system.

Unimodal biometric systems perform person recognition based on a single source of biometric information. Such systems are often affected by some problems such as noisy sensor data and non-universality. Thus, due to these practical problems, the error rates associated with unimodal biometric systems are quite high and consequently it makes them unacceptable for deployment in security critical applications [5]. Some of these problems can be alleviated by using multimodal biometric systems. However, multispectral palmprint images were taken as a kind of multimodal biometrics. Furthermore, in the multimodal system design, the different spectral bands operate independently and their results are combined using an appropriate fusion scheme. Thus, the fusion can be performed at different levels.

The final goal of this chapter is to build a reliable biometric system, using information from palmprint images captured under visible and infrared light resulting into four spectral bands {*Red, Green, Blue* and *Near-InfRared (NIR)*}. For that, four biometric systems are proposed based on contourlet Transform (*CT*), Karhunen-Loeve transform (*KL*), Correlation Filter (*CF*) and log-Gabor Filter (*GF*). As mentioned above, the system performance could be improved by using the integration or fusion of information from spectral band modalities. Therefore, in this study, information presented by different spectral bands is fused to make the system efficient using both image level fusion and matching score level fusion.

13.2 Related Works

A multispectral imaging consists of capture and stacks the images of the same object or scene, each at a different spectral narrow band (each band representing the intensity image at a given wavelength). Multispectral imaging has been related to diverse fields such as remote sensing and medical imagery, as well as in biometric analysis (e.g., face, fingerprint). Han et al. [6] have presented multispectral palmprint recognition considering different illumination, including Red, Green, Blue and Infrared for characterizing the palmprint images. The Competitive Coding Scheme is adopted as matching algorithm. By fusing all spectrums (using wavelet-based image fusion method), the verification test results show better effort on motion blurred source images than single channel. Guo et al. [7] have compared palmprint recognition accuracy using white light and other six different color lights. The experiments on a large database have shown that white light is not the optimal illumination for palmprint recognition. In [8], Ying Hao et al. have proposed a method to improve the verification performance of contract-free palmprint recognition system by means of feature-level image registration and pixel-level fusion of multispectral palm images. By using various image fusion methods, the obtained . results have demonstrated the improvement in the recognition performance compared to only use the monochrome images. Zhang et al. [9] have proposed an online system of multispectral palmprint verification. In their works, a data acquisition device is designed to capture the palmprint images under different illuminations and then use the orientation-based coding feature representation for multispectral images of the palm. However, score level fusion scheme is used. The palmprint verification experiments have demonstrated the superiority of multispectral fusion to each single spectrum, which results in both higher verification accuracy and anti-spoofing capability. On the other hand, Khan et al. [10] have developed a multispectral palmprint prototype system based on the Contour Code representation. These codes are derived from the Non-sub-sampled Contourlet Transform. The multispectral palmprint verification results have shown that the Contour Code achieves a good performance compared to state-of-the-art methods. In [11], Cui et al. have used the Image-Based Linear Discriminant Analysis (*IBLDA*) and fusing the four bands palmprint images. The four bands palmprint images are divided into two groups and then construct a complex matrix using the two bands of every group and did the feature level fusion for each group respectively. The experiments have shown that a higher recognition rate can be achieved when IBLDA method is used. Xu et al. [12] have proposed a new method for multispectral images based on a quaternion model. Thus, multispectral palmprint images represented by a quaternion matrix, then Principal Component Analysis (PCA) and Discrete Wavelet Transform (DWT) have applied respectively on the matrix to extract palmprint features. Experimental results have shown that the quaternion matrix can achieve a higher recognition rate.

13.3 Multimodal Biometric System

Multimodal approaches to biometric recognition tasks, that is, approaches that combine two or more biometric traits to perform personal identification, have been found to produce better results than single biometrics using a single trait alone. The information fusion strategies employed in multimodal biometric systems can be categorized into four levels; (i) fusion at image level, (ii) fusion at feature extraction level, (iii) fusion at matching score level, and (iv) fusion at decision level. The first and third fusion schemes are considered in this work.

13.3.1 Fusion at Image Level

The goal of image fusion is to integrate complementary information coming from different sensors (e.g., visible and infrared images, multispectral satellite images) into one new image containing more information of which cannot be achieved by individual images. The aim of this technique, apart from reducing the amount of data, is to create new images that are more suitable for the purposes of image-processing tasks such as segmentation, object detection or target recognition in applications such as remote sensing and medical imaging [13]. The most important issue concerning image fusion is to determine how to combine the sensor images. In recent years, several image fusion techniques, include principal component analysis and multi-resolution methods, have been proposed. The image fusion based on multi-resolution method can perform by the pyramid decomposition (such as *Laplace* pyramid, *contrast* pyramid, *gradient* pyramid) or by wavelet decomposition, such as discrete *wavelet* transform [14]. In this chapter, we will explore the efficiency of the above techniques, by fusing multispectral palmprint images, in biometric system performance.

13.3.2 Fusion at Matching Score Level

Fusion at the matching-score level is the most common approach in the field of biometrics due to the ease in accessing and combining the scores generated by different matchers [15]. In our system we adopted the combination approach, where the individual matching scores are combined to generate a single scalar score, which is then used to make the final decision. During the system design we experimented five different rules. These rules consist of the *sum* (*SUM*) and *WeigHTed-sum* (*WHT*) of the two similarity measures, their *MINimum* (*MIN*) and *MAXimum* (*MAX*) of both and finally their MULtiplication (*MUL*). The final decision of the classifier is then given by choosing the class, which maximizes the fused similarity measures between the sample and the matching base.

During the identification process, the characteristics of the test spectral bands are analyzed and then the distance d between this feature vector and all of templates (models) in the database are computed, therefore the vector, D, given all these distance is given as:

$$D = [d_0 \quad d_1 \quad d_2 \quad d_3 \cdots\cdots d_N] \tag{13.1}$$

where N represents the size of the system database.

The matching scores output by the various modalities are heterogeneous, score normalization is needed to transform these scores into a common domain, prior to combining them. Thus, a *Min-Max* normalization scheme [16] was employed to transform the scores computed into similarity scores in the same range. Thus,

$$\tilde{D} = \frac{D - Min(D)}{Min(D) - Max(D)} \tag{13.2}$$

where $\tilde{D}$ represent the normalized vector. However, these scores are compared, and the highest (or lowest, with respect to the feature matching criteria) score is selected. Therefore, the best matching score is D_o and its equal to:

$$D_o = \max_i(\tilde{D}) \quad \text{or} \quad D_o = \min_i(\tilde{D}) \tag{13.3}$$

Finally, this score is used for decision making.

13.4 Multispectral Palmprint Identification

13.4.1 Multispectral Palmprint Image

A palmprint is defined as the unique inner surface of a hand between the wrist and the fingers. Palmprint biometric has several advantages as compared with other biometrics. As palmprint contains more information they are more distinctive, they

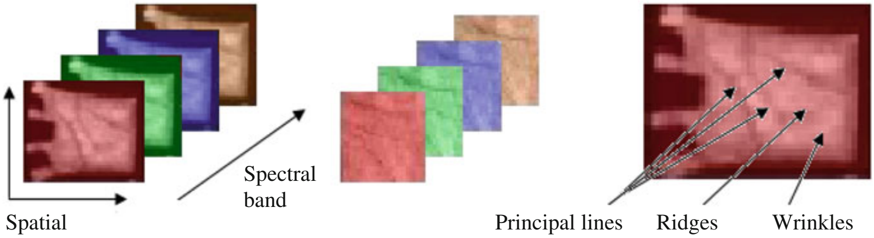


Fig. 13.1 Sample of multispectral palmprint image, with four bands, in our database

can be captured using low resolution devices, invariant with time and widely accepted by users [17]. Palmprint contains many features (See Fig. 13.1) like principal lines, wrinkles (secondary lines), ridges, minutiae points, singular points and texture etc. All these factors make the palmprint as an important biometric trait, especially, for low cost and medium security applications. The multispectral imaging technique can be give different information from the same palmprint using an acquisition device to capture the palmprint images under different wavelengths resulting into several spectral bands. For example, Fig. 13.1 shows a sample of multispectral palmprint image contain four spectral bands.

13.4.2 *Multispectral Palmprint Database*

The proposed methods are validated on multispectral palmprint database from the Hong Kong polytechnic university (PolyU) [18]. The database contains images captured with visible and infrared light. Four palmprint images for each person, including *Red*, *Green*, *Blue* and *near-infrared* spectral band, are collected 6000 multispectral palmprint images were collected from 500 persons. These images were collected in two separate sessions. In each session, the person provide 6 images for each palm, so there are 12 images for each person. Therefore, 48 spectrum images of all illumination from 2 palms were collected from each person. The average time interval between the first and the second sessions was about 9 days.

13.4.3 *Palmprint Preprocessing*

In order to localize the palm area, the first step is to preprocess the palm images; we use the preprocessing technique described in [19] to align the palmprints. In this technique, Gaussian smoothing filter is used to smoothen the image before extracting the *ROI* sub-image and its features. After that, Otsu's thresholding is used for binarized the hand. A contour-following algorithm is used to extract the hand contour. The

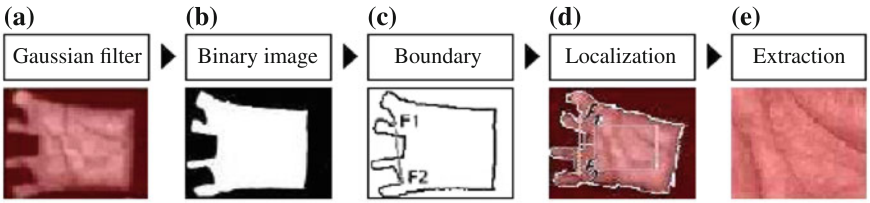


Fig. 13.2 Various steps in a typical region of interest extraction algorithm. **a** The filtered image, **b** The binary image, **c** The boundaries of the binary image and the points for locating the *ROI* pattern, **d** The central portion localization, and **e** The preprocessed result (*ROI*)

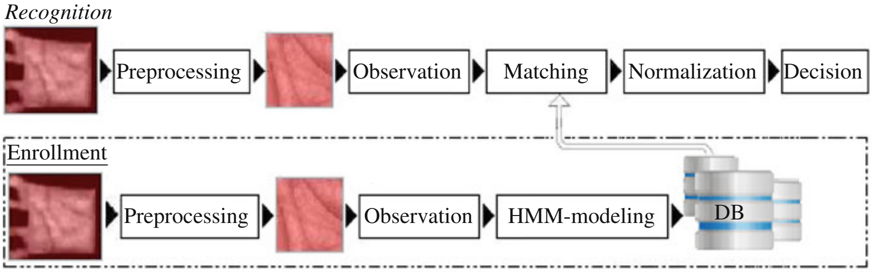


Fig. 13.3 contourlet based unimodal palmprint identification system

tangent of the two stable points on the hand contour (they are between the forefinger and the middle finger and between the ring finger and the little finger) are computed and used to align the palmprint. The central part of the image, which is 128 × 128, is then cropped to represent the whole palmprint. Figure 13.2 shows the palmprint pre-processing steps.

13.4.4 Unimodal Identification System Test Results

13.4.4.1 Contourlet Based Identification System

Figure 13.3 illustrates the various modules of the proposed contourlet based multispectral palmprint identification system (unimodal system). The proposed system consists of preprocessing, feature extraction, matching and decision stages. For enrollment phase, an observation vector is extracted from each spectral band which describes certain characteristics of the palmprint images using *CT* and modeling using an *HMM* model. Finally, the models parameters are stored in the system database. For identification, the same features vectors are extracted from the test spectral bands and the log-likelihood is computed using all of models references in the database. For the matching score level fusion based multimodal system, each sub-system computes its own matching score and these individual scores are finally combined

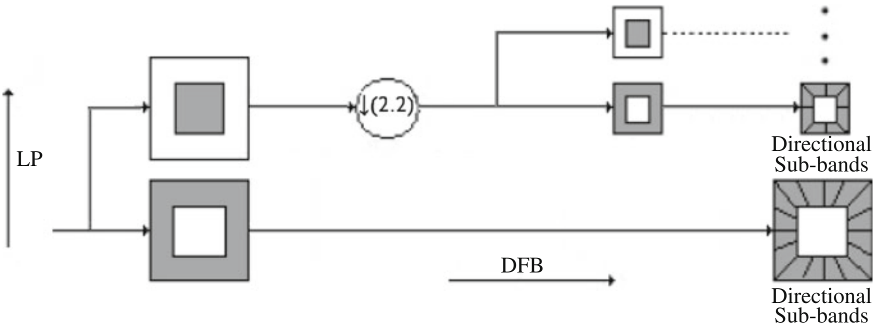


Fig. 13.4 The contourlet transform: first, a multiscale decomposition into bands by the *LP* is computed, and then a *DFB* is applied to each bandpass channel

into a total score, which is used by the decision module. Based on this matching score a decision about whether to accept or reject a user is made. In the image level fusion based multimodal system, two or more spectral bands are fused.

Discrete contourlet transform is a multi-scale and directional image representation that uses first a wavelet like structure for edge detection, and then a local directional transform for contour segment detection [20]. In general the application of *CT* to an image involves two stages. A Laplacian Pyramid (*LP*) is first used followed by the application of a Directional Filter Bank (*DFB*). The Fig. 13.4 illustrates the contourlet transformation.

Let a_o be the input image, the output after the LP stage is J bandpass images b_j, $j = 1, 2, \ldots, J$ and a lowpass image a_J. That means, the j-th level of the LP decomposes the image a_{j-1} into a coarser image a_j and a detail image b_j. Each bandpass image b_j is decomposed by an l_j-level DFB into 2^{l_j} bandpass directional images $c_{j,k}^{(l_j)}$, $k = 0, 1, \ldots, 2^{l_j} - 1$. Since the multiscale and directional decomposition stages are decoupled in the discrete contourlet transform, we can have a different number of directions at different scales, thus offering a flexible multiscale and directional expansion.

The feature extraction module processes the acquired biometric data (each spectral band) and extracts only the salient information to form a new representation of the data. In our method, the spectral band is typically analyzed using the contourlet transform. After the decomposition transform of the *ROI* sub-image, some of the contourlet bands are selected and compressed using *PCA* to construct observation vectors. Since an *HMM* model of each observation vector is constructed.

To create an observation vector, the spectral band is transformed into a contourlet sub-bands form. Then the palmprint feature vectors are created by combining some contourlet bands extracted using contourlet decomposition. Several vectors can be extracted using 1–4 levels of decomposition for each spectral band. Figure 13.5 shows an example of feature vector extraction methods using a contourlet decomposition with 4 levels. In this example, 4 vectors, υ_1^S, υ_2^S, υ_3^S and υ_4^S, can be extracted using

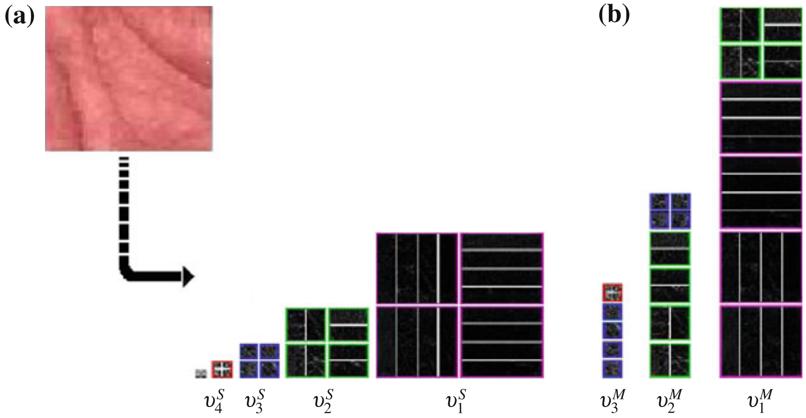


Fig. 13.5 Observation vector generation using 4 levels contourlet decomposition. **a** Single level based feature vector extraction and **b** Multiple level based feature vector extraction

a single level. Thus, we can also combined some levels to obtain another 3 vectors (multiple levels), υ_1^M, υ_2^M and υ_3^M. Then, each feature vector, υ_i^X with $X = S$ or M, is compressed using *PCA* technique and some of principal components are selected for representing the final observation vectors ν_i^X.

An *HMM* is a Markov chain with a finite number of states. Although the Markov states are not directly observable, each state has a probability distribution associated with the set of possible observations. Thus, an *HMM* is characterized by [21]: a state transition probability matrix (A), an initial state probability distribution (π) and a set of probability density functions associated with the observations for each state (B). A compact notation $\lambda = (A, B, \pi)$ can represent the parameter set of the model. Finally, forward-backward recursive algorithm, Baum-Welch, and Viterbi algorithm are used to solve evaluating, training, and decoding, respectively [22].

The *HMM* parameters are estimated from a set of training vectors υ_i^X, with the objective of maximizing the likelihood. So, the *HMM* training problem is the estimation of the model parameters $\lambda = (A, B, \pi)$ for a given data set υ_i^X.

For the matching process, after extracting the observation vectors corresponding to the test spectral band, the probability of the observation sequence given a *HMM* model is computed via a viterbi recognizer [23]. The model with the highest log-likelihood is selected and this model reveals the identity of the unknown palmprint. Thus, during the identification process, the characteristics of the test image are extraction. Then the log-likelihood score of the observation vectors given each model, $d_i = P(\nu_i^X|\lambda_i) = \ell(\nu_i^X, \lambda_i)$, is computed. Therefore, the score vector is given by:

$$D(\nu_i^X) = [\ell(\nu_i^X, \lambda_1)\ \ell(\nu_i^X, \lambda_2)\ \ell(\nu_i^X, \lambda_3) \cdots\cdots \ell(\nu_i^X, \lambda_N)] \tag{13.4}$$

Table 13.1 Unimodal identification system performance using single level based vector

Number of levels	ν_1^S		ν_2^S		ν_3^S		ν_4^S	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER
1	0.9230	4.651	x	x	x	x	x	x
2	0.8064	11.450	0.9262	0.805	x	x	x	x
3	0.7379	19.912	0.7921	0.658	0.9190	0.274	x	x
4	70.7777	4.242	0.7366	1.031	0.8171	0.418	0.9186	0.313

Table 13.2 Unimodal identification system performance using multiple level based vector

Number of levels	ν_1^M		ν_2^M		ν_3^M	
	T_o	EER	T_o	EER	T_o	EER
2	0.7390	0.969	x	x	x	x
3	0.7161	1.053	0.8064	0.581	x	x
4	0.7819	0.695	0.8186	1.055	0.7164	0.188

where N represents the size of model database.

In order to find the best parameters of the ergodic *HMM* model, we choose empirically the number of gaussian in the Gaussian Mixture Model (*GMM*) equal to 1 and the number of states of the *HMM* equal to 3. Finally, in contourlet transformation, we apply the orthogonal wavelets '*pkva*'. In all experiments, three samples of each modality (spectral band) were randomly selected to construct a training set (enrollment) and the other was taken as the test set. Thus, there are total of $3 \times 400 = 1{,}200$ training images and $9 \times 400 = 3{,}600$ test images for each modality, respectively. Therefore, there are totally 3,600 genuine comparisons and 718,200 impostor comparisons are generated, respectively.

(1) Observation vector selection

The PCs vectors reflect the compact information of different column vectors of υ_i^X. Most of these vectors they become negligible, as result; the vector derived from the initial vectors computation (υ_i^X) is limited to an array of summed vectors within all components. The test was repeated for various spectrum images, only 8, 16, 32 and 64 PCs vectors for υ_i^S, with $i = 1, \ldots, 4$, are enough to achieve good representation (100 % of the total information). In other hand, only 8, 24 and 64 PCs vectors for υ_i^M, with $i = 1, \ldots, 3$, give a good representation (100 % of the total information).

At the first stage, we conducted several experiments to investigate the effectiveness of the observation vector extraction methods at different number of contourlet decomposition levels. Our goal is to choose the observation vector extraction method and the number of the contourlet decomposition levels yield the best performance. For this, we found the performance under different vectors against the contourlet

Table 13.3 Unimodal identification system test results under different spectral bands

Observation vectors	Blue		Green		Red		NIR	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER
Single	0.9190	0.274	0.9173	0.243	0.9154	0.317	0.8905	0.719
Multiple	0.7164	0.188	0.7969	0.414	0.7983	0.378	0.7746	0.944

decomposition levels. Tables 13.1 and 13.2 presents the average identification results in terms of Equal Error Rate (*EER*) as a function of the feature vector extraction method and the number of decomposition levels. From Table 13.1 (single decomposition level based observation vector extraction) we can observe that the ν_3^S vector at three decomposition level offers the best identification with *EER* equal to 0.247 % at $T_o = 0.9190$. Once, as shown in Table 13.2, the identification system performance can be improved over 34 % with respect to the multiple decomposition level based observation vector extraction (ν_3^M vector at four decomposition level). The system can achieve an *EER* equal to 0.188 % at $T_o = 0.7164$. In the rest of experiments, we tested the performance of the best vector extraction methods, ν_3^S and ν_3^M, for each spectral band separately.

(2) Test Results

The goal of this experiment was to evaluate the system performance when we using information from each modality (spectral band). For this, we found the performance under different modalities {*Blue*, *Red*, *Green*, and *NIR*}. However, the two best observation vector (single and multiple levels based vector extraction methods) are selected. Thus, in order to see the performance of the identification system, we usually present, in Table 13.3, the results for all spectral bands.

This table shows that the *Green* spectral band offers better results in terms of the *EER*, for the first methods (single level based vector extraction method). In this case, the identification system can achieve an *EER* of 0.243 % for $T_o = 0.9173$. Thus, as the table show, the second identification system (multiple level based vector extraction method) produces the best accuracy (0.188 % at $T_o = 0.7164$) in the case of *Blue* spectral band. Note that, all of the methods can give a small to medium errors and it's compared with other methods presented in literatures. Finally, the results expressed as a Receiver Operating Characteristics (*ROC*) curves, which is a plot of False Reject Rate (*FRR*) against False Accept Rate (*FAR*), obtained by the different spectral bands with single decomposition level based vectors extraction are plotted in Fig. 13.6a, b plot the *ROC* curves for the different spectral bands with multiple decomposition levels based vectors extraction. By the analysis of the previous results, it can be observed that the performance of the unimodal identification system is significantly improved by using the multiple decomposition levels based vectors extraction methods. In addition, experiments also demonstrate that *Blue* spectral band performs better results.

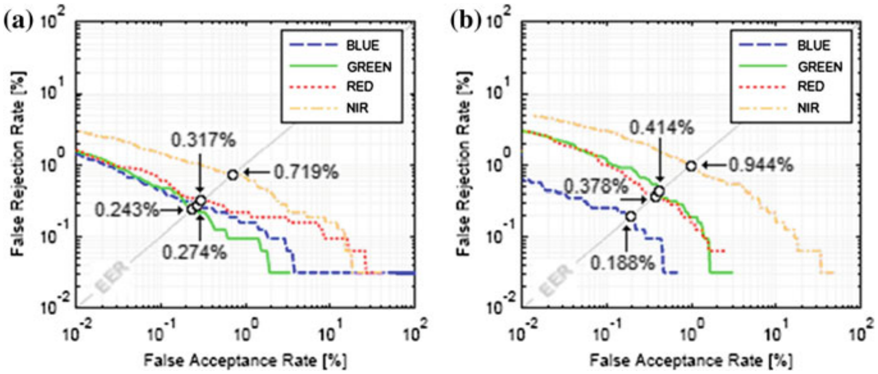


Fig. 13.6 Contourlet based unimodal identification system performance. **a** The *ROC* curves for all spectral bands using single decomposition level based observation vectors and **b** The *ROC* curves for all spectral bands using multiple decomposition levels based observation vectors

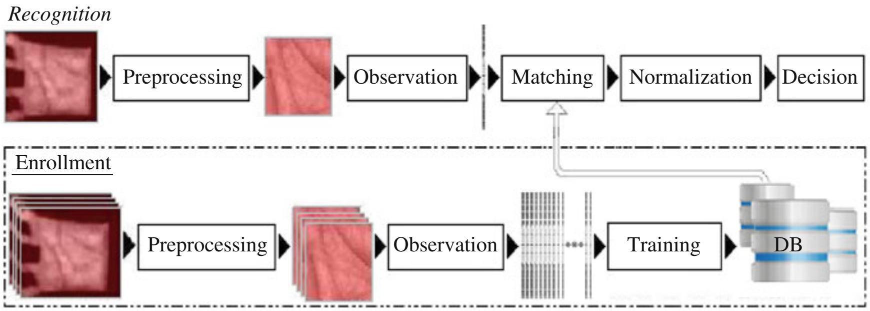


Fig. 13.7 Karhunen-Loeve based unimodal palmprint identification system

13.4.4.2 Karhunen-Loeve Based Identification System

Figure 13.7 shows the block-diagram of the proposed unimodal biometric identification system based on the palmprint image. In the preprocessing module, the *ROI* sub-image is localized. For the enrolment phase, each *ROI* sub-image is mapped into one dimensional signal (Observation vector). After that, these vectors are concatenated into two dimensional signal. This vector is transformed by the *KL* transform into feature space called eigenpalms space (Training module). For identification phase, the same feature vector is extracted from the test palmprint image and projecting into corresponding subspace. Then euclidian distance is computed using all of the references in the database (Matching module). Finally, after a normalization process, decision which person accepted or rejected is made.

The KL transform, also known as *PCA*, applied to a set of images, can be used to find the subspace that is occupied by all of the images from the analyzed set. We can calculating principal component by the following method [24]:

Let the training set of vectors of original data (each vector with dimension M), X, be x_1, x_2, x_3,..., x_N. First, compute the mean of original data of the set by: $\widetilde{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Second, subtract the mean from each original data to generate the mean removed data by $\psi_i = x_i - \widetilde{X}$. Third, form the matrix using mean removed data of $(M \times N)$ dimension, $D = [\psi_1 \;\; \psi_2 \;\; \psi_3 \cdots \psi_N]$. Fourth, compute the sample covariance matrix (C) of dimension $(M \times M)$, $C = \frac{1}{N}\sum_{n=1}^{N} \psi_n \psi_n^T = DD^T$ and compute the eigen values of the covariance matrix and of the eigen vectors for the eigen values. Finally, keep only the eigen vectors corresponding to L largest eigen values. These eigen values are called as principal components.

The *KL* applied to a set of images, can be used to find the subspace that is occupied by all of the images from the analyzed set. When the images are encoded into this subspace and then returned to the original space, the error between the reconstructed and the original images is minimized. To begin, we have a training set of N_I *ROI* sub-images. By reordering these *ROI* sub-images into one dimensional vector, x_i, and concatenate all x_i, with $i = [1 \cdots N]$, for obtaining a two dimensional vector, $X = [x_1, x_2, x_3, \ldots, x_{N_I}]$. The process of obtaining a single subspace consists of finding the covariance matrix C of the training set of *ROI* sub-images, X, and computing its eigenvectors. Each original *ROI* sub-image can be projected into this subspace. The eigenvectors spanning the palm-space can be represented as images with the same dimensionality as the palm *ROI* sub-images used to obtain these eigenvectors. These sub-images are called eigenpalms.

In order to identify a user, the matching process between the test template, ψ_t, and the templates from the database, ψ_i, has to be performed. The matching between corresponding feature vectors is based on the Euclidean distance. In this step, the following distance is obtained:

$$d(\psi_t, \psi_i) = \sqrt{(\psi_t - \psi_i)(\psi_t - \psi_i)^T} \tag{13.5}$$

Therefore, the score vector is given by:

$$D(\psi_t) = [d(\psi_t, \psi_1) \quad d(\psi_t, \psi_2) \quad d(\psi_t, \psi_3) \cdots\cdots d(\psi_t, \psi_N)] \tag{13.6}$$

where N is the total number of templates in the database.

At the first stage, we conducted several experiments to investigate the effectiveness of each spectral band. For this, experiment was conducted using different palmprint modalities (*Red*, *Green*, *Blue* and *NIR*). Our goal is to choose the modality yield the best performance. Therefore, by varying the palmprint modalities we can choose the modality which minimize the systems *EER*. Thus, in the case of identification, the *ROC* curves for four distinct modalities are shown in Fig. 13.8a. This figure compares the identification performance of the system varying palmprint modalities. After analyzing this figure we were able to conclude that the *NIR* band based system achieved the best performance, it can achieve an *EER* equal to 0.056 % at a threshold $T_o = 0.1152$. Poor results are obtained when using Green modality, in this case the system work with an *EER* equal to 7.709 % at a $T_o = 0.1251$. The *ROC* curve, which is a plot of Genuine Acceptance Rate *(GAR)* against FAR for all possible thresholds,

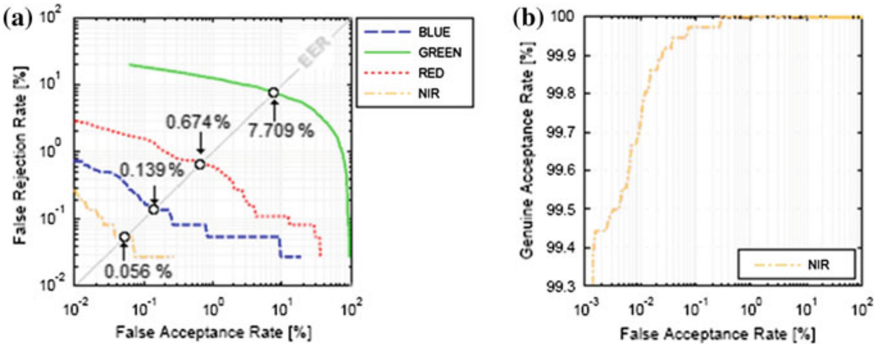


Fig. 13.8 Karhunen-Loeve based unimodal identification system performance. **a** The *ROC* curves with respect to the spectral bands and **b** The *ROC* curve for the *NIR* modality based system

Table 13.4 Unimodal identification test results under different spectral bands

DB	Blue			Green			Red			NIR		
	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR
	0.001	0.002	1.194	0.025	0.245	16.31	0.001	0.008	3.000	0.010	0.001	0.611
400	0.122	0.139	0.139	0.125	7.709	7.709	0.134	0.674	0.674	0.115	0.056	0.056
	0.300	5.712	0.056	0.180	19.65	5.500	0.250	7.583	0.111	0.200	0.432	0.000

for the best case is displayed in Fig. 13.8b. Compared with other existing unimodal systems, the proposed identification system has achieves better results expressed in terms of the *EER*. Finally, Table 13.4 present the experiments results obtained for all modalities.

13.4.4.3 Correlation Filter Based Identification System

The proposed system is composed of two different sub-systems exchanging information in image level or matching score level. Each sub-system exploit a different modalities (spectral bands). Each unimodal biometric system (for example Fig. 13.9 show a unimodal biometric identification system based on *Red* spectral band) consists of preprocessing, matching (correlation process), normalization and decision process. Modality (each spectral band) identification with correlation filters is performed by correlating a test image {transformed into the frequency domain via a Discrete Fourier Transform (*DFT*)} with the designed filter (enrollment) also in the frequency domain. The output correlation is subjected to an Inverse Discrete Fourier Transform (*IDFT*) and reordered into the dimensions of the original training image, prior to being phase shifted to the center of the frequency square. The resulting correlation plane is then quantified using performance measures (peak-to-sidelobe (*PSR*) ratio or max peak size ratio). Based on this unique measure, a final decision is made.

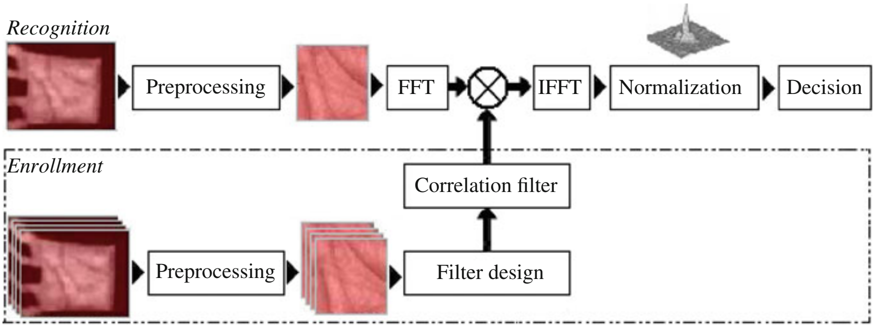


Fig. 13.9 Correlation filter response based unimodal palmprint identification system

In the matching process, each class is synthesized by a single (Unconstrained) Minimum Average Correlation Energy, (U)MACE, filter. Once the (U)MACE filter $H(u, v)$ has been determined, the input test image f is cross correlated with it in the following manner:

$$c(x, y) = IFFT\{FFT(f(x, y)) * H^*(u, v)\} \tag{13.7}$$

where the test image is first transformed to frequency domain and then reshaped to be in the form of a vector. The result of the previous process is convolved with the conjugate of the *(U)MACE* filter. This operation is equivalent with cross correlation with the (U)MACE filter. The output is transformed again in the spatial domain. Essentially *MACE* filter is the solution of a constrained optimization problem that seeks to minimize the average correlation energy while at the same time satisfy the correlation peak constraints. As a result the output of the correlation planes will be close to zero everywhere except at the locations of the trained objects that are set to be correct where a peak will be produced.

MACE filter, H, is found using Lagrange multipliers in the frequency domain and is given by [25]:

$$H = D^{-1}X(X^{\star}D^{-1}X)^{-1}u \tag{13.8}$$

D is a diagonal matrix of size $d \times d$, (d is the number of pixels in the image) containing the average correlation energies of the training images across its diagonals. X is a matrix of size $N \times d$ where N is the number of training images and $\star$ is the complex conjugate. The columns of the matrix X represent the discrete fourier coefficients for a particular training image X_n. The column vector u of size N contains the correlation peak constraint values for a series of training images. These values are normally set to 1.0 for images of the same class.

The *UMACE* filter like the MACE filter minimizes the average correlation energy over a set of training images, but does so without constraint (u), thereby maximizing the peak height at the origin of the correlation plane. The *UMACE* filter expression,

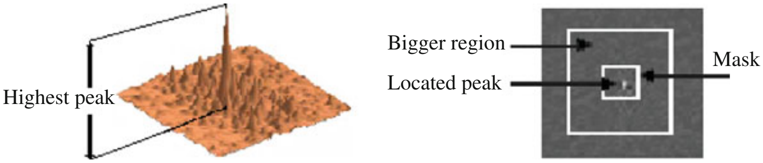


Fig. 13.10 Similarity matching. (*left*) Max peak size and (*right*) Peak-to-sidelobe ratio

Table 13.5 Unimodal identification test results under different spectral bands

Filter	Match	Blue		Green		Red		NIR	
		T_o	EER	T_o	EER	T_o	EER	T_o	EER
MACE	Peak	0.7654	0.004	0.9180	0.003	0.7352	0.002	0.7800	0.003
	PSF	0.9101	0.000	0.7840	0.000	0.9580	0.000	0.8565	0.000
UMACE	Peak	0.9160	0.000	0.9762	0.000	0.9619	0.000	0.9121	0.001
	PSF	0.7153	0.012	0.9960	0.000	0.6290	0.111	0.7725	0.069

H, is given by:

$$H = D^{-1}X \tag{13.9}$$

The maximum peak value (see Fig. 13.10a is taken as the maximum correlation peak value over a correlation plane. The height of this peak (Max peak size) can be used as a good similarity measure for image matching. Another parameter (see Fig. 13.10b, the *PSR* is used as a performance measure for the sharpness of the correlation peak. *PSRs* are typically large for true class and small for false category. Thus, the *PSR* is used to evaluate the degree of similarity of correlation planes. The significance of the *PSR* is that it measures the sharpness of the correlation function. *PSR* can be calculated as [26]:

$$d = PSR = \frac{Peak - \mu_{SR}}{\sigma_{SR}} \tag{13.10}$$

Peak is the maximum located peak value in the correlation plane, μ_{SR} is the average of the sidelobe region surrounding the peak (e.g., 40 × 40 pixels, with a mask region: 5 × 5 excluded zone around the peak) and σ_{SR} is the standard deviation of the sidelobe region values.

Table 13.5 compares the performance of the palmprint identification system under the two filters (*MACE* and *UMACE*) and the two performance measures (Peak size and *PSR*). The experimental results indicate that the *MACE* filter with the *PSR* matching perform better result than the other cases in terms of *EER*. Our identification system can achieve a best *EER* of 0.000 % for all spectral bands. Therefor, the *UMACE* filter and peak size matching done a *zeroEER* for the *Red*, *Green* and *Blue* spectral bands. For the *NIR* band, the described identification system can recognize palmprints quite accurately with an *EER* of 0.001 % and T_o = 0.9121. In general, the results show the benefits of using the palmprint modalities (all spectral bands gives a best

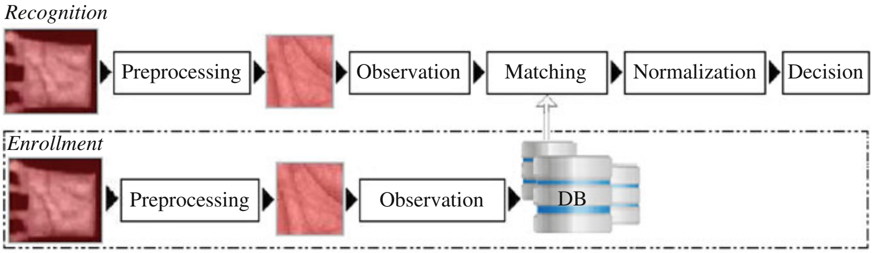


Fig. 13.11 Log-Gabor filter response based unimodal palmprint identification system

identification rate). Thus, the performance of the unimodal identification system is significantly improved by using the modalities of the multispectral palmprint.

13.4.4.4 Gabor Filter Based Identification System

Figure 13.11 is an overview of typical biometric system. It can be divided into two modules: (i) enrollment and (ii) identification. The enrollment module scans the palmprint (each spectral band) of a person through a sensing device and then stores a representation (called template or feature) of the spectral band in the database. The identification module is invoked during the operation phase. The same representation which was used in enrollment phase is extracted from the input spectral band and matched with a large number of palmprints in the database and as a result, palmprint is accepted or rejected.

The most discriminating information present in a spectral band pattern must be extracted. Only the significant features must be encoded so that comparisons between templates can be made. 1D Log-Gabor filter is able to provide optimum conjoint representation of a signal in space and spatial frequency [27]. Thus, the features are generated from the *ROI* sub-images by filtering the image with 1D Log-Gabor filter (see Fig. 13.12). Gabor features are a common choice for texture analysis. They offer the best simultaneous localization of spatial and frequency information. One weakness of the Gabor filter in which the even symmetric filter will have a DC component whenever the bandwidth is larger than one octave [28]. To overcome this disadvantage, a type of Gabor filter known as log-Gabor filter, which is Gaussian on a logarithmic scale, can be used to produce zero DC components for any bandwidth. The frequency response of a log-Gabor filter is given as:

$$G(f) = \exp\left[\frac{-(\log(f/f_o))^2}{2(\log(\sigma/f_o))^2}\right] \tag{13.11}$$

where f_o represents the center frequency, and σ gives the bandwidth of the filter. The ROI sub-images (rows) were unwrapped to generate 1D vector for feature extraction. These signals were convolved with 1D Log-Gabor filter. The resulting convolved

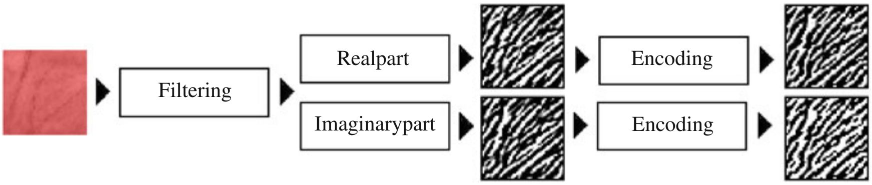


Fig. 13.12 Extraction of the biometric features from the hand image by log-Gabor filtering

form of the signal is complex valued. We then apply the following inequalities to extract binary response templates (encoding process) for both, real and imaginary part.

$$\begin{aligned} b_r &= 1 \quad if \;\; Re[\bullet] \geq 0 \quad b_r = 0 \quad if \;\; Re[\bullet] < 0 \\ b_i &= 1 \quad if \;\; Im[\bullet] \geq 0 \quad b_i = 0 \quad if \;\; Im[\bullet] < 0 \end{aligned} \tag{13.12}$$

Feature extraction method stores the real and imaginary part in the feature vector.

Matching the obtained and the stored features is based on normalized Hamming distance between both representations. Let $T_P(i, j)$ and $T_Q(i, j)$ be the input and stored templates of size $N_1 \times N_2$. Then the Hamming Distance (*HD*) between T_P and T_Q can be defined as [29]:

$$HD = \frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} T_P(i, j) \oplus T_Q(i, j)}{N_1 * N_2} \tag{13.13}$$

It is noted that HD is between 0 and 1. for perfect matching, the matching score is zero. When the Hamming distance of two templates is calculated, one template is shifted left and right bit-wise and a number of hamming distance values are calculated from successive shifts. As the result, our palmprint matching process can handle different translation. Matching scores from both log-Gabor response parts (real and imaginary) are combined into a unique matching score using MIN rule fusion. Based on this unique score, a decision about whether to accept or reject a user is made. The unique matching score is obtained as:

$$d = min\{HD_{Real}, HD_{Imag}\} \tag{13.14}$$

where HD_{Real} and HD_{Imag} is the Hamming distance of the real and imaginary part, respectively.

In this experiment, our goal is to choose the best spectral band who's minimizing the system *EER*. For this, we found the performance under the different spectral bands. Figure 13.13 presents the average identification results from four repeated tests in terms of *EER*. As can be observed from this figure, *Red* band consistently achieved the best performance among all others spectral bands that were used in our experiments. It can achieve an *EER* equal to 0.315 % at the threshold $T_o = 0.2571$,

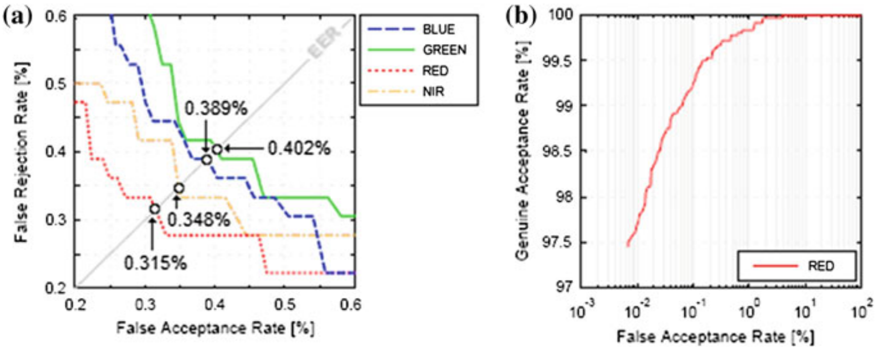


Fig. 13.13 Log-Gabor filter response based unimodal identification system performance. **a** The *ROC* curves with respect to the spectral bands and **b** The *ROC* curve for the RED modality based system

Table 13.6 Unimodal identification test results under different spectral bands

DB	BLUE			GREEN			RED			NIR		
	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR	T_o	FAR	FRR
	0.001	0.009	3.611	0.001	0.009	3.528	0.001	0.006	2.583	0.001	0.006	2.056
400	0.249	0.389	0.389	0.259	0.402	0.402	0.257	0.315	0.315	0.286	0.348	0.348
	0.450	3.021	0.000	0.500	4.635	0.000	0.500	4.980	0.000	0.500	3.311	0.000

while the other cases, *Green*, *Blue* and *NIR* spectral bands, show error rates in the range from 0.340 to 0.410 % with the thresholds 0.2592, 0.2493 and 0.2864, respectively. Also, as can be seen from Fig. 13.13, all spectral bands can gives a feasible *EER*. Finally, Table 13.6 shows the *FAR* and the *FRR* values of the four spectral bands for different thresholds.

For the comparison of all techniques, Fig. 13.14 summarizes the results in terms of the *GAR*. We can observe that, firstly, all techniques can gives an Higher identification rate (> 97 % accuracy). Second, the *CF* based unimodal provide the best result, *zeroEER*, than the other cases.

13.4.5 Multimodal Identification System Test Results

13.4.5.1 Fusion at Image Level

Image fusion is the process by which two or more images are combined into a single image. Therefore, information presented by different spectral bands is fused to make the system efficient. For that, a series of experiments were carried out using the multispectral palmprint database to selection the best combination {*Red-Green-Blue* (*RGB*) and *Red-Green-Blue-NIR* (*RGBN*)} and fusion technique (see Sect. 13.3.1) that minimize the *EER*. Thus, to determine the best combination and their fusion

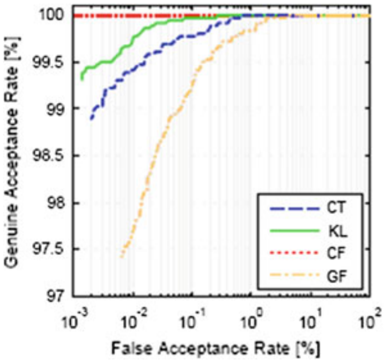


Fig. 13.14 Comparison between all unimodal systems

Table 13.7 Multimodal identification test results using fusion at image level (RGB combination)

MTD	DWT		PCA		Contrast		Gradiant		Laplacian	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
CT	0.7912	0.515	0.7953	0.347	0.8156	0.281	0.8147	0.327	0.8043	0.344
KL	0.1231	0.470	0.1257	1.153	0.1280	0.204	0.1140	1.028	0.1286	0.226
CF	0.7840	0.000	0.8160	0.000	0.6420	0.000	0.5957	0.000	0.6924	0.000
GF	0.2948	0.585	0.2909	0.500	0.2451	0.332	0.2554	0.402	0.2772	0.438

Table 13.8 Multimodal identification test results using fusion at image level (RGBN combination)

MTD	DWT		PCA		Contrast		Gradiant		Laolacian	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
CT	0.7923	0.688	0.7963	0.344	0.8234	0.279	0.8247	0.302	0.8007	0.399
KL	0.1390	0.593	0.1263	0.572	0.1389	0.194	0.1191	1.147	0.1351	0.194
CF	0.7812	0.000	0.6373	0.000	0.6555	0.000	0.6022	0.000	0.7202	0.000
GF	0.3251	0.770	0.2875	0.474	0.2512	0.312	0.2711	0.458	0.2811	0.438

technique, we usually give, in Table 13.7, the results for all the feature extraction methods and the image fusion rules in the case of *RGB* combination. Thus, the result suggests that, first, the *contrast* technique has performed better than the other techniques for all feature extraction methods. Second, *CF* based multimodal system give the best error, $EER = 0.000\,\%$ at threshold $T_o = 0.6420$. The *ROC* curves comparing all techniques, using *contrast* based image fusion, are plotted in Fig. 13.15a.

We also performed *RGBN* combination case by applying all image fusion rules on the spectral bands and calculated *EER*. Results are shown in Table. 13.8. From this figure, we can see that the *contrast* fusion rule is always efficiency than the other fusion rules. The *ROC* curves are plotted in Fig. 13.15b. By comparing the two combinations, *RGB* and *RGBN*, in the case of the *contrast* fusion rule, it can be seen that by combining all spectral bands, the performance is improved.

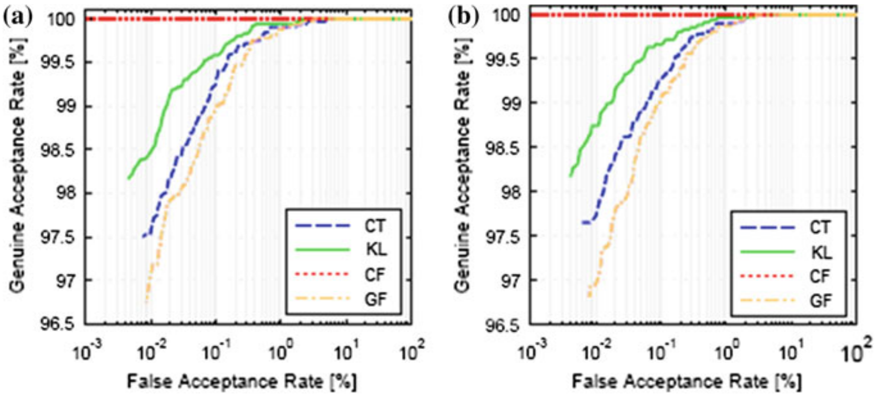


Fig. 13.15 Multimodal identification system performance using fusion at image level. **a** The *ROC* curves for *RGB* combination with respect to all techniques and **b** The *ROC* curves for *RGBN* combination with respect to all techniques

Table 13.9 Multimodal identification test results using fusion at score level (RGB combination)

MTD	SUM		WHT		MUL		MAX		MIN	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
CT	0.8068	0.030	0.7833	0.034	0.5874	0.031	0.8710	0.156	0.8026	0.048
KL	0.1228	0.088	0.1381	0.110	0.0225	0.153	0.2099	1.028	0.0003	0.183
CF	0.6408	0.000	0.6408	0.000	0.1640	0.000	0.8380	0.000	0.4560	0.000
GF	0.2583	0.153	0.2573	0.156	0.0180	0.156	0.3224	0.189	0.1988	0.250

13.4.5.2 Fusion at Matching Score Level

In the case of using *RGB* combination, to find the better of all the score fusion rules, with the lowest *EER*, table showing the results were generated (see Table. 13.9). From this table, we can observe the benefits of using the *SUM* rule for all techniques. For example, if *CT* is used, we have *EER* $= 0.030\,\%$ at the threshold $T_o = 0.8068$. In the case of using *KL*, *EER* was $0.088\,\%$ with $T_o = 0.1228$. Using *GF*, *EER* was $0.153\,\%$ ($T_o = 0.2583$). A *CF* remain the efficiency technique ($0.000\,\%$ at the threshold $T_o = 0.6408$) for a database size equal to 400. Therefore, the system can achieve higher accuracy at the fusion of the two matching score compared with a single matching score. The *ROC* curves, for all techniques in the case of *SUM* rule, are plotted in Fig. 13.16a.

In order to see the performance of the multimodal identification system, in the *RGBN* combination case, we usually present, in Table. 13.10, the results for all techniques and fusion rules. This table shows that the *WHT* rule offers better results in terms of the *EER*, for all techniques (excepted the case of *CF* technique, all rules give a *zeroEER*). In Fig. 13.16b, we plot the system performance as a function of all techniques for the *WHT* rule case.

Table 13.10 Multimodal identification test results using fusion at score level (RGBN combination)

MTD	SUM		WHT		MUL		MAX		MIN	
	T_o	EER	T_o	EER	T_o	EER	T_o	EER	T_o	EER
CT	0.8183	0.024	0.5837	0.017	0.7917	0.031	0.9000	0.156	0.7732	0.086
KL	0.0278	0.131	0.1439	0.028	0.0029	0.478	0.2281	0.639	0.0143	0.111
CF	0.6220	0.000	0.6220	0.000	0.5600	0.000	0.8870	0.000	0.3600	0.000
GF	0.2503	0.156	0.2521	0.105	0.0042	0.162	0.3367	0.219	0.1950	0.281

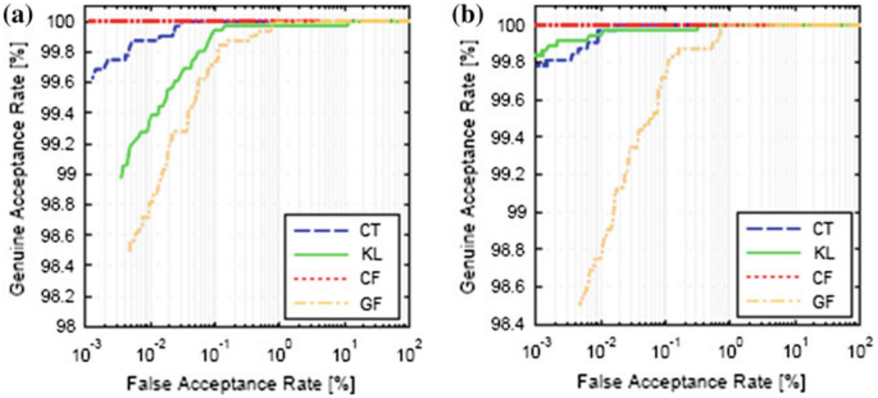


Fig. 13.16 Multimodal identification system performance using fusion at score level. **a** The *ROC* curves for *RGB* combination with respect to all techniques and **b** The *ROC* curves for *RGBN* combination with respect to all techniques

13.5 Reliability of Multispectral Imaging

Several studies results have shown that near-infrared image based biometrics, e.g., face recognition, offers many benefits. Multispectral image can provides more information than just color, and improves the accuracy of the color. It is not limited to visual range, rather can also be used in near-infrared spectral band. Thus, the near-infrared and visible image fusion has been successfully used for visualization purposes. In general, the information in the different spectral bands (e.g., near-infrared and visible image) is independent and complimentary. According to some specific fusion schemes, these spectral bands are fused in order to construct a more effective biometric systems. The purpose of this section is to establish the higher performance of the multispectral imaging based biometric identification system. For that, by respect to each technique used for feature extraction vector, we have developed a comparative study between the single spectral band based identification system (unimodal system), color image (*RGB*) based identification system (multimodal system) and all spectral bands (*RGBN*) based identification system (multimodal system) in the cases of two fusion schemes: image level and matching score level. Thus, the results of the multimodal systems are compared with those of unimodal systems to illustrate the

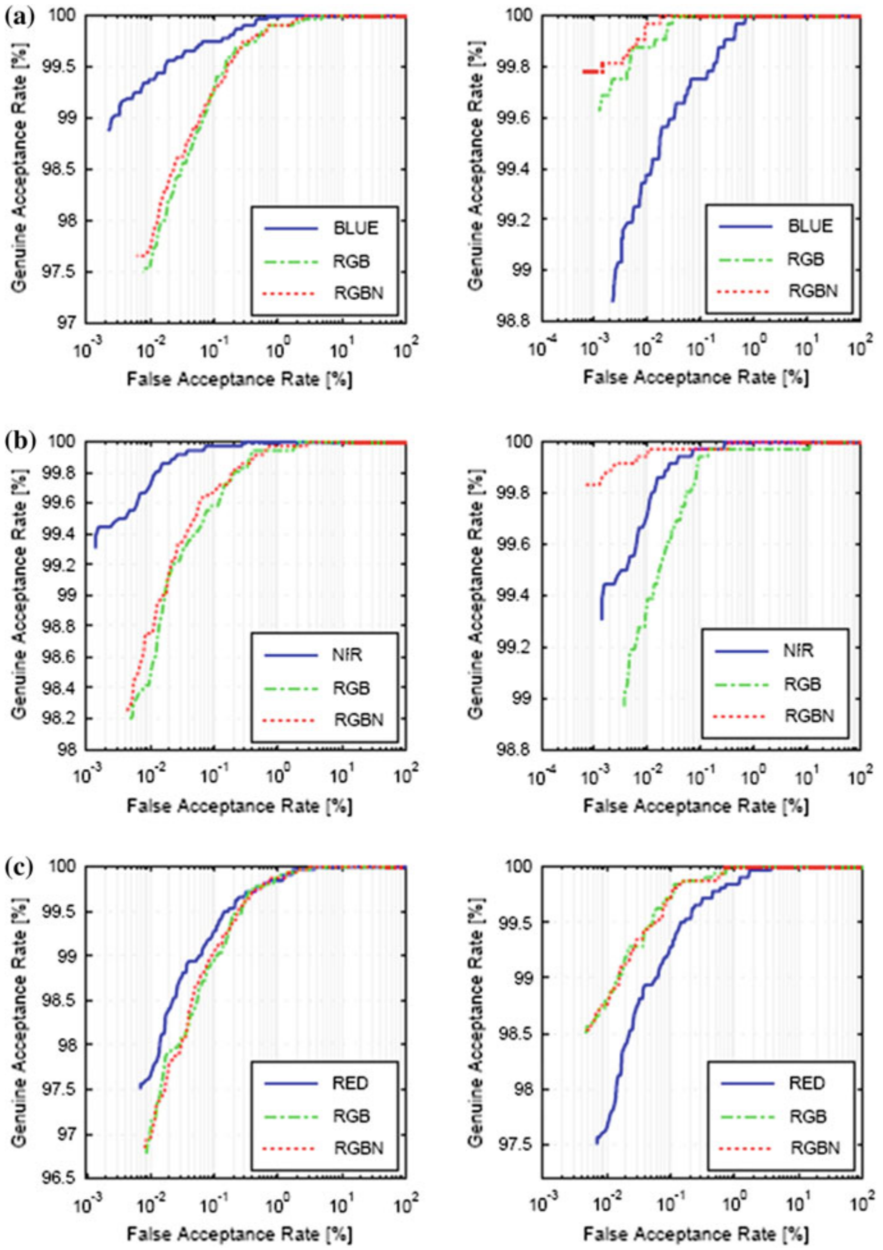


Fig. 13.17 Comparison of the unimodal and multimodal systems (first column for the image level fusion scheme and the second column for the score level fusion scheme) **a** The *ROC* curves using *CT* technique, **b** The *ROC* curves using *KT* technique, and **c** The *ROC* curves using *GF* technique

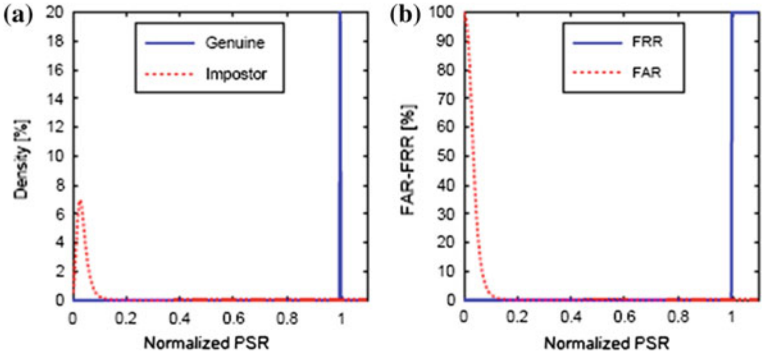


Fig. 13.18 Multispectral palmprint based multimodal identification test results (using all spectral bands (*RGBN*) fused at matching score level and *CF* based feature matching). **a** The genuine distribution and the imposter distribution and **b** the dependency of the *FAR* and the *FRR* on the value of the threshold

advantages of combining the results provided by these spectral bands. *ROC* curves for different feature extraction techniques and by different fusion schemes are shown in Fig. 13.17.

For all three feature extraction methods, Fig. 13.17 (*left*) shows that unimodal system based on single band is able to obtain higher accuracy than multimodal system based on *RGB* and *RGBN* combinations. Also, Comparing the curves in the Fig. 13.17 (*right*), we can see that both combinations (color and multispectral image) can give a considerable gain (except the case of *KT* based system, the RGB combination achieve the poor performance). This is probably because the fusion at matching score level provide in general the best result than the other fusion level. Also, they would feel more convenience to use the palmprint image with all spectral bands than with the color bands, which consequently leads to a best result.

Finally, among the four feature extraction methods proposed before, the *CF* method give the best results in our experiments, so it can achieve a *zeroEER* for both cases, unimodal and multimodal systems. The score distributions for genuine and impostor are estimated and are shown in Fig. 13.18a, b show the dependency of the *FAR* and the *FRR* on the threshold value.

13.6 Summary and Conclusions

Biometrics technology that relies on the physiological and/or behavioral human characteristics can provide increased security over standard forms of identification. This type of characteristics provides enhanced security levels because different individuals are unlikely to have similar characteristics. This work describes the design and development of a multimodal biometric personal identification system based on features extracted from multispectral palmprint. The use of palmprint modality is selected for several main reasons. First, it is more user-friendly than other biometric-based identifier, e.g., fingerprint and iris. Second, they can be captured using low resolution devices. Finally, due their great surface, it can extract several characteristics like wrinkles, ridges, principal lines etc. Furthermore, the unimodal systems suffer from various challenges like noise in the sensor data and non-universality etc, affecting the system performance. These problems are effectively handled by multimodal systems. Several studies were shown that multimodal biometric approach can be a possible solution for increased accuracy of the biometric based identification systems. Multispectral imaging can be integrated in these systems, to enrich their ability of identification, by adding new patterns containing different spectral band. The features are extracted using the contourlet transform, karhunen-loeve transform, correlation filter and log-gabor filter. Finally, information presented by different spectral bands (sub-system) is fused to make the system efficient using both image level fusion and matching score level fusion. Several studies were shown that multimodal biometric approach can be a possible solution for increased accuracy of the biometric based identification systems. Multispectral imaging can be integrated in these systems, to enrich their ability of identification, by adding new patterns containing different spectral band.

The experimental results, obtained on a database of 400 users, show a very high identification accuracy. They also demonstrate that combining different spectral bands does significantly reduce the accuracy of the system. In addition, our tests show that: First, from among the four methods used for feature extraction, the correlation filter offers the best identification rate ($EER = 0.000\,\%$), while the other methods show error rates lowest than $0.300\,\%$. Second, the multimodal system provides better identification accuracy than the best unimodal systems for all the tested spectral bands combinations in the case of fusion at matching score level. Third, combining the spectral bands give a considerable performance improvement. For further improvement, our future work will project to use other biometric modalities (Face and Iris) as well as the use of other fusion level like feature and decision levels. Also we will focus on the performance evaluation in both phases (verification and identification) by using a large size database.

References

1. Arun, A., Ross, A., Nandakumar, K., Jain, A.K.: Handbook of multibiometrics. In: Springer Science+Business Media, LLC, New York (2006)
2. Wayman, J., Jain, A., Maltoni, D., Maio, D.: Biometric Systems, Technology, Design and Performance Evaluation. Springer, London (2005)
3. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. IEEE Trans. Inf. Forensics Secur. **1**(2), 125–143 (2006)
4. Meraoumia, A., Chitroub, S., Ahmed, B.: Multimodal biometric person recognition system based on multi-spectral palmprint features using fusion of wavelet representations. In: Advanced Biometric Technologies. Published by InTech, pp. 21–42 (2011). ISBN 978-953-307-487-0

5. Zhang N.: Face recognition based on classifier combinations. In: International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM), Guiyang, China, 267–270, (2011)
6. Han, D., Guo, Z., Zhang, D.: Multispectral palmprint recognition using wavelet-based image fusion. In: proceedings of the 9th International Conference on Signal Processing, pp. 2074–2077 (2008)
7. Guo, Z., Zhang, D., Zhang, L.: Is white light the best illumination for palmprint recognition? In: Computer Analysis of Images and Patterns Lecture Notes in Computer Science, vol. 5702, 50–57 (2009)
8. Singh, R., Vatsa, M., Noore, A.: Hierarchical fusion of multispectral face images for improved recognition performance. Inf. Fusion **9**(2), 200210 (2008)
9. Zhang, D., Guo, Z., Guangming, L., Zhang, L., Zuo, W.: An online system of multispectral palmprint verification. IEEE Trans. Instrum. Measur. **59**(2), 480–490 (2010)
10. Khan, Z., Mian, A., Hu, Y.: Contour code: robust and efficient multispectral palmprint encoding for human recognition. In: ICCV2011 (2011)
11. Cui, J.-R.: Multispectral palmprint recognition using Image? Based linear discriminant analysis. Int. J. Biometrics **4**(2), 106–115 (2012)
12. Xu, X., Guo, Z., Song, C., Li, Y.: Multispectral palmprint recognition using a quaternion matrix. Sensors **12**(4), 4633–4647 (2012)
13. Bogoni, L., Hansen, M.: Pattern-selective color image fusion. Pattern Recogn. **34**(8), 1515–1526 (2006)
14. Simone, G., Farina, A., Morabito, F.C., Serpico, S.B., Bruzzone, L.: Image fusion techniques for remote sensing applications. Inf. Fusion **3**(1), 3–15 (2002)
15. Jain, A.K., Ross, A.: Learning user-specific parameters in a multibiometric system. In: Proceedings of IEEE International Conference on Image Processing (ICIP), pp. 57–60, Rochester, NY (2002)
16. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **38**, 2270–2285 (2005)
17. Jiaa, W., Huang, D.-S., Zhang, D.: Palmprint verification based on robust line orientation code. Pattern Recogn. **41**, 1504–1513 (2008)
18. PolyU Database. The Hong Kong Polytechnic University (PolyU) Multispectral Palmprint Database (2003). http://www.comp.polyu.edu.hk/biometrics/MultispectralPalmprint/MSP.htm
19. Zhang, D., Kong, A.W.K., You, J., Wong, M.: On-line palmprint identification. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1041–1050 (2003)
20. Singh, A.P., Mishra, A.: Image de-noising using contoulets (a comparative study with wavelets). Int. J. Adv. Networking Appl. **03**(03), 1210–1214 (2011)
21. Rabiner, L.R., Juang, B.H.: An introduction to hidden Markov models. In: IEEE ASSP Magazine, pp. 4–16 (1986)
22. Uguz, H., Arslan, A., Turkoglu, I.: A biomedical system based on hidden Markov model for diagnosis of the heart valve diseases. Pattern Recogn. Lett. **28**, 395–404 (2007)
23. Viterbi, A.J.: A personal history of the Viterbi algorithm. In: IEEE Signal Processing Magazine, pp. 120–142 (2006)
24. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. IEEE Trans. Neural Networks **13**(6), 1450–1464 (2002)
25. Hussain, A., Ghafar, R., Samad, S.A., Tahir, N.M.: Anomaly detection in electroencephalogram signals using unconstrained minimum average correlation energy filter. J. Comput. Sci. **5**(7), 501–506 (2009)
26. Ghafar, R., Hussain, A., Samad, S.A., Tahir, N.M.: Umace filter for detection of abnormal changes in eeg: a report of 6 cases. World Appl. Sci. J. **5**(3), 295–301 (2008)
27. Senapati, S., Saha, G.: Speaker identification by joint statistical characterization in the Log-Gabor wavelet domain. In: International Journal of Intelligent Systems and Technologies, Winter (2007)

28. Wang, F., Han, J.: Iris recognition method using Log-Gabor filtering and feature fusion. J. Xian Jiaotong Univ. **41**, 360–369 (2007)
29. Meraoumia, A., Chitroub, S., Saigaa, M.: Person's recognition using palmprint 2 based on 2D gabor filter response. In: Advanced Concepts for Intelligent Vision Systems. International conference, ACIVS 2009, Bordeaux, France, September 28 October 2, 2009. Proceedings. Berlin, Springer, LNCS 5807, 720–731 (2009)

Chapter 14
Electrocardiogram (ECG): A New Burgeoning Utility for Biometric Recognition

Manal Tantawi, Kenneth Revett, Abdel-Badeeh Salem and M. Fahmy Tolba

Abstract Recently, Electrocardiogram (ECG) has been emerged as a new biometric trait. ECG as a biological signal has the advantage of being an aliveness indicator. Moreover, it is difficult to be spoofed or falsified. In this chapter, a comprehensive survey on the employment of ECG in biometric systems is provided. An overview of the ECG, its benefits and challenges, followed by a series of case studies are presented. Based on the survey, ECG based biometric systems can be fiducial or non-fiducial according to the utilized features. Most of the non-fiducial approaches relax the challenging fiducial detection process to include only the R peak yielding to more reliable features. However, the drawback of such approaches is that they usually resulted in high dimension feature space. Hence, a non-fiducial ECG biometric system based on decomposing the RR cycles in wavelet coefficient structures using discrete biorthogonal wavelet transform is introduced. These structures were reduced through a proposed two-phase reduction process. The first phase globally evaluates the different parts of the wavelet structure (five details and one approximation parts) and maintains those parts that preserve the system performance. However, the second phase excludes more coefficients by locally evaluating the coefficients of each part based on an information gain criterion. Our experiments were carried out with four Physionet datasets using Radial basis functions (RBF) neural network classifier. Critical issues like stability over time, ability to reject impostors and generalization

M. Tantawi (✉) · K. Revett · A.-B. Salem · M. F. Tolba
Faculty of Computer and Information Sciences, Ain shams University, Cairo, Egypt
e-mail: manalmt2012@hotmail.com

K. Revett
e-mail: biomodelling@aol.com

A.-B. Salem
e-mail: abmsalem@yahoo.com

M. F. Tolba
e-mail: fahmytolba@gmail.com

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_14, © Springer-Verlag Berlin Heidelberg 2014

to other datasets have been addressed. The results indicated that with only 35 % of the derived coefficients the system performance not only can be preserved, but also it can be improved.

14.1 Introduction

Biometrics is a process of human identification and/or verification based on physiological or behavioral characteristics of individuals [1]. Biometric systems have become incorporated into the fabric of everyday life. It is employed where and whenever secure access is needed. In the last century, many approaches to person authentication have been emerged. The efficacy of a biometric trait is usually measured according to fundamental requirements that encompass: (1) universality, everyone must have the measured property; (2) collectability, the property can be measured quantitatively; (3) user acceptability; (4) uniqueness, the measured property is different among individuals; (5) circumvention, how easy it is to fool the system; (6) permanence which meansstability over time. Thus, any biometric trait should be rigorously assessed according to these requirements before making any claims about its utility [1].

The existing biometric approaches can be generally categorized into physiological or behavioral approaches. Physiological approaches are the most familiar approaches which rely on the constancy of fingerprints, iris scans, retinal pattern and etc. The advantage of this approach is that the measurable features are unique and stable over time. On the other hand, behavioral biometrics depends on the way an individual interacts with the authentication device. For instance, signature and voice are examples of behavioral biometrics. The potential benefit of behavioral biometrics is its simplicity, where in many cases a software only based approach is needed (i.e. keystroke). Moreover, they are more acceptable by users. Physiological approaches need specialized hardware which is usually expansive and difficult to employ for an internet based trusted application. The main issue with behavioral biometrics is the high intra-subject variability (we cannot make our signature exactly the same every time) [1].

An alternative and new branch of biometrics which has been gaining momentum over the past decade is the utilization of biological signals [1]. Biological signals are like electroencephalogram (EEG) and electrocardiogram (ECG) our interest in this chapter. ECG has been used for decades as a reliable diagnostic tool. Recently, the possibility of using ECG as a biometric trait has been introduced. Compared to other physiological traits (e.g. iris, fingerprint, face…etc.) and behavioral traits (e.g. signature, gait…etc.), the potential benefit of deploying ECG in the field of biometrics isits difficulty to be spoofed or falsified and being a life indicator [1, 2].

The existing ECG based biometric systems can be generally categorized according to the nature of the utilized features as fiducial or non-fiducial based systems. Fiducial based approach requires the detection of 11 fiducial points from the three complex waves labeled: P, QRS and T displayed for each normal heartbeat in an ECG trace

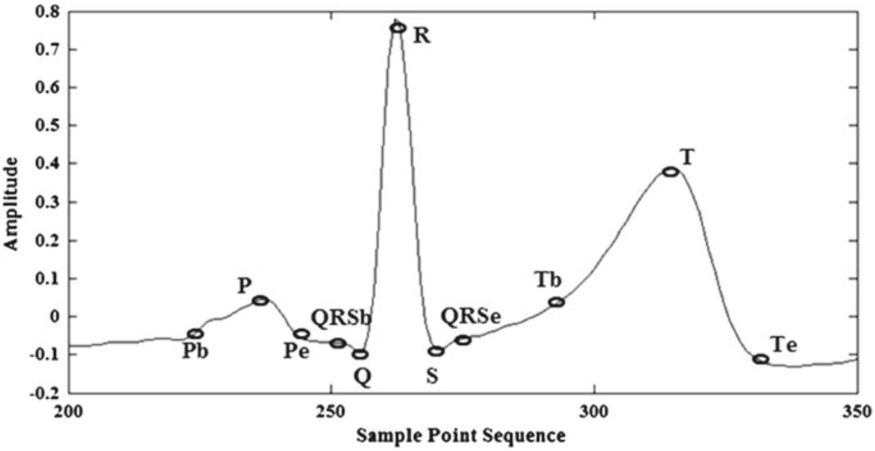


Fig. 14.1 Shows fiducial points extracted from each ECG heartbeat

and occurred in this temporal order (Fig. 14.1). These 11 fiducial points include three peak points (P, R and T), two valleys (Q and S) and the six onsets and offsets for the three waves. Hence, fiducial based features represent the temporal and amplitude distances between fiducial points along with angle features [2, 3].

On the other hand, non-fiducial based approaches investigate the frequency contentof ECG data. For example, non-fiducial features can be wavelet coefficients, discrete cosine transform coefficients…etc. it is worth mentioning that most of non-fiducial approaches need only the detection of the R peak which is considered the easier point to detect due to its strong sharpness and for some approaches no detection is needed at all [2, 3].

In this chapter, not only the existing ECG based approaches are investigated but also a wavelet based approach is proposed. Astructure of wavelet coefficients derived from the decomposition of extracted R-R intervals using 5-level biorthogonal wavelet (bior 2.6) transform was examined for subject identification using Radial Basis Function (RBF) neural network classifier. Moreover, a two-phase reduction stage for the wavelet coefficients is introduced. The first phase is a global phase inspired from that the frequency content of the ECG after filtering is concentrated in low frequencies [1–40 Hz]. Hence, the impact of successively excluding each of details parts of the structure on the classification accuracy was evaluated, in order to maintain only the indispensable details parts. Further reduction was done through the second phase which provides more local deep insight to the significance of each of the preserved coefficients from the first phase using information gain feature selection approach [4]. For more reliability and robustness, the validation of all our experiments was done using four Physionet datasets [5–8] and the evaluation was drawn on the basis of measuring quantities, such as subject identification (SI), heartbeat recognition (HR), false acceptance/false rejection rate (FAR/FRR), and generalizability to other datasets.

14.2 Heart Fundamentals

In this section, a brief discussion is given about the heart and its function as follows:

14.2.1 Heart Anatomy

The heart is a powerful muscle that lies in the chest cavity between the lungs and is slightly larger than the one's clenched fist. It consists of four chambers, many large arteries, many veins and valves. A muscle in some form divides the heart into two cavities: the left cavity of the heart receives oxygenated blood from the lungs and pumps it to the rest of the body, while the right cavity of the heart receives deoxygenated blood from all parts of the body except for the lungs. Each side is partitioned into two chambers, the atrium (upper chamber) and ventricle (lower chamber). Thus, the four chambers are called atria and ventricles. A heartbeat is the physical contraction of the heart muscles for pumping blood [9].

14.2.2 Flow of Blood

Blood veins carry blood to the heart from the rest of the body. Carbon dioxide and cellular waste products are hold by the blood. The blood goes into the right atrium and then to the right ventricle, where it is then pumped to the lungs to dispose of wastes and receive a fresh oxygen supply. From the lungs, the blood returns to the heart. It returns to the left atrium and then to the left ventricle. The blood is then pumped out of the heart by the left ventricle into the aorta (Fig. 14.2). Thereafter, the aorta sends this blood to small arteries, which carry the oxygen-rich blood to the rest of the body [9].

14.2.3 The Electrical Activity of the Heart

The heart rate per minute is normally between 60 and 100 times. This rate is set by a small collection of specialized heart cells called the Sinoatrial (SA) or sinus node, which is located in the right atrium. The *SA node* is the heart's "natural pacemaker." It discharges by itself without control from the brain. Thereafter, with each discharge two events occurred: both atria contract and then an electrical impulse passes through the atria to reach another area of the heart called the Atrioventricular (AV) node, which exists in the wall between the two ventricles. The AV node it acts as a relay point to further propagate the electrical impulse. From the AV node, an electrical wave passes to ventricles, causing them to contract and pump blood. The blood from the right ventricle goes to the lungs, and the blood from the left ventricle goes to the body [10].

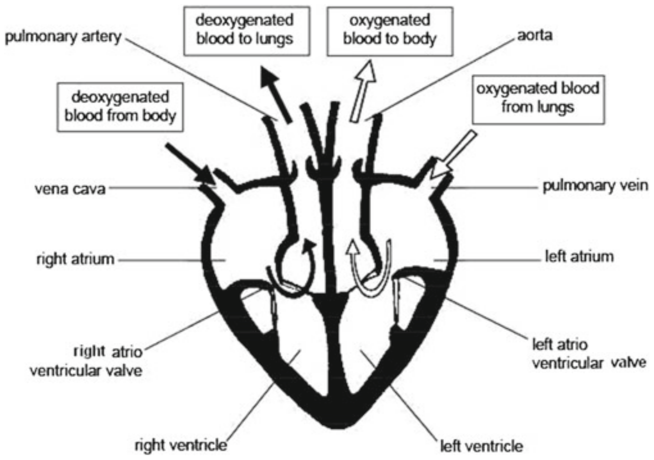

deoxygenated blood to lungs
oxygenated blood to body
pulmonary artery
aorta
deoxygenated blood from body
oxygenated blood from lungs
vena cava
pulmonary vein
right atrium
left atrium
right atrio ventricular valve
left atrio ventricular valve
right ventricle
left ventricle

Fig. 14.2 A heart diagram that illustrates the flow of blood in and out of the heart

14.3 Electrocardiogram

In 1893, Willem Einthoven introduced the term electrocardiogram (ECG) for the first time. ECG records the electricalactivity of the heart by placing electrodes (up to 12 electrodes) at various strategic body points such as chest, neck, arms, and legs. The results of the impulses are displayed on a computer monitor and can be printed onto graph paper. For each heartbeat, the ECG traces three complex waves: P, QRS and T waves.

The Atria contractions (both right and left) show up as the P wave, its duration is less than 120 milliseconds. The spectral characteristic of a normal P wave is usually considered to be low frequency, below 10–15 Hz; while the ventricular contractions (both right and left) show as a series of three waves, Q-R-S, known as the QRS complex, its durationis about 70–110 milliseconds in a normal heartbeat, and has the largest amplitude of the ECG waveforms. Due to its steep slopes, the frequency content of the QRS complex is considerably higher than that of the other ECG waves, and is mostly concentrated in the interval of 10–40 Hz; finally, the third and last common wave in an ECG is the T wave which reflects the electrical activity produced when the ventricles are recharging for the next contraction. The T wave duration is about 300 milliseconds after the QRS complex. The position of the T wave is strongly dependent on heart rate, becoming narrower and closer to the QRS complex at rapid rates [2, 10].

Hence, ECGis a valuable source of information, and it has been used as a reliable diagnostic tool. It carries information about heart rate, heart rhythm and morphology. Interpreting ECG reveals if the electrical activity is normal or slow, fast or irregular [2, 9, 10]. Moreover, it can indicate the weaknesses or the damage in various parts of the heart muscle [9, 10]. Finally, in the last two decades, it has been discovered that crucial information about human identity can be also inferred from an ECG recording .which in turns, open up the possibility of using ECG as a new biometric trait for human identification/verification [2, 3].

14.3.1 ECG and Biometrics

Biometrics is a scientific approach for personality identification. There are two general classes of biometrics: physiological and behavioral. Physiological biometrics encompasses anatomical traits that are unique to an individual such as: retina, iris and fingerprint, while behavioral biometrics considers with functional traits such as: signature, keystroke and gait. Although physiological biometrics has recorded lower error rates than behavioral biometrics, physiological biometrics requires dedicated and expensive hardware. In addition, they are less user acceptable than behavioral biometrics. Recently, a new alternative branch of biometrics which has been growing up over the past two decades is the utilization of biological signals such as the electroencephalogram (EEG) and electrocardiogram (ECG) (our interest in this chapter) as biometric traits [1].

14.3.2 ECG Motivation

The main potential benefits of utilizing ECG as a biometric trait can be summarized as follows:

(1) Uniqueness, The nature of an ECG waveform makes it carry the physiological and geometrical differences of the heart yielding a significant inter-variability among individuals [2, 3].
(2) As a biological signal, ECG not only provides a distinctive mark for different individuals as the other biometric traits, but also it distinguishes itself by providing an aliveness indicator [1–3].
(3) Circumvention, by comparing ECG to other biometric traits, it is found that it is more difficult to be spoofed or falsified and more universal since it is a vital signal [1–3].
(4) ECG is less expensive and more user acceptable than physiological traits. In addition,it is more reliable than *behavioral* traits [1–3].

14.3.3 ECG Challenges

A central challenge regarding utilizing ECG is its stability over time, in other words, the intra-subject variability which has a significant impact on the performance of large-scale real time systems. Hence, sources of intra-subject variability must be investigated, in order to successfully explore ECG based approaches that are invariant to them [2, 3]. The main known sources can be described as follows:

(1) ECG is susceptible to variety of noise sources which sometimes overlap with the ECG spectra [11]. Hence, filtering the ECG records with band path filter with cutoff frequencies 0.5 or 1 and 40 (expected ECG spectra) may not properly reduce the noise. Thus, increasing the intra-subject variability, especially when extraction of fiducial features is considered.
(2) Extracting fiducial features from an ECG heartbeat requires the detection of the 11 fiducial points: P, Q, R, S and T along with the onsets and the offsets of the three complexes. Actually, there is no universal acknowledged rule for defining exactly where the onset and the offset of each complex wave exist. Hence, for the same ECG record, even cardiologists cannot provide exactly the same locations for the marks of the wave boundaries [2]. Thus, developing accurate detection algorithms for those points without well-known standards is very challenging. One solution is to consider approaches that can dispense with these points detection (e.g. wavelets, autocorrelation. . . etc.).
(3) Heart rate variability: Heart rate is the number of heartbeats per unit of time, typically expressed as beats per minute (bmp). Variations in heart rate yield increasing or decreasing in the QT interval duration (the interval between the beginning of the QRS wave and the end of the T wave) which in turns has impact on the accuracy of the extracted features. Thus, developing ECG based approaches that are invariant to Heart rate variability is needed [12].
(4) Cardiac irregularities: the most challenging issue is not in the low-frequently severe cardiac disorders which require sensitive medical care in hospitalsor those disorders which cause permanent changes in the individual ECG and can be resolved by updating the user stored template in the system in some way [3, 13]. The challenging issue is to develop an ECG based approaches that is invariant to the non-critical everyday cardiac irregularities, for example premature ventricular contraction (PVC) which was considered by Agrafioti and Hatzinakos [13] in their proposed system.

14.4 Existing ECG Based Biometric Systems

A typical ECG based biometric system (identification or verification) usually includes three stages: (1) Pre-processing stage, where noise is reduced and fiducial points are detected if necessary; (2) Feature extraction, where fiducial or non-fiducial features areextracted; (3) Classification, where theconsidered features are fed into a classifier. However, a feature reduction stage can be added before the classification stage as shown in Fig. 14.3 to evaluate the available feature set and select only the significant relevant features with the aim of reducing the feature dimension, while preserving the system performance.

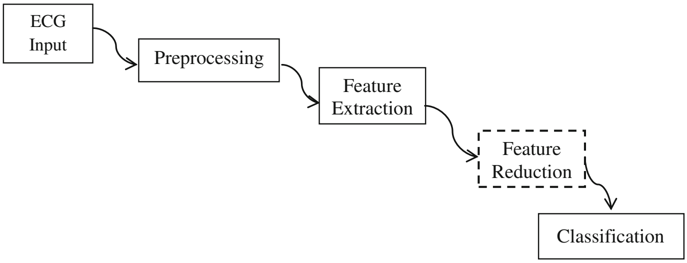


Fig. 14.3 General ECG based system architecture

Moreover, regarding classification there are two cases: (1) human identification, a 1: N mapping, the assumption is that the subject exists within the database and the classification task is to find the closest match to the user requesting access; (2) human verification, a 1:1 mapping, the task is to compare the access request with the stored model of the user, a much easier task than identification.

According to the utilized features, ECG based biometric systems have two main classes: fiducial [2, 11, 14–24] or non-fiducial systems [2, 13, 25–46]. However, few existing systems combined both fiducial and non-fiducial features together [2, 25, 45, 46]. A brief discussion about the key existing systems in the literature is provided in the next sections along with an analysis about their advantages and drawbacks.

14.4.1 Fiducial Based Systems

Fiducial features represent duration, amplitude differences along with angles between 11 fiducial points detected from each heartbeat. These points are three peaks (P, R, and T), two valleys (Q and S) and six onsets and offsets as shown in Fig. 14.1. The first fiducial based system was that developed by Forsen et al. [14] in 1977. Later on, many fiducial based systems have been emerged. A brief review of the key published works is provided below along with a summary is given in Table 14.1 for comparison.

Biel et al. [15] were among the first to examine the applicability of analyzing ECGs for human identification. A set of 30 temporal and amplitude features were extracted from heart beats. The features are directly extracted from SIEMENS ECG equipment which is considered a lack of automation (a major drawback of the system). Features are then reduced through experiments and correlation matrix analysis. A multivariate analysis-based method was utilized for classification and a 100 % human identification rate was achieved on 20 subjects.

Kyoso and Uchiyama [16] proposed detecting the locations of the fiducial points of the ECG signal using the second derivative waveform. Only four duration features were extracted from each heartbeat, which are P wave duration, PQ interval, QRS interval and QT interval. Every possible two feature combination from the four extracted features was used with Discriminant analysis for classification,

Table 14.1 Summarizes the key existing fiducial based systems discussed in this study

References	Normalization of features	# Of fiducial features	# Of reduced features	Classifier	# Of subjects	Results (%)	Comments
[15]	–	30	10	multivariate analysis-based method	20	SI = 100	lack of automation
[16]	–	4	2	Mahalanobis distance criterion	9	SI = 100 HR = 76−99	Only two features
[17]	–	6	–	MLP and SFA	10	MLP, HR = 97.6 SFA, HR = 86.4	None of the P or T wave features were considered
[11]	Features were divided by full heartbeat length	15	12	LDA	29	SI = 100 HR = 81	Only interval features were considered
[18]	Features were divided by full heartbeat length	19	–	Template matching	50	100	Short records were used for testing the system
[19]	–	14	–	Bayes classifier	502	97.4	Large dataset
[20]	–	24	9	Mahalanobis Distance criterion	16	100	–
[21]	–	7	–	Deviation threshold	20	97	Only amplitude features were considered

(continued)

Table 14.1 (continued)

References	Normalization of features	# Of fiducial features	# Of reduced features	Classifier	# Of subjects	Results (%)	Comments
[22]	Features were divided by full heartbeat length	28	–	MLP and RBF	13	Both classifiers SI = 100 RBF provided better HR = 97.8	Stability, generalization and imposter rejection tests were considered
[23]	Features were divided by full heartbeat length	28	21	RBF	13	SI = 100	Stability, generalization and imposter rejection tests were considered
[24]	Features were divided by full heartbeat length	36	23	RBF	13	SI = 100	Only Features derived from peaks and valleys points were considered

Mahalanobis' generalized distance was applied as a criterion for discrimination. Experiments were conducted using nine subjects. The highest accuracy was achieved using the combination of QRS interval and QT interval features and also the combination of QRS interval and PQ interval features provided accurate results. These results encouraged the authors to suggest the use of a combination of these three features for better accuracy.

Palaniappan and Krishnan [17] proposed an approach that utilizedonly six features: R-R interval, R amplitude, QRS interval, QR amplitude, RS amplitude and finally a form factor of the QRS segment was introduced. Hence, except for the RR interval, all features were concentrated in only the QRS wave. The Form factor is a measure of the complexity of the QR signal. The performance of Multilayer Perceptron and Simplified Fuzzy ARTMAP (SFA) neural network classifiers was compared. Experiments were conducted using long-term records of 10 subjects. 97.6 % was the best identification accuracy achieved by the MLP classifier, while 84.5 % was the best result achieved by SFA. However, the SFA has the advantage of incremental learning ability, no need to be retrained if new subject is added.

Israel et al. [11] proposed first a set of descriptors that characterize the fiducial points of the ECG. After detecting the fiducial points from each heartbeat, 15 distance featureswere then extracted from distances between those points. Since The distances between the fiducial points and the R position varies with heart rate. Distances were normalized by dividing them by the full heart beat duration. A Wilk's Lambda method was applied for reduction of feature number to 12 features.Classification was performed on heartbeats using standard linear discriminant analysis (LDA). The system was tested on dataset of 29 subjects, and a 100 % human identification accuracy was achieved, while the heartbeat recognition accuracy was 81 %. Moreover, the effect of varying ECG lead placement and the invariance of the normalized extracted features to individual's anxiety states (i.e. reading aloud, mathematical manipulation and driving) was successfully tested and validated.

Singh and Gupta [18] developed a method in a series of steps: (1) a preprocessing step where successive low and high pass filters were applied to reduce noise artifacts; (2) an existing QRS and new P and T delineators developed by the authors were utilized to detect QRS, P and T waves, respectively from ECG records; (3) from each heartbeat,19 stable interval, amplitude and anglefeatures were computed; (4) template matching and adaptive thresholding were utilized for classification. The accuracy achieved by the system was 99 % with FAR 2 % and FRR 0 % on the data set of 50 healthy individual ECG. FRR was tested on data from the same training record, while the FAR was tested by data from another database.

Fourteen features were extracted from each heartbeat by Zhang and Wei [19]. Thereafter, principle component analysis PCA was applied to reduce dimensionality. Bayes classifier was utilized to maximize the posterior probability given prior probabilities and class-conditional densities. A data base including 502 ECG recordings are used for development and evaluation. Each ECG recording is divided into two segments: a segment for training, and a segment for performance evaluation. The proposed method was found superior to Mahalanobis' distance by 3.5 to 13 % according to the considered lead and achieved 97.6 % subject identification accuracy.

Gahi et al. [20] applied a Bessel filter with a cut-off frequency of 30 HZ to reduce noise. The filtered ECG data were then processed to extract 24 temporal and amplitude features. The extracted features were ranked and then reduced to nine significant features using information gain ratio criteria. Mahalanobis distance-based classifier was used to identify the individuals. For each heartbeat, the Mahalanobis distance between it and the set of templates stored in the system database was computed. The template resulting in the smallest distance was considered to be a match and hence, increments the matching score by one. This process was repeated for 150 heartbeats (approximately two minutes). The template (individual) with the highest score was considered to be a match. The results showed that the system can achieve a 100 % identification rate with the reduced set of features. The system was tested using dataset of 16 subjects.

Singla and Sharma [21] suggested utilizing only seven amplitude features without any duration features. A deviation threshold was employed for classification. The system was tested on a dataset of 20 subjects and a 97 % subject identification accuracy was achieved. Moreover, the system was tested for FAR and FRR, the results were 3.21 and 3 % respectively. For each subject there are ten samples for testing. The FRR is the number of rejected samples from the ten, while the FAR is derived by comparing template of each person with 190 samples of the others. Hence, the FRR and FAR were tested from the same database.

Tantawi et al. [22] proposed the extraction of a set of 28 features. Thereafter, both feed-forward (FFN) and radial basis (RBF) neural networks were utilized as classifiers for comparison. The RBF was superior and provided the best results. Critical issues like stability over time, generalization and imposter rejection test (FAR and FRR) were all considered. The system was successfully tested using a set of 13 subjects from PTB database [5] who have more than one record for each (stability test), while the remaining 38 subjects of PTB database (only one record for each), 40 subjects of fantasia database [8] and 24 subject of MIT_BIH database [6, 7] were utilized for the imposter rejection and the generalization tests. In addition, the same authors [23] also investigated reducing the cardinality of the feature set using a variety of established methods, such as principle component analysis, linear discriminant analysis and selection methods based on information gain and rough sets, which was proposed for the first time in literature. The RBF neural network was used as a classifier. The system was tested in the same way as in [22] using the same databases and the same critical issues were considered. The 21 features selected by rough sets (25 % reduction)and outperformed the other reduction methods.

Tantawi et al. [3] introduced the PV set of fiducial features. This set includes 23 features and distinguishes itself by including only interval, amplitude and angle features that are derived only from the peaks (P, R and T) and valleys (Q and S). Hence, the fiducial detection process was relaxed to include only the five prominent points (peaks and valleys) which are easier and can be more accurately discerned than onsets and offsets. RBF neural network was utilized as a classifier. The efficacy of the proposed set was examined against a super set of 36 features and comparable results were achieved. The system was tested using 13 subjects of PTB database (those of more than one record for each) and a generalization test was done using the remaining 38 subject of PTB database.

14.4.2 Non-Fiducial Based Systems

Non-fiducial features capture the holistic patterns by examining the ECG data in the frequency domain. Non- fiducial features are like wavelets coefficients, autocorrelation coefficients, polynomial coefficients, etc. Hence, no fiducial detection is needed, except for the R peak which is the sharpest fiducial point and the easiest to detect. The R peak is needed by some approaches to define heartbeats by considering RR intervals. However, some non-fiducial approaches don't need any fiducial detection. They operate on an ECG record after dividing it into segments. The duration of such segments is usually found empirically [2]. A brief review for the key existing non-fiducial systems organized according to the utilized approach for non-fiducial feature extraction is provided below along with a summary is given in Table 14.2 for comparison.

14.4.2.1 Raw ECG Samples Based Approaches

A two dimensional heart vector known as the characteristic of the electrocardiogram was proposed by [26]. This heart vector has been constructed by QRS data from three Leads: I, II and III. Moreover, the first and the second derivative of the heart vector were also calculated. The distance between two heart vectors as well as their first and second temporal derivatives was considered for classification. The achieved verification accuracy was 99 %. In addition, 0.2 % false acceptance and 2.5 % rejection rate were achieved on a database of 74 subjects.

Mai et al. [27] proposed a system that also dispense with the P and the T wave. The data samples of the QRS wave were solely fed into Multilayer perceptron (MLP) and Radial basis functions (RBF) for comparison. Eighteen subjects from MIT_BIH dataset were utilized in the experiments. The results showed that both classifiers achieved 100 % subject identification accuracy with QRS classification accuracy 99.6 % and 97.5 % for MLP and RBF respectively.

14.4.2.2 Wavelets Based Systems

Chan et al. [28] introduced a novel distance measure based on the wavelet transform for identifying P, QRS and T waveforms (heartbeats). The waveforms were detected automatically using the multiplication of backward differences algorithm and temporally aligned using a cross-correlation measurement,and the signal-to-noise ratio

Table 14.2 Summarizes the key existing non-fiducial based systems discussed in this study

References	Non fiducial approach	# Of reduced features	Classifier	# Of subjects	Results (%)	Comments
[26]	Heart vector constructed by QRS data from three leads along with its 1st and 2nd derivatives		Euclidean distance	74	SI = 99	P and T waves were ignored
[27]	Raw samples of QRS wave		MLP and RBF	18	SI = 100 MLP: HR = 99.6 RBF: HR = 97.5	P and T waves were ignored
[28]	Wavelet distance measure proposed to examine the similarity between extracted heartbeats			50	95	Fiducial detection is needed
[29]	256 coefficients resulted from biorthogonal wavelet decomposition applied to 256 averaged RR intervals	–	Feed forward Neural network	23	100	High feature dimension
[30]	512 coefficients resulted from applying 9-level Haar wavelet decomposition to ECG segments	–	Euclidean distance	35	Normal, SI = 100 Patient, SI = 81	Very high feature dimension and not all subjects were normal

(continued)

Table 14.2 (continued)

References	Non fiducial approach	# Of reduced features	Classifier	# Of subjects	Results (%)	Comments
[31]	Reconstructed signals at scale 3 after applying dyadic wavelet transform (DWT)	–	correlation	27	99.6	Each subject is represented by one heartbeat only
[32]	5-level discrete wavelet transform using db3 wavelet from Daubechies was applied to RR intervals	–	Random forest method	80	SI = 100	High feature dimension
[2]	The AC sequence of ECG segments	First C DCT coefficients of auto-correlated ECG segments	Euclidean distance	13	SI = 100 HR = 94.4	Generalized to MIT_BIH database (14 subjects)
[13]	The AC sequence of ECG segments	LDA coefficients	Euclidean distance	56	SI = 96.2	The system can discard abnormal beats (PVC or APC)
[33]	Fourier transform applied to three complex waves P, QRS and T		Neural network	20	SI = 97.15	Heart rate variation was considered

(continued)

Table 14.2 (continued)

References	Non fiducial approach	# Of reduced features	Classifier	# Of subjects	Results (%)	Comments
[34]	Morphological synthesis technique			10	SI = 98	Heart rate variation was considered
[35]	Polynomial approximation for the 3 complex waves P, QRS and T		Euclidean distance	15	SI = 100	Fiducial detection is still needed
[36]	Extended Kalman filter		Log-likelihood ratio	13	SI = 87.5	System robust for signal to noise ratio above 20 dB
[37]	8-bit uniform quantization to map the ECG samples to strings from a 256-symbol alphabet		Ziv-Merhav cross parsing algorithm	19	SI = 100	Emotional state variation is considered
[38]	An AR model of order 4 applied to ECG 50 % overlaped segments		KNN classifier	12	SI = 100	–
[39]	Fusing temporal and Cepstral information		SVM classifier	18	SI = 98.26	–

was improved by ensemble averaging. Three different quantitative measures: percent residual difference, correlation coefficient, and a novel distance measure based on the wavelet transform were used and compared for classification. The system was tested on a dataset of 50 healthy subjects, three records per subject, one used as enrollment data and the others for testing. A classification accuracy of 95 % was achieved by the wavelet distance measure, outperforming the other methods by over than 10 %.

Wan and Yao [29] suggested a verification system that utilized a set of 40 heartbeats which were extracted for each subject. By averaging every four heartbeats, the 40 heartbeats (RR intervals) were reduced to a set of 10 heartbeats. Each of the 10 heartbeats/subject was decomposed into 256 biorthogonal wavelet coefficients. These coefficients were used as input vectors to a 3 layer feed-forward neural network. The network input layer accepts two heartbeats wavelet coefficients (512-element vector) as the input vector and trained to verify if they are for the same person or not. The system was trained on a database of 23 persons and tested by ECG records for 15 subjects recorded after few months of the training records. All the 15 subjects in the experiments were successfully verified.

Chiu et al. [30] applied 9-level Haar wavelet decomposition to input signal segments which were constructed by concatenating data points form the backward 43rd point of the R peak to the forward 84th point from four heart beats yielding input segments of 512 points for each. The resulted 512 wavelet coefficients represented the feature vector employed for verification. The system was trained on a database of 35 subjects and 10 arrhythmia patients, 100 % verification rate was achieved for normal subjects and 81 % for arrhythmia patients using Euclidean distance as a criterion for verification.

A new wavelet based framework was introduced and evaluated by Fatemian and Hatzinakos [31]. The proposed system utilized a robust preprocessing stage that was directly applied on the raw ECG signal for noise handling. Furthermore, one of the novelties of this system was the design of personalized heartbeat template so that the gallery set consists of only one heartbeat per subject. A dyadic wavelet transform (DWT) was applied to the raw ECG signals, and then the signals were reconstructed at the third scale where most of the signal energy is retained. Further smoothing via moving average was applied. The heartbeats were then resampled, normalized and using the median of the aligned heartbeats, the heartbeat template for each subject is constructed. Finally, classification was accomplished based on the correlation among templates. The system was evaluated over two common databases: MIT-BIH (13 subjects) and the PTB (14 subjects), and an accuracy of 99.6 % was achieved.

A5-level discrete wavelet transform using db3 wavelet from Daubechies familywas applied by Belgacem et al. [32] on normalized R-R cycles. The random forests method was used for verification. A dataset of 80 subjects were used for testingthe system. The dataset includes 60 subjects from Physiobank databases and 20 subjects from dataset collected by the authors. The proposed system has achieved 100 % verification rate.

14.4.2.3 AC Based Systems

Wang et al. [2] proposed a new approach for human identification using ECG signals based on utilizing the coefficients from the Discrete Cosine Transform (DCT) of the Autocorrelation (AC) sequence of ECG data segments. The main advantage of this approach, it doesn't require any waveform detections (fiducial locations) or even extracting of individual ECG pulses (heartbeats). The ECG records were first filtered, for noise reduction. The normalized autocorrelation function of each record was estimated over a considered windowof arbitrary length N, origin and M autocorrelation lags. The DCT of the windowed autocorrelation was calculated and the first C number of significant coefficients is selected. Normalized Euclidean distance was considered for classification. By experiments, the values of M and C were empirically selected and it was found that the length N must be longer than the average heartbeat length so that multiple pulses are included. The system was tested on 13 subjects from the PTB database and generalized on 14 subjects from the MIT_BIH database. The subject identification accuracy was 100 % for both and the window recognition accuracy 94.4 % for PTB and 97.8 % for MIT_BIH.

Agrafioti and Hatzinakos [13] proposed an identification system that is robust to common cardiac irregularities such as premature ventricular contraction (PVC) and atrial premature contraction (APC). ECG records were filtered using Butterworth band pass filter for noise removal. Criteria concerning the power distribution and complexity of ECG signals were defined using DCT to discern abnormal ECG recordings, which were not employable for identification. Thereafter, Features were extracted from autocorrelation (AC) coefficients of healthy ECG records using a windowing technique to eliminate the shortcoming of localizing fiducial points. PCA and LDA were tested for dimension reduction and Euclidian distance was utilized as a criterion for classification. Experiments were carried out using 56 subjects and results indicated a recognition accuracy of 96.2 % using LDA for dimension reduction.

14.4.2.4 Other Non-Fiducial Approaches

Saechia et al. [33] investigated the heart rate variations in their work. After the heartbeats were normalized and divided into P wave, QRS complex and T wave, the Fourier transform was computed globally on a heartbeat itself and all three waves for comparison. Thereafter, the spectrum was then fed into a neuralnetwork for classification. The experiments were conducted using 20 subjects. The best result achieved was 97.15 and It was revealed that false rate was significantly lower (17.14 to 2.85 %) by using the three waves instead of the original heartbeat.

Instead of locating P, Q, R, S, and T peaks of heartbeats in filtered ECG signals, Molina et al. [34] used the morphology of R-R segments.The R-peaks were taken as reference since the R peak appeared in all electrodes and its sharpness make it easy to detect. Furthermore,all the elements of a PQRST-cycle were contained within an R-R segment. The authenticity of a given R-R segment was decided by comparing it to a matching R-R segment morphologically synthesized from a model characterizing the identity to be authenticated. The morphological synthesis process used a set of R-R segments (templates), recorded during enrolment at different heart-rates, and a time alignment algorithm. This ensures the authentication to be independent of the, usually variable, heart-rate. The optimum average equal-error-rate obtained in our experiments is 2 %. The experiments were done using 10 subjects.

A polynomial distance measure (PDM) method for ECG based biometric authentication was proposed by Sufi et al. [35]. After the three complexes P, QRS and T were detected from each heartbeat and differentiated, an approximated polynomial equation was generated and the coefficients were stored for each wave. Thereafter, the coefficients of the three waves were concatenated to form a feature vector for a specific heartbeat. A match was achieved when the Euclidean distance between two feature vectors was below certain threshold. The system was tested on a database of 15 subjects with a 100 % subject identification accuracy. For each subject, both training and testing beats were extracted from the same ECG record.

Ting and Salleh [36] utilized an Extended Kalman filter to represent the ECGin a state space form with the posterior states inferred by it. The Log-likelihood score was employed for matching between stored templates and testing ones. The best accuracy achieved was 87.5 % on a dataset of 13 subjects.

After extracting ECG heartbeats, Coutinho et al. [37] applied 8-bit uniform quantization to map the ECG samples to strings from a 256-symbol alphabet. Ziv-Merhavcross parsing algorithm was employed for classification. The stored template of minimum description length from the test input was found (given the strings in the particular template). 100 % identification accuracy was achieved using 19 subjects.

Ghofrani and Bostani [38] segmented the ECG signals with 50 % overlap and an AR model of order four was derived. The resulted coefficients were utilized as features along with the mean power spectral density(PSD) of each segment. An accuracy of 100 % was achieved using KNN classifier and a 12-subject dataset. This method was superior to utilizing nonlinear features such as Lyapunovexponent, ApEn, Shanon Entropy, and the Higuchi chaotic dimension.

Li and Narayanan [39] introduced a hybrid approach by fusing temporal and Cepstral information. The Hermite polynomial expansion was utilized to transform heartbeats into Hermite polynomial coefficients which were then fed into SVM with alinear kernel for classification. By simple linear filtering and modeling by GMM/GSV (GMM supervector), Cepstral features were derived. The proposed fusion was at score level through a weighted sum strategy. The best result achieved was 98.26 % with a 0.5 % ERR. Experiments were conducted on a dataset of 18 subjects from MIT_BIH [6, 7] dataset.

14.4.3 Combined Fiducial and Non-fiducial Based Systems

Some of the existing systems in the literature proposed combining both fiducial and non-fiducial features in some hierarchical scheme for more robustness and efficacy.

Table 14.3 Summarizes the key existing combined fiducial and non-fiducial systems discussed in this study

Ref.	Feature extraction approach	Classifier	# of subjects	Results %	Comments
[45]	QRS complexes for first stage7 QRST temporal and amplitude features for second stage	Template matching and DBNN for second stage	20	SI = 100	P wave was ignored
[46]	QRS complexes for first stage17 temporal and amplitude features for second stage	Template matching and distance measure classifier	168	SI = 4 95..3	# of needed features increased when # of subjects increased.
[2]	9 analytical features in the first stage PCA coefficients for second stage	LDA for first stage nearest neighborhood classifier for second stage	13	SI = 100 HR = 98.9	The system was generalized to another 14 subjects SI = 100 HR = 99.4

In this section, the key published studies are presented below and they are summarized in Table 14.3 for comparison.

Shen et al. [45] proposed a two-step schemefor identity identification. First, a template matching method was employed to compute the correlation coefficient for comparison of two QRS complexes. Thereafter, a decision-based neural network (DBNN) approach was then applied with seven temporal and amplitude features extracted from QRST waveas an input to give final decision from the possible candidates selected with template matching. The best result was 95 % for template matching, 80 % for the DBNN, and 100 % for combining the two methods on a dataset of 20 subjects.

Later on, a larger database of 168 normal healthy subjects was utilized by Shen [46]. For prescreening,template matching and mean square error (MSE) methods were compared and distance classification and DBNN compared for second-level classification. For the second-level classification, 17 temporal and amplitude features were used. The best identification rate was 95.3 % using template matching anddistance classification.

A two-stage hierarchical scheme that combines both analytical and appearance features was proposed by Wang et al. [2]. The first stage employed nine analytical features selected from a set of 21 features using Wilk's Lambda method and LDA for classification, while, the second stage employed PCA features extraction and a nearest neighborhood classifier. As a first step,only analytic features wereused to provide us the potential classes that the entry might belongs to. If all the heartbeats wereclassified as one subject, decision module output this result directly. If the heartbeats were classified as a few different subjects, the PCA based classification module which was dedicated to classify these confused subjects was then applied. The system was tested on 13 subjects of PTB database and it was generalized on 14 subjects of MIT_BIH database with 100 % identification accuracy for both and heartbeat recognition accuracy of 98.9 and 99.4 % for PTB and MIT_BIH respectively.

14.4.4 Fiducial Versus Non-fiducial Based Systems

Fiducial based systems require the computation of duration and amplitude differences between fiducial points along with the angles between them. Although such fiducial featuresare simply computed, they imply the accurate detection of the 11 fiducial points which is a very challenging task by itself. Fiducial points are prone to error due to the impact of noise or deficiencies in the detection algorithm itself, especially the onsets and the offsets since there are no universally acknowledged rules for detecting them. Such errors in the detection process may decrease the quality of the derived features, causing an increase in the intra-subject variability which in turns weakens the classifier performance. One solution suggested by Tantawi et al. [24] is to preserve features that need only peaks (P, R and T) and valleys (Q and S) in their derivation (since such points are easier and more accurately detected for their prominence) and dispense with the onsets and the offsets based features. The results of low-scale datasets revealed that excluding these features causes a slight decrease in the HR accuracy, but the SI is still preserved.

On the other hand, non-fiducial based approaches resolves the fiducial detection problem by considering only the detection of the R peak which is considered the easier point to detect due to its strong sharpness and for some approaches no detection is needed at all. However, non-fiducial approaches may result in a high dimension feature space (hundreds of coefficients), which in turn increases the computational overhead and requires more data for training (the size of data needed for training grows exponentially with the classifier input dimension [47]). Moreover, high dimension data usually encompass superfluous and irrelevant information that may weaken the performance of the classifier. In the literature, none of the existing systems have addressed this issue, especially in case of wavelet based approaches.

As mentioned in Sect. 14.4.3, some of the existing systems proposed combining fiducial and non-fiducial features in a hierarchical manner to improve performance. However, combining both fiducial and non-fiducial feature not only increases the computational load, but also it invalidates the main advantage of non-fiducial approaches which is relaxing the fiducial detection process to include the R peak or no detection at all is needed. Moreover, these studies didn't provide significant improvement comparing to the other studies that considered either fiducial or non-fiducial approaches solely.

Table 14.4 shows the consideration of the mentioned critical issues of ECG by the existing systems in the literature

References	Heartbeat (window) recognition accuracy	Training and testing from different records	Testing for FAR and FRR	Cardiac irregularities	Generalization to other databases
[15, 28, 44]	×	√	×	×	×
[16, 17, 27]	√	×	×	×	×
[19, 20, 30–32, 35, 38, 45, 46]	×	×	×	×	×
[18, 25, 33, 34, 36, 37, 39]	×	×	√	×	×
[11]	√	√	×	×	×
[2, 40]	√	√	×	×	√
[13]	√	√	×	√	×
[21]	√	√	√	×	×
[24, 29]	×	√	×	×	×
[41]	√	×	√	×	×
[22, 23]	√	√	√	×	√
[24]	√	√	×	×	√
[26, 43]	×	√	√	×	×

Finally, both fiducial and non-fiducial systems presented in the previous sections indicate that the accuracies are well above 95 % in most cases, giving more evidence on the reliability of ECG as a biometric trait. However, for robust large-scale real time biometric system, there are some critical issues that must be considered. By checking Table 14.4, it is very clear that none of the existing systems considered them all and some of the systems did not address any of them at all. These issues are such as: (1) heartbeat recognition (HR) accuracy, which is the number of beats or windows correctly classified as belonging to a particular subject, it gives indication to the degree of efficacy of the system; (2) stability over time, which means measuring the reliability of developed system, when it is tested by data recorded after some time (days, months or years) from the enrollment data, many of the existing systems utilized data for enrollment (training) and testing extracted from the same ECG record, one reason for this is the limited sources sometimes for data; (3) imposter rejection test (FAR and FRR test), it is a very crucial biometric issue which test the ability of the system to reject impostors (FAR), while allowing the authenticated users without much effort (multiple log in trials) (FRR). Except for [22, 23], the few existing systems that considered such issue didn't acquire the FAR and FRR in a manner that preserves robustness. For instance, the studies reported in [21, 26, 34, 39] measured the FAR and FRR from the same training database. While, Singh and Gupta [18] reported a 2 % FAR using one database (different from the one used for training), but reported their FRR results using data from the same records used for enrollment stage; (4) every day non-critical cardiac irregularities,none of the systems except the one proposed by [13] considered such issue which can affect the HR accuracy by increasing the intra-subject variability; (5) Generalization, which measures the ability of the system to preserve its efficacy on different datasets rather than the one used for training the system, few systems considered also this issue and limitation of data sources can be the main reason.

14.5 Methodology

The proposed methodology can be broken down into five main steps: (1) Data preparation; (2) preprocessing; (3) feature extraction; (4) feature reduction; (5) classification (subject identification).

14.5.1 Data Preparation

The famous published Physionet databases: PTB [5], MIT_BIH [6, 7] and Fantasia [8] were deployed for training and testing purposes. These databases provide one long ECG record for each subject except for the PTB database; it encompasses 2 sets of subjects: (1) PTB_1 includes 13 subjects with more than one record and some of them are few months (years) apart; (2) PTB_2 includes 38 subjects with only one record. Duration of PTB records varies from 1.5–2 min (minimum 100 beats). Thus, for more reliability and robustness PTB_1 dataset was employed for training and testing, while PTB_2 along with fantasia database (40 subjects) and MIT_BIH (24 subjects) were utilized to measure the ability of the proposed system to generalize to other datasets and to reject impostors.

The PTB_1 dataset was partitioned in three partitions: one set for training and two sets for testing. The training set contains 13 records, one for each subject. While the testing sets: ‘Test set 1’ contains 13 records (each of them is 100 beats) recorded on the same day of recording the training records but in different sessions, while ‘Test set 2’ contains nine records (belongs to six subjects) recorded after few months (years) of recording the training records.

14.5.2 Preprocessing

A Butterworth filter of second order with cutoff frequencies of 1 and 40 Hz was applied for noise reduction and baseline line removal.Regarding the detection of fiducial points, only the R peak is needed. The Pan and Tompkins algorithm [48] was applied for that purpose. Subsequently, R-R cycles were extracted from each record. All the R-R cycles were interpolated to the same length of 128 points. The amplitude of all points for each R-R cycle was normalized by the value of R peak into the range of [0–1]. From each training record, 40 cycles were chosen randomly. In order to avoid signal variations, each four cycles were averaged to one yielding 10 cycles for each subject utilized for training.

14.5.3 Feature Extraction

5-level discrete wavelet decomposition was applied to the selected RR cycles using discrete 'Bior 2.6' wavelet. 'Bior 2.6' belongs to bi-orthogonal wavelet family.This family of wavelets exhibits linear property which is necessary for signal reconstruction and has excellent localization properties both in time and frequency domains [49]. For the first level, thespectrum of each RR cycle is decomposed to low frequency region (approximation part) and high frequency region (details part). Thereafter, for each level, the approximation of the previous level is further decomposed into new approximation and details regions and so on until the last level. The resulted wavelet coefficient structure consists of six parts: five parts for the coefficients derived for the details region from each level and one part for the coefficients derived for the remaining approximation region in the last level. Figure 14.4 demonstrates the discretewavelet decomposition levels and the formation of the resulted wavelet coefficient structure. Meanwhile, Figure 14.5a, b illustrates an original averaged R-R cycle and its corresponding structure.

14.5.4 Feature Reduction

This step is crucial in order to reduce the high dimension of the resulted coefficient structure from the wavelet transformation and remove any superfluous or irrelevant information that may weaken the classifier performance. Furthermore, it decreases the computational overhead and time consumption which are very critical issues for real time systems. In this work, the feature reduction process was applied in a two-phase approach as follows:

(a) **Global Phase (Significance of the Structure Parts)**
After filtering, the frequency content of an ECG signal is usually concentrated in low frequencies [1–40 Hz]. Hence, are all the details parts of the wavelet coefficient structure (Fig. 14.2) indispensable? To answer this question, we followed a backward elimination approach, where we began with evaluating the whole structure (190 coefficients) as a feature vector then the details parts from level $1 - 15$ were excluded one by one until we have only the approximation part (16 coefficients) for evaluation as a feature vector in the end.

(b) **Local Phase (Significance of Coefficients)**

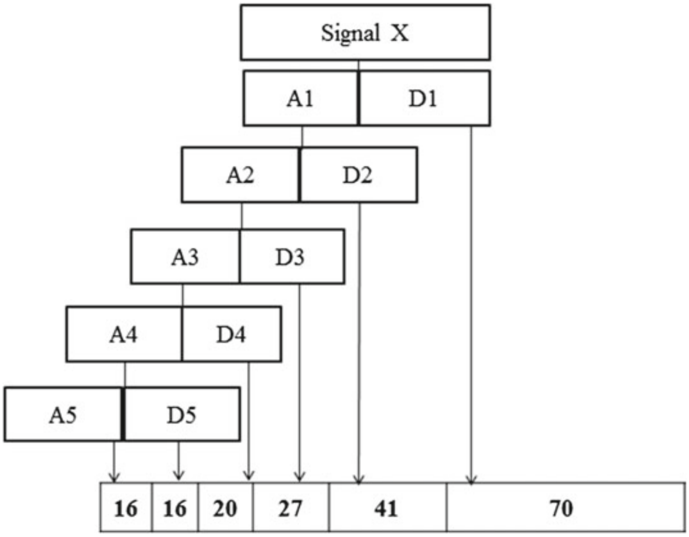


Fig. 14.4 The discrete wavelet decomposition levels and the formation of the resulted wavelet coefficient structure (190 coefficients from six parts)

In this phase, the significance of each coefficient of the preserved parts from the previous phase was independently evaluated using information gain ratio (IGR) criterion. The IGR [4] is a statistical measure that has been utilized in decision trees learning algorithms to provide a basis for selecting amongst candidate attributes in each step while growing the tree. IGR depends on the concept of entropy, which, in this context,characterizes the impurity of an arbitrary collection of examples S [4]. If the target attribute can take on c different values, then the entropy of S relative to this c-wise classification is defined as,

$$\text{Entropy (S)} = \sum_{\text{i}=1}^{\text{c}} -\text{p}_\text{i} \log_2 \text{p}_\text{i} \tag{14.1}$$

where p_i is the proportion of S belonging to class i. Using this formulation (1), one can simply define IGR of an attribute as the expected reduction in entropy caused by partitioning the examples according to this attribute [4]. More precisely, the information gain, Gain (S, A) of an attribute A, relative to a collection of examples S, is defined as

$$Gain(S, A) = Entropy(S) - \sum_{v\in} Values(A) \frac{|S_v|}{|S|} Entropy(S_v) \tag{14.2}$$

where Values(A) is the set of all possible values for attribute A, and is the subset of S for which attribute A has value v (i.e., = {s ε S|A(s) = v}). Thus, the features

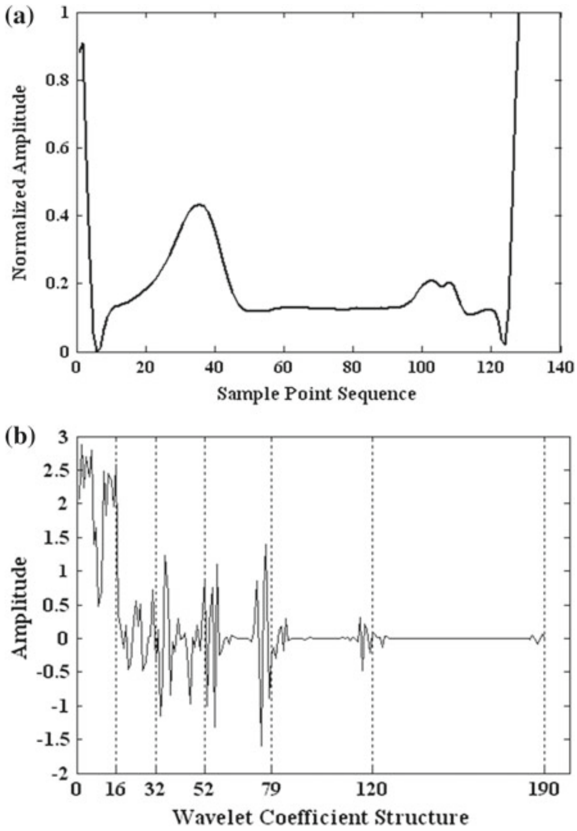


Fig. 14.5 **a** An original averaged RR cycle and **b** its wavelet coefficient structure (the bounds of each part are defined by *dashed vertical lines*)

are ranked according to their IGR. Initially, the selection algorithm begins with an empty set F of best features and then features are added from the ranked set of features until the classification accuracy begins to decrease or it becomes a specific selected value [4].

14.5.5 Classification

After the feature sets have been acquired, the considered set of features of a heartbeat is fed into the classifier in order to perform the classification task (i.e. associate a heartbeat to a particular subject ECG record). Due to its superiority in our previous work [22–24], Radial Basis Functions (RBF) neural network was employed here as a classifier. The RBF network is based on the simple idea that an arbitrary function y(x) can be approximated as the linear superposition of a set of localized basis functions φ(x) [47]. The RBF is composed of three different layers: the input layer in which the number of nodes is equal to the dimension of input vector. In the hidden layer, the input vector is transformed by a radial basis activation function (Gaussian function):

$$\varphi\left(x, c_j\right) = \exp\left(\frac{-1}{2\sigma^2}||x - c_j||^2\right) \tag{14.3}$$

where $||x - c_j||$ denotes the Euclidean distance between the input data sample vector x and the center c_j of Gaussian function of the jth hidden node; finally the outer layer with a linear activation function, the kth output is computed by equation

$$F_k(x) = \sum_{j=1}^{m} W_{kj}\varphi\left(x, c_j\right) \tag{14.4}$$

w_{kj} represents a weight synapse associates with the *j*th hidden unit and the *k*th output unit with m hidden units [47]. The orthogonal least square algorithm [50] is applied to choose the centers, which is a very crucial issue in RBF training due to its significant impact on the network performance. This algorithm was chosen for its efficiency and because there are very few parameters to be set or randomly initialized [50].

14.6 Results and Discussion

In this work, a two-phase feature reduction approach was applied. Experiments were conducted to establish which wavelet coefficients are most relevant to the task in hand. The efficacy of our approach was assessed in terms of Subject Identification accuracy (SI) and Heartbeat Recognition accuracy (HR), along with typical biometric quality indicators, such as false acceptance/false rejection rate (FAR/FRR).

SI accuracy is defined as the percentage of subjects correctly identified by the system, and the HR accuracy is the percentage of heartbeats correctly recognized for each subject. Note that when computing the average HR accuracy, only HRs of identified subjects were considered. A subject was considered correctly identified if more than half of his/her beats were correctly classified to him/her and a heartbeat is recognized by majority voting of the classifier outputs.

FAR/FRR is typically acquired by adjusting one or more acceptance thresholds. The acceptance thresholds are varied, and for each value, the FAR and FRR are computed. In this work, two thresholds were employed. One is typically used (call it Θ (1) is the value above which a heartbeat is considered classified, while another threshold (call it Θ (2) represents the minimum percentage of correctly classified beats needed for a subject to be considered identified. The set of imposters for computing FAR (approximately 100 subject ECG trails) were gathered from three different databases: including 38 subjects of PTB_2, 24 subjects of MIT_BIH and 40 subjects of Fantasia.

Table 14.5 The achieved SI and HR accuracies

	Test set 1		Test set 2	
	SI accuracy (%)	HR accuracy (%)	SI accuracy (%)	HR accuracy (%)
190 Coefficients	100	97.3	100	61.37
79 Coefficients	100	98	100	85.6

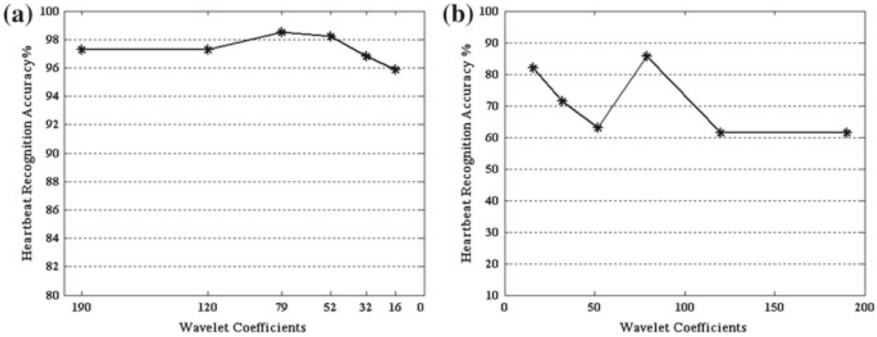


Fig. 14.6 The mean of the HR accuracy for each of the resulted reduced set from the first phase for **a** test set 1 and **b** test set 2

The following three subsections provide detailed discussion about the experiments and their results. The first two subsections are concerned with the experiments of the two phases of the reduction process respectively. Meanwhile, the last subsection is concerned with generalizing the results to other datasets.

14.6.1 The Global Phase (Significance of the Structure Parts)

(a) SI and HR Accuracies

The PTB_1 was processed as described earlier in the previous section. The wavelet coefficient structure (190 coefficients) of each of the 10 averaged RR cycles/subject were fed into an RBF classifier with spread 1 and SSE 0.7 for training. These values were found empirical during the experiment. Table 14.5 shows that the SI and HR achieved for test set 1 and test set 2. This experiment was repeated after excluding each of the details parts from level 1–5 yielding reduced sets of 120, 79, 52, 32 and 16 coefficients respectively. The results showed that the SI for both test sets is 100 % even when the structure became 16 coefficients only. While for HR, Fig. 14.6a, b shows the impact of reducing the details parts on the HR for both test sets.

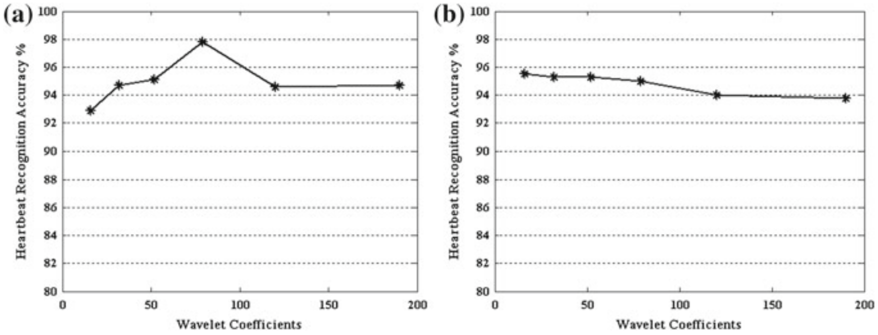


Fig. 14.7 The mean of the HR accuracy for each of the resulted reduced set from the first phase for **a** PTB and **b** Fantasia

From the achieved HR results, we can have the following observations:

(1) Excluding the details of the first level (d1) has no impact on the HR for both test sets which is expected since most of d1 coefficients are nearly zero values (Fig. 14.5b).

(2) For test set 1 (same day records), excluding the details parts has no significant impact on the HR average value. Even when the approximation part has been solely considered, the average HR is only 3 % less than the maximum value achieved, which shows the significance of the approximation part (low frequency region) as a rich source of information for the task in hand.

(3) Testing with test set 2 provides insight into the stability of the system over time, since it includes ECG signals recorded after few months (years) of recording the training signals. In this case, the HR showed less accuracies and more sensitivity to the exclusion of the details parts than that achieved for test set 1. Stability overtime (test set 2) is a challenging task since the one's ECG may vary due to changes in the one's physiological state or changes in the environment and the conditions of recording which usually contribute to the added noise. Excluding d1 and d2, while preserving d3, d4, d5 and a5 (79 coefficients) has achieved the maximum HR mean (Table 14.5).This result reveals that d1 and d2 are not informative for the task in hand and may carry information that confuses classifier rather than providing valuable information to it.

These results were achieved using only 13 subjects of PTB_1. What if we change the dataset or increase the number of subjects, can we maintain the same results? To answer this question, the experiment was repeated two times: once using the whole 51 subjects of the PTB dataset (13 subjects of PTB_1 + 38 of PTB_2) and the other with the 40 subjects of fantasia datasets. Since only one record is available for PTB_2 and fantasia subjects, the record for each subject will be partitioned into training and testing parts. The results showed that the SI for both datasets is also 100 % even when the structure becomes 16 coefficients only. Figure 14.7a, b shows the impact of reducing the number of coefficients on the HR. The results reveal that adding or removing d1 has no effect on the HR. the best HR was achieved also with

Table 14.6 The achieved values of the HR after thresholding, FRR and FAR

	The optimum values for thresholds	HR accuracy (%)	FRR (%)	FAR (%)
190 Coefficients	$\Theta_1 = 0.57$ $\Theta_2 = 80\,\%$	93.6	0	3.9
79 Coefficients	$\Theta_1 = 0.57$ $\Theta_2 = 85\,\%$	94.47	0	3.9

79 coefficients. Finally, for fantasia dataset where training and testing beats are from the same 2-h records for all subjects, excluding details parts provided stable HR. Thus, we have the same observations even after changing the dataset and increasing number of subjects.

(b) FRR and FAR

As mentioned at the beginning of the section, the thresholds Θ_1 and Θ_2 were adjusted for measuring the FAR and FRR in such a way to have the minimum FAR while preserving zero value for FRR as much as possible. The best result again was achieved with a reduced set of 79 coefficients (without d1 and d2). Table 14.6 shows the optimum values for Θ_1 and Θ_2 along with the best values achieved for FAR, FRR and the HR accuracy after thresholdingwith the whole number of coefficients (190) and the reduced set of 79 coefficients.

14.6.2 Local Phase (Significance of Coefficients)

(a) SI and HR Accuracies

After excluding d1 (70 coefficients) and d2 (41 coefficients) in the previous reduction phase, we are trying in this second phase to answer a question: are all the coefficients in the remaining parts significant to the classification process? To answer this question, the information gain was calculated for each of the 79 coefficients in the same way discussed in Sect. 14.5.4. A new threshold α is employed. It is defined as the information gain value below which the corresponding coefficients are excluded. We began with a very small value for α (no coefficients were out) and then it was increased gradually as long as the SI and HR accuracies were preserved. For comparison, the experiment was done with PTB_1 using RBF classifier in the same way done in Sect. 14.6.1. The results showed that the optimum value for α is 0.7 which excludes 13 coefficients. Seven of them from d3, while, the locations of the others were distributed over the remaining parts (d4, d5 and a5).This result gave more evidence that the significant information is concentrated in low frequencies. Thus, the length of the final set of coefficients is 66. As shown in Table 14.7, this final set improved the HR of test set 2 by 2.5 % compared to the result achieved by 79 coefficients (Table 14.5).

Table 14.7 The SI and HR achieved by the final reduced set of 66 coefficients

	Test set 1		Test set 2	
	SI accuracy (%)	HR accuracy (%)	SI accuracy (%)	HR accuracy (%)
66 Coefficients	100	98.5	100	87.87

Table 14.8 The generalization results using the whole PTB and Fantasia databases

	PTB (51 subjects)			Fantasia (40 subjects)		
	HR accuracy	FRR (%)	FAR (%)	HR accuracy	FRR (%)	FAR (%)
79 Coefficients	94	2	4.6	95.3	0	4
66 Coefficients	94	2	3	95	0	2.67

(b) FRR and FAR
The thresholds Θ_1 and $\Theta_2$were re-adjusted for measuring the FAR and FRR of the system after reducing the number of coefficients to 66. The results showed that no changes are needed to the thresholds values presented in Sect. 14.4.1. Moreover, the reduced set of coefficients retained the same values for FAR, FRR and averaged HR achieved by 79 coefficients (Table 14.2).

14.6.3 Generalization

The parameters of the RBF classifier and the thresholds (Θ_1 and Θ_2) were fixed to their optimum values according to PTB_1 dataset and then the system was trained using each of the considered databases for generalization (PTB or Fantasia). The set of imposter for the FAR test in this experiment encompasses subjects of PTB_1, MIT_BIH and Fantasia (PTB_2), if the PTB_2 (Fantasia) was used for training.The average HR accuracy, FRR and FAR were computed. The results showed that our final reduced set of 66 coefficients has achieved better results, as shown in Table 14.8.

14.7 Conclusion and Future Work

This chapter discussed a new burgeoning branch of biometrics, which is utilizing ECG signals not only as a diagnostic tool but also as a biometric trait. ECG has the advantage of being universal, unique, aliveness indicator and difficult to be falsified or spoofed. However, the intra-subject variability is a central challenge which has been investigated by the developed systems to be overwhelmed. A brief description is given in this chapter about the existing ECG fiducial and non-fiducial based systems along with their advantages and drawbacks.

Based on the survey, fiducial features represent the temporal and amplitude distances between fiducial points along with angle features. Hence, they require the detection of 11 fiducial points from each heartbeat: three peak points (P, R and T), two valleys (Q and S) and the six onsets and offsets for the three heartbeat waves. Thus, the efficacy of the fiducial approach significantly relies on the accuracy of the fiducial detection process, which is a big challenge by itself especially for the onsets and the offsets points, since they are susceptible to error and there is no universally acknowledged rule for defining exactly where the wave boundaries lie. On the other hand, non-fiducial based approaches usually investigate the ECG spectra. Only the R peak is needed for such approaches and for some of them, no detection is needed at all. However, non-fiducial approaches usually result in a high dimension feature space (hundreds of coefficients), which in turn has its limitations. Hence, in order to minimize the intra-subject variability and provide a more compact scheme, a non-fiducial based system that employed biorthogonal wavelet coefficients derived from RR cycles as non-fiducial features along with RBF neural network for classification was proposed. Moreover, a reduction procedure of two phases for the wavelet coefficients was successfully introduced to resolve the problem of high input dimension and improve the performance. The results revealed that 35 % of the wavelet coefficients are enough for the task in hand. The experiments were conducted with RBF classifier using four Physionet databases. Critical issues like stability over time, impostor rejection test and generalization to other databases were addressed by the proposed system.

Finally, crucial aspects such as: cardiac irregularities and variability in heart rate will be considered in our future work. In addition, this work will be explored further by applying this approach to additional datasets (we are compiling our own longitudinal dataset), in order to provide a more thorough evaluation of the stability and scalability of the proposed approach.

References

1. Revett, K.: Behavioral Biometrics: A Remote Access Approach. Wiley, ISBN: 978-0-470-51883-0 (2008)
2. Wang, Y., Agrafioti, F., Hatzinakos, D., Plataniotis, K.: Analysis of human electrocardiogram for hiometric recognition. EURASIP J. Adv. Sig. Process. Article ID 148658 (2008). doi:10.1155/2008/148658
3. Agrafioti, F., Gao J., Hatzinakos, D.: Heart Biometrics: Theory, methods and applications. In: Yang, J. (eds.) Biometrics: Book 3, pp. 199–216. IntechOpen (2011)
4. Mitchel, T.: Machine Learning, 2nd edn. McGraw-Hill, New York (1997)
5. Oeff, M., Koch, H., Bousseljot, R., Kreiseler, D.: The PTB Diagnostic ECG Database (National Metrology Institute of Germany). http://www.physionet.org/physiobank/database/ptbdb/. Accessed 22 Sept 2013
6. The MIT-BIH Normal Sinus Rhythm Database: http://www.physionet.org/physiobank/database/nsrdb/. Accessed 22 Sept 2013
7. The MIT-BIH Long Term Database: http://www.physionet.org/physiobank/database/ltdb/. Accessed 22 Sept 2013

8. The Fantasia Database: http://www.physionet.org/physiobank/database/fantasia/. Accessed 22 Sept 2013
9. Cherry, E., Fenton, F.: Heart Structure, Function and Arrhythmias. Department of Biomedical Sciences, College of Veterinary Medicine, Cornell University, Ithaca, http://thevirtualheart.org/3dpdf/Heart_3d.pdf
10. S̈ornmo, L., Laguna, P.: Bioelectrical Signal Processing in Cardiac and Neurological Applications. Elsevier, Amsterdam (2005)
11. Israel, S.A., Irvine, J.M., Cheng, A., Wiederhold, M.D., Wiederhold, K.: ECG to identify individuals. Pattern Recogn. **38**(1), 133–142 (2005)
12. Karjalainen, J., Viitasalo, M., Mänttäri, M., Manninen, V.: Relation between QT intervals and heart rates from 40 to 120 beats/min in rest electrocardiograms of men and a simple method to adjust QT interval values. J. Am. Coll. Cardiol. **23**(7), 1547–1553 (1994)
13. Agrafioti, F., Hatzinakos, D.: ECG Biometric Analysis in Cardiac Irregularity Conditions, Signal, Image and Video Processing, pp. 1863–1703. Springer (2008)
14. Forsen, G., Nelson, M., Staron, R.: Personal attributes authentication techniques. In: Griffin, A.F.B. (ed.) RADC Report RADC-TR-77-1033 (1977)
15. Biel, L., Petersson, O., Philipson, L.P., Wide, P.: ECG Analysis: a new approach in human identification. IEEE Trans. Instrum. Meas. **50**(3), 808–812 (2001)
16. Kyoso, M., Uchiyama, A.: Development of an ECG identification system. In: Proceedings of the 23rd Annual International Conference of the IEEE Engineering Medicine and Biology, vol. 4, pp. 3721–3723 (2001)
17. Palaniappan, R., Krishnan, S.M.: Identifying individuals using ECG beats. In: Proceedings of the International Conference on Signal Processing and Communications (SPCOM'04), pp. 569–572. Bangalore, India (2004)
18. Singh, Y.N., Gupta, P.: Biometrics method for human identification using electrocardiogram. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009, pp. 1270–1279. LNCS 5558 (2009)
19. Zhang, Z., Wei, D.: A new ECG identification method using bayes' teorem. In: TENCON 2006, IEEE Region 10 Conference, pp. 1–4 (2006)
20. Gahi, Y., Lamrani, A., Zoglat, A., Guennoun, M., Kapralos, B., El-Khatib, K.: Biometric Identification System Based on Electrocardiogram Data, New Technologies, Mobility and Security NTMS'08, pp. 1–5 (2008)
21. Singla, S., Sharma, A.: ECG based biometrics verification system using LabVIEW. Songklanakarin J. Sci. Technol. **32**(3), 241–246 (2010)
22. Tantawi, M., Revett, K., Tolba, M.F., Salem, A.: On the applicability of the physionet electrocardiogram (ECG) repository as a source of test cases for ECG based biometrics. Internat. J. Cogn. Biometrics **1**(1), 66–97 (2012)
23. Tantawi, M., Revett, K., Tolba, M.F., Salem, A.: Fiducial feature reduction analysis for electrocardiogram (ECG) based biometric recognition. internat. J. Intell. Inf. Sys. Springer, **40**(1) 17–39 (2013)
24. Tantawi, M., Revett, K., Tolba, M.F.,Salem, A.: A novel feature set for deployment in ECG based biometrics. In: Proceeding of the 8th (IEEE) Conference on Computer Engineering and Systems (ICCES), pp. 186–191 (2012)
25. Venkatesh, N., Jayaraman, S.: Human electrocardiogram for biometrics using DTW and FLDA. In: 20th International Conference on Pattern Recognition (ICPR), pp. 3838–3841 (2010)
26. Wübbeler, G., Stavridis, M., Kreiseler, D., Bousseljot, R., Elster, C.: Verification of humans using the electrocardiogram. Pattern Recogn. Lett. **28**(10), 1172–1175 (2007)
27. Mai, V., Khalil, I., Meli, C.: ECG biometric using multilayer perceptron and radial basis function neural networks. In: 33rd Annual International Conference of the IEEE EMBS, pp. 2745–2748. Boston, USA (2011)
28. Chan, A., Hamdy, M., Badre, A., Badee, V.: Person identification using electrocardiograms. IEEE Trans. Instrum. Measur. **57**(2), 248–253 (2008)
29. Wan, Y., Yao, J.: A neural network to identify human subjects with electrocardiogram signals. in: Proceedings of the World Congress on Engineering and Computer Science. San Francisco, USA (2008). doi:10.1.1.148.5220

30. Chiu, C., Chuang, C., Hsu, C.: A novel personal identity verification approach using a discrete wavelet transform of the ECG signal. In: Proceedings of the 2008 International Conference on Multimedia and Ubiquitous Engineering (MUE'08), pp. 201–206. Washington, DC, USA (2008)
31. Fatemian, S., Hatzinakos, D.: A new ECG feature extractor for biometric recognition. In: Proceedings of the 16th international conference on Digital Signal Processing, pp. 323–328. IEEE Press Piscataway, NJ, USA (2009)
32. Belgacem, N., Ali, A., Fournier, R., Bereksi-Reguig, F.: ECG based human authentication using wavelets and random forests. Internat. J. Crypt. Inf. Secur. **2**(2), 1–11 (2012)
33. Saechia, S., Koseeyaporn, J., Wardkein, P.: Human identification system based ECG signal. TENCON **2005**, 1–4 (2005)
34. Molina, G., Bruekers, F., Presura, C., Damstra, M., van der Veen, M.: Morphological synthesis of ECG signals for person authentication. In: Proceedings of the 15th European Signal Procesing Conference, pp. 738–742 (2007)
35. Sufi, F., Khalil, I., Habib, I.: Polynomial distance measurement for ECG based biometric authentication. Secur. Commun. Netw. (Wiley Interscience) (2008). doi:10.1002/sec.76
36. Ting, C., Salleh, S.: ECG based personal identification using extended kalman filter. In: 10th International Conference on Information Sciences Signal Processing and their Applications, pp. 774–777 (2010)
37. Coutinho, D., Fred, A., Figueiredo, M.: One-lead ECG-based personal identification using Ziv-Merhav cross parsing. In: 20th International Conference on Pattern Recognition, pp. 3858–3861 (2010)
38. Ghofrani, N., Bostani, R.: Reliable features for an ECG-based biometric system. In: 17th Iranian Conference of Biomedical Engineering, pp. 1–5 (2010)
39. Li, M., Narayanan, S.: Robust ECG biometrics by fusing temporal and cepstral information. In: 20th International Conference on Pattern Recognition, pp. 1326–1329 (2010)
40. Plataniotis, K., Hatzinakos, D., Lee, J.K.M.; ECG biometric recognition without fiducial detection. In: Proceedings of Biometrics Symposiums (BSYM'06), Baltimore, Md, USA (2006)
41. Agrafioti, F., Hatzinakos, D.: ECG based recognition using second order statistics. In: Sixth Annual Conference on Communication Networks and Services Research (CNSR), Halifax, Canada, 5–8 May 2008
42. Wao, J., Wan, Y.: Improving computing efficiency of a wavelet method using ECG as a biometric modality. Int. J. Comput. Netw. Secur. **2**(1), 15–20 (2010)
43. Odinaka, I., Lai, P., Kaplan, A., O'Sullivan, J., Sirevaag, E., Kristjansson, S., Sheffield, A., Rohrbaugh, J.: Ecg biometrics: A robust short-time frequency analysis. IEEE International Workshop on Information Forensics and Security, pp. 1–6 (2010)
44. Tawfik, M., Selim, H., Kamal, T.: Human identification using time normalized QT signal and the QRS complex of the ECG. In: Proceedings of the 7th International Symposium on Communication Systems Networks and Digital Signal Processing, CSNDSP (2010)
45. Shen, T.W., Tompkins, W.J., Hu, Y.H.: One-lead ECG for identity verification. In: Proceedings of the 2nd Joint EMBS/BMES Conference, pp. 62–63 (2002)
46. Shen, T.W.: Biometric identity verification based on electrocardiogram (ECG). Ph.D. thesis, University of Wisconsin, Madison (2005)
47. Haykin, S.: Neural Networks: A Comprehansive Foundation, 2nd edn. Prentice Hall, Upper Saddle River (1999)
48. Pan, J., Tompkins, W.: A real time QRS detection algorithm. IEEE Trans. Biomed. Eng. **33**(3), 230–236 (1985)
49. Misiti, M., Misiti, Y., Oppenheim, G., Poggi, J.: Wavelet Toolbox 4 User Guide, 4.1 edn. The MathWorks Inc. (2007)
50. Chen, S., Chng, E.: Regularized orthogonal least squares algorithm for constructing radial basis function networks. Internat. J. Control **64**(5), 829–837 (1996)

Chapter 15
Image Pre-processing Techniques for Enhancing the Performance of Real-Time Face Recognition System Using PCA

Behzad Nazarbakhsh and Azizah Abd Manaf

Abstract In the last decade face recognition has made significant advances, but it can still be improved by applying various techniques. The areas that have high promise of improvement are those that utilize preprocessing techniques. The main objective of this study is to improve the auto face recognition system performance using off-the-shelf image library. Face detection technique plays a significant role in recognition process. The process chains used to detect human face are those that comprises of color segmentation, localization using Haar-like cascade algorithm and geometry normalization. Subsequently, one half portion of the facial image was selected to be used as the calculated average half-face image. The high-dimensionality of the image value is further reduced by generating Eeigenfaces. This is followed by the classification process that was achieved by calculating the Eigen distances values and comparing values of image in the database with the captured one. Finally, the verification tests are carried out on images obtained from VidTIMIT database to evaluate the recognition performance of the proposed framework. The resultant tests from the data set yielded the following results: true acceptance rate at 91.30 % and false acceptance rate at 33.33 %. The obtained experimental results illustrates the proposed image preprocessing framework improves the recognition accuracy as compared to not applying it.

B. Nazarbakhsh (✉) · A. A. Manaf
Advanced Informatics School (UTM AIS), Universiti Teknologi Malaysia,
Wilayah Persekutuan, 54100 Kuala Lumpur, Malaysia
e-mail: nbehzad2@live.utm.my

A. A. Manaf
e-mail: Azizaham.kl@utm.my

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_15, © Springer-Verlag Berlin Heidelberg 2014

15.1 Introduction

The idea of using physical attributes for proving human identity recently is demand of many systems. Face is one of the human attributes that clearly distinguishes different individuals. In fact, face is the attribute that is most commonly used by human visual system to identify people. Philosophically, "identity" is whatever makes an entity definable and recognizable, in terms of possessing a set of qualities or characteristics that distinguish it from entities of a different type. Hence, "Identification", is the act of establishing that identity.

With the advent of the Artificial Intelligence knowledge area and Automatic Identification and data capture (AIDC) technologies (such as RFID, Smart Cards, Barcode System, Biometrics and Optical Character Recognition); data are captured or collected by using automated mechanism without the need of manual input. Biometrics is common technology which will be used for human face identification [1].

The Automatic face recognition technology (AFRT) is a relatively new concept. The face detection, face recognition, facial expression recognition, facial gender determination, facial age estimation and facial ethnicity estimation are the details that can be extracted from the face by using AFRT. Face recognition is based on having an unknown face image and matching it against a database of known images. It has emerged as an attractive solution to address many contemporary needs for identification and the verification of identity claims. It brings together the promise of other biometric systems, which attempt to tie identity to individually distinctive features of the body, and the more familiar functionality of visual surveillance systems.

The AFRT was developed in the 1960s; the first semi-automated system for face recognition required the administrator for localizing the facial features such as eyes, ears and mouth on photographs before it calculated distance and ratios to a common reference point, which were then compared to reference data [2]. In the 1970s, the manual computation was the problem of feature measurements and locations. In 1988, Kriby and Sirovich applied principle component analysis, a standard linear algebra methods, to the face recognition problem. This was considered somewhat of a milestone as it shows that less than one hundred values were required to accurately code a suitably aligned and normalized face image [3].

In 1991, Turk and Pentland [4] revealed that while using the Eigenfaces method, the residual error could be used to detect faces in images, a discovery that enabled reliable real-time automated face recognition systems. Although the tactic was slightly constrained by environmental factors, it nonetheless created significant interest in furthering development of automated face recognition machineries. The technology first captured the public attention from the media reaction to trial implementation at the January 2001 Super Bowl, which captured surveillance images and compared them to a database of a mug shot [5]. Different approaches have been tried by several groups, worldwide, during the decades to solve the current problems in face recognition, but so far no system or technique exists which have shown the satisfactory results in all the circumstances.

In general, the AFRT compares the rigid features of the face which does not change over the period of time. Its inputs can be visible light, infrared, and an image or video stream from stereo or other range-finding technologies. The output is an identification or verification of the subject that appears. Among those variant types of entry data, imaging in the visible light spectrum is considered as the best entry data for face recognition due to the gigantic quantity of legacy data and the ubiquity and cheapness of photographic capture equipment.

Basically, the current facial recognition systems can be classified into verification, identification and watch list. Verification is a one-to-one match that compares a query face image against a template face image whose identity is being claimed [6]. Identification compares the given individual to all the other individuals in the database and gives a ranked list of matches. Indeed, it is a one-to-many matching process that compares a query face image against all the template images in a face database to determine the identity of the query face [7]. Lastly, the watch list is presented as an open-universe test by the Face Recognition Vendor Test which was conducted by the National Institute of Standards and Technology. In this system, the test individual may or may not be in the system database. Thus, the person is compared to the others in the system's database and a similarity score is reported for each comparison [8].

Although, diverse algorithms are available to perform the comparisons of facial features in the face recognition systems, the basic processes should be sequentially well-defined for achieving any type of facial recognition goals. The entire face recognition process has been applied by cutting the entire process to variant basic processes in order to overcome the difficulties in face recognition which are very real-time and natural.

The detecting and recognizing faces always are the two main problems in any automatic face recognition. Because any face images can have head pose, illumination or occlusion, those cause face detection and recognition have not done perfectly. Indeed those problems are quite critical for recognition system which cause of reducing the system accuracy and performance.

In order to provide the solution for those problems, it was required that image to be processed before recognition. Consequently, the detection of faces had been done by separated multiple computations processes. Indeed, these processes are divided into small ones for attenuating the huge problem by eliminating the cluttered background, localization of the face, extraction of facial features in given image and face verification as the certain steps that must be consecutively followed [9]. The National Polytechnic Institute Mexico researchers have pursued the above perception by defining the different recognition terminology [10], as shown in Fig. 15.1. They defined the pre-processing step including some kind of cutting, filtering, or some methods of image processing techniques among others were applied. These methods helped them to acquire a portion of facial image before beginning to verify the face by eliminating unnecessary information from the image. Besides, the classifier defines a set of training as a set of elements, each being formed by a sequence of data for a specific object.

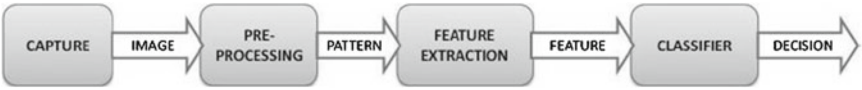


Fig. 15.1 General structure of a face recognition system

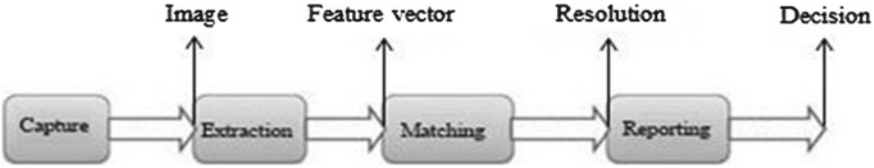


Fig. 15.2 Biometric recognition flow diagram

The Automation Chinese Academy of Sciences survey illustrates searching to grab needed information from those features must be the first major process for face detection [11]. Subsequently, the alignment process is defined for determining a head-like shape, head's position, size and pose. Later, normalization process was the next step that was followed in order to overcome the illumination and pose issue in face recognition [12, 13]. In this step, the image of the head is scaled and rotated so that it can be registered and mapped into an appropriate size and pose. Normalization is performed regardless of the head's location and distance from the camera and environment illumination does not impact on the normalization process. In the representation process, the facial data were translated into a unique code, for easily evaluating, the acquired facial data in comparing to stored facial data.

One of the major challenges in face perception lies in determining how different facial attributes contribute to judgments of identity. The matching process are defined, for acquiring facial data and comparing it with the stored data in order to ideally linked to at least one stored facial representation. Even if the reporting process is defined after the matching process for aiding to return more than facial matches based on the score and user preferences [10] (Fig. 15.2).

This project is to develop in-door facial recognition system. The facial images with very small changes in head pose and tilting will be analyzed. The appropriate practices for regionalizing facial image area are applied and dimensionality of an image will be reduced for enhancing the recognition accuracy. The camera is positioned at the same distance from the face and the samples will be in the similar lighting condition, without any physical obstruction.

15.2 Literature Survey

15.2.1 Image Acquisition

Image acquisition is an essential part of image processing consists of two stages, image capturing and preprocessing. The result of capturing image will be a two-dimensional, ordered matrix of integer or floating-point values. The pre-processing the image properties impacts on the image size, resolution, coordinate system, pixel values, format and types which after or before any preprocessing do impact on face recognition process. Usually the performance of recognition system is depended to image preprocessing.

The resolution of an image specifies the spatial dimensions of the image in the real world and is given as the number of image elements per measurement. Although, the spatial resolution of an image will not relevant in many basic image processing steps but precise resolution information is important in cases where geometrical elements need to be drawn on an image or when distances within an image need to be measured. A coordinate system is needed, due to understanding which position on the image corresponds to which image element.

The exact bit-level layout of an individual pixel depends on the image's type. Five types of image have been introduced including gray scale images, binary images, color images, indexed images and special images which have different bit-level layout of an individual pixel [14]. Most of recent image processing methods focus on color dots. The color depth and color spaces are two main specifications for color image besides what are mentioned above.

Allocating a certain number of required bits in an image represents color depth. Color spaces in the simplest terms mean image visualization based on the color receptors which are red, green and blue colors. For representing a color, there are variant color spaces, or models such as RGB, CMYK, HSI, YUV, and LAB [15]. Some literatures asserted, most of software application related to machine vision use HSI in identifying the color of different objects and creating better interaction with humans [16], due to usual inclination to distinguish altered colors (hue) at different shades (saturation and intensity). Moreover, some image processing techniques such as histogram operations, intensity transformations and convolutions are performed with much ease on an image in the HSI color space since their operation only effects on image intensity.

15.2.2 Segmentation

The basic required phase for face detection is facial face segmentation. Numerous techniques relates to these phases will be used to cluster the pixels into visible image regions. The principal advantage of segmentation and corresponds to the reduction of the dimensionality and then, the computing time. Color Information, Boundary, Eigen Basis, Neural Network, Genetic Algorithm, and Voronoi Diagram are common segmentation approaches recently proposed [17]. The color based technique is used to track objects defined by a set of colored pixels whose chrominance values satisfy a range of thresholds. The effectiveness of this method will be evaluated in terms of quality of segmentation results and the behavior of transformation from input. Usually, in the color information model, the facial image is going to be segmented through the brightness of the skin color or skin textures.

In 1994, color space segmentation techniques classified into three sets: histogram based techniques, segmentation by clustering data in color space and segmentation by fuzzy clustering [18]. Lately, in 2012, face segmentation is defined based on HSI color information conversion, determining the classes of colors and treating each chromatic color [19]. During decades, even though color information method has been used for image segmentation, but it gives a large amount of false results. Moreover, it is very difficult to segment face in real time from the color images with cluttered background, camera parameter changes, or light condition [18–20].

Beside color skin thresholding technique, the skin color pixel classification method is the other well-known practice that can be used for segmenting an image based on color pixels. This method is categorized into several algorithms as follow: The piecewise linear classifiers, Bayesian classifier with the histogram technique and Gaussian classifiers and the multilayer perceptron. The major drawbacks of these algorithms are due to sensitivity to camera parameter changes; being sensitive to light condition, enhancing detection complexity and finally having destructive influence on detection accuracy [21].

Boundary segmentation method is another method that can be used for image segmentations. It refers to the process of identifying and locating sharp discontinuities in an image. The majority of diverse methods for applying boundary segmentation are grouped into Gradient based Edge Detection and Laplacian based Edge Detection [22]. After studying numerous reviews, it is understood that Sobel operator, Robert's cross operator, D. Prewitt Operator, Laplacian of Gaussian and Canny edge detection approaches are the renowned practices. Those algorithms are compared based on noise sensitivity, complexity, detection accuracy and missing orientation process (Table 15.1).

In the Eigen-basis models, the image will be compared with the predefined face template. This method has tried to use the set of Eigenvectors as the set of features for characterizing the variation between the face images. Each image location contributes more or less to each eigenvector so the eigenvector can be displayed as the sort of ghostly face which is called Eigen Face [4]. Some researchers have used the same concept and applied it on facial features. Indeed, they used Eigen Feature that concentrates on the restricted face area to obtain the main components of features point of the face. The main issue of this approach is that, it cannot effectively deal with scale, pose, and shape change of faces/facial features.

Neural Network is the solution to recognize the patterns, for training the network input data comprise three forms, face data, non-face data and face mask which creates a mask that cuts off surrounding edges of the image rectangle to give an oval shape to the face image. Although, advantages of this approach is referred back to its

Table 15.1 Comparison of edge segmentation algorithms (a) H = High (b) M = Medium (c) L = Low

Edge segmentation methods	Missing orientation process	Noise sensitivity	Complexity	Detection accuracy
Sobel, E. Kiresh and D. Prewitt	L	H	L	L
Canny Shen-Casten	H	L	H	H
Laplacian of Gaussian	M	L	M	M

Table 15.2 Comparison between skin color and Eigen-basis method

Criterion	Skin color based approach	Eigen-basis approach
Handling more than one face	It detects each face	It fails
Rotation, profile and titled face	It is dependent on the skin color not on the orientation	In case if the Eigenspace contains such information then works
Size of the face	Not necessarily, since neck can be included	It depends on the size of the Eigenbasis
Complex background	If it has color similar to the skin color than it suffers	Usually it handles this situation

large difficult dataset but the main drawback of this segmentation approach is the huge search space at run time. Moreoever, it is very computationally demanding and therefore are not very interesting in applications where real-time or near real-time performance is required [23].

The Genetic Algorithm was conceived for object recognition in machine vision but it can be also applied to search for possible facial region and features in an image. The main idea behind this method is selecting the worst design solutions by estimating how close it came to meeting the overall specification and to breed new ones from the best design solutions. Even though genetic algorithm is precise but it suffers from being a time consuming technique [24].

Finally, Voronoi diagrams divides a space into disjoint polygons and triangles from an image and then segment it based on these features. Kosagi proposed mosaic model which follows the same theory to detect the head region by matching the input mosaic image and six clustered mosaic face patterns [25]. The facial feature detection correct rate of this approach is high but it is too sensitive to noise. Moreover, it has limitations in cases of rotation [26].

Variant surveys illustrate, Eigen-basis and skin color segmentation are the most used in current automatic face recognition techniques. After comparison of those methods as shown in Table 15.2, color segmentation is implied as the suitable approach, due to having less complexity.

15.2.3 Localization

Face localization is the process used for identifying and tracking the main facial features and also analyzing facial expressions by using feature detectors. It aids to solve the pose problem and the multi-view face detection issue. Furthermore, it improves the detection process by uniforming the background, fixing the scales, poses and etc.

The feature detectors are divided into three categories: contour based, intensity based and parametric model based. A contour based technique is based on hydrocodes practices by parameterizing the contour using algorithms such as Lagranglan and Eulerian. An intensity based feature detector extracts interest points in an image to identify the interested image section. Finally, the parametric modeling denotes the use of parameters to control the dimensions and shapes [27]. The parametric geometry is one way of applying parametric model base. Linear or angular dimensions, datum line, points, surface or coordinate system, geometric constraints are the primitive elements of parametric geometry. Variational geometry solvers are nonlinear solution for applying parametric modeling, it is claimed that they are more general and powerful than the parametric geometry [28].

The facial landmarks, joint optimization and Viola-Jones face detector, SVM classifier and Haar_Like Cascade are common methods for applying face localization. A facial landmark is to invent heuristics that are experimentally validated on a particular dataset. For instance, the closest facial feature point to the camera can be selected as the tip of the nose in 3D facial feature localization. The joint optimization is kind of interactions between landmark locations and local feature constraints which are considered as distances to feature templates. Viola-Jones method patches around facial landmarks are detected in the face area with a boosted cascade of simple classifiers. This approach is usually used for the 2D coarse-scale detection. SVM classifiers will constraint the face area searching by choosing the specific facial feature location such as the nose, eyes and etc. This method is one of the most expensive solutions which are proposed by now [29]. Haar-like features is the core foundation for Haar classifier object detection. Instead of using the intensity values of pixel, these features use the change in contrast values between adjacent rectangular groups of pixels. In this technique contrast variances between the pixel groups are used to determine relative light and dark areas. Haar-like feature is made up of two or three adjacent groups with a relative contrast variance. By increasing or decreasing the size of the pixel group being examined, Haar features can easily be scaled. This allows features to be used to detect objects of various sizes [30]. The Fig. 15.3 illustrates the use of Haar-like features in order to detect the facial features.

15.2.4 Normalization

Imaging condition, pose variation, presence or absence of face structural component are the major challenges of face recognition process [31]. Normalization aids the

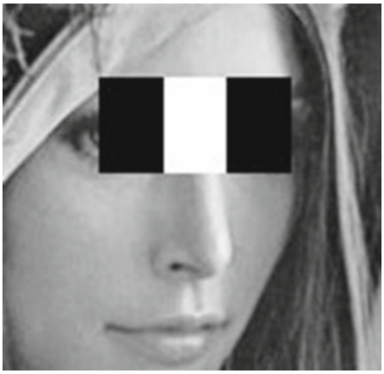


Fig. 15.3 Use of Haar-like features in order to detect the facial features

recognition process by standardizing detected image in terms of size, pose, illumination, relative to the images in the database. Indeed, those factors can have the bad or good effect on performance of face recognition.

Normalization types were categorized into geometric and illumination [32]. The geometric normalization involves bringing the faces to a standard size and rotating it in-plane. The illumination normalization controls the image capturing environment and imposes strict requirements regarding lighting conditions. Indeed, illumination normalization is a good compromise between execution speed and robustness to lighting variations. Histogram Equalization (HE), Contrast Limited Adaptive Histogram Equalization (CLAHE), Logarithm Transformed combined with suppressing DCT coefficients (LogDCT) or retinex are the common algorithms which will be used for minimizing the effect of lighting.

15.2.5 Feature Selection and Extraction

The main aim of feature extraction is feature selection in order to reduce the number of features provided to the classification task. Technically, face detection and extraction are combined with each other because most of feature extraction algorithms and techniques will be based on facial feature detection. Indeed, the feature extraction is the task of generating the set of elementary features; convert it into the data and passing it directly to the classification task. Usually, the annoying facial features for recognition process will be eliminated or higher-order features will be determined in extraction process. Generating accurate approximations, representing the extracted structures compactly, supporting subsequent classification, and being domain independent are characteristics of suitable feature extractor [33]. The eyes and mouth are marked as the suitable features, because finding them are faster, easier and more precise than faces not only during identification but also for analyzing the extracted feature [34]. As the result eyes are the most important facial features and the hair is considered as an annoying factor. So, it must be filtered before sending image vectors to the classifier.

It is well-known that the face is symmetrical, for this reason numerous studies utilize the half face to assist in computing a similarity measure between faces and feature extraction. It is claimed using the average half face on two dimensional face images requires two steps, including: (1) the facial image must be centered properly oriented face to represent the data as symmetric as possible, (2) the facial image must be divided into two symmetric halves and to be averaged together by reversing the columns of one of the halves first [35]. The feature selection as well as subspace computation can be performed on the set of average-half-faces just as is done on a set of full faces and any recognition algorithm can be applied on it.

15.2.6 Classification

After minimization of the feature extraction effort by performing the feature extraction, the selected features will be relegated to the classification. Indeed, the goal of a classifier is helping verification by defining a set of training as a set of elements, each being formed by a sequence of data for a specific object. The classifier measures an approximation of the live model with the reference model of the known person. If the live model exceeds a threshold verifying is successful. If not, the verification is unsuccessful. Classification comprises training and testing stages. In training, prototypes which will be built from single or multiple face samples are constructed for each person in the database. The prototype will be represented by a series of feature vectors that rearranged the intensities of their pixels. In testing stage, the test samples are compared with each person prototype and similarity scores are computed.

Applying the optimized algorithms to refine the results and produce a better and quick outcome has been a goal of proposing variant algorithms during decades. The classifier algorithms define a model for each object specific, so that the class to which belongs an element can be calculated from the data values that define the object. Various algorithms combinations can be used for classification task. The computational complexity of most classifiers algorithms are dependent on the dimensionality of the input features, therefore if all the pixel values of the face image are used as features for classification the time required to finish the task will be excessively large. This prohibits direct usage of pixel values as features for face recognition. To overcome this problem, different dimensionality reduction techniques have been proposed. Those classification algorithms are categorized based on three different approaches [36]:

(1) Holistic based (2) Geometry based (3) Hybrid.

The holistic based algorithms classify the image by applying a pattern classifier in training stage. A template can be a pixel image or a feature vector obtained after processing the face image as a whole [7]. Geometry based computes the distances between feature points and the relative sizes of the major face components to form a feature vector [9] and hybrid algorithms apply both the local and holistic features model. The variant survey shows holistic algorithms are reducing a higher dimensional training data set of face images to a lower dimensional one, while preserving the key information contained in the data set. They do not destroy the data in the images due to concentrating on limited regions or points of interest which is the advantages of this approach in comparing with geometry based method but they are sensitive to variation in pose, because global features are highly receptive to translation and rotation changes of the face. In contrary, the geometry based approach is, to compensate for pose changes by allowing a flexible geometrical relation between the face components in the classification stage [37].

Linear subspace, none linear subspaces and transformation based algorithms are well known groups of algorithms belong to holistic category. Discrete Cosine Transform (DCT) from transformation base, Eigenfaces from linear subspace are two renowned methods which were introduced as the well-known holistic based approach for feature extraction [38]. Linear subspaces algorithms are viewed as finding a set of vectors which effectively represent the information content of an observation while reducing the dimensionality.

Principle component analysis (PCA based Eigenface) is at the easiest and simplest Linear subspaces algorithm. Correlation between components of the data vector will be clearly viewed and the second order correlation value will be captured by PCA. This Algorithm identifies the linear combinations of variables and ignores the high order Correlation value. The 2D-PCA algorithm is based on 2D image matrices instead of 1D vector. The image covariance matrix is constructed directly from the original image. In the recent proposed, 2D Linear Discriminant Analysis, the image is not reordered as a column vector [39].

Independent Component Analysis (ICA) is has been utilized to find a subspace of the data. It captures the high order statistics of the data but if the data sources are independent then it works well [40]. Linear discriminant analysis (LDA) algorithm developed to find the subspace that best discriminates different face classes by maximizing between class scatter, while minimizing the within-class scatter [41]. Indeed, linear mapping, dimensionality of the subspace is limited by the number of classes of the data but it cannot handle data in which the individual classes are far from Gaussian, moreover it suffers from small sample size problem.

Another linear algorithm is Locality preserving projections (LPP). This algorithm is a general method for manifold learning. It is obtained by finding the optimal linear approximations to the eigenfunctions of the Laplace Betrami operator on the manifold [42]. It shares many of the non-linear properties for data representation. In this case, Gaussian weights are used to reduce the space but the occurrence of parameter sensitive is a disadvantage of this algorithm. Singular value decomposition (SVD) is the robust method of storing large images as smaller but its computation is hard and does not work for subsequence indexing. The main difference between PCA, LDA, and LPP is that PCA and LDA focus on the global structure of the Euclidean space, while LPP focuses on the local structure of the manifold [43].

The algorithms combinations can be used for classification. Image classification algorithms at test stage are divided into parametric and Non-parametric classifiers approaches. Parametric algorithms need an intensive training phase of the classifier parameters for instance parameters of SVM, Boosting, parametric generative models, decision trees, fragments, object parts and etc. Non-parametric classifiers are base their classification decision directly on the data, and require no learning of parameters [44].

The Nearest Neighbor (NN), Bayes' decision; Neural Networks, Gaussian MixtureModel and Support Vector Machine are the well-known existent classifier algorithms using in testing stage [45]. The k-nearest-neighbors decision rule is common method to be used which classifies an object based on the class of the k data points nearest to the estimation point. Nearness is commonly measured using the Euclidean distance metric space. Bayesian algorithm is based on computing the tradeoffs between several classification verdicts using likelihood that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. The artificial neural networks algorithms such as Multilayer perception (MLP), Dynamic link architecture (DLA), Probabilistic decision based neural network (PDBNN) and etc. attempt at modeling the information processing capabilities of nervous systems. The adjustment of the parameters will be done through a learning algorithm, of automatic adaptive method.

A Gaussian mixture Model (GMMs) is commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum a Posteriori (MAP) estimation from a well-trained prior model. Besides, classification can be done by matching the marginal density distributions. EM + MML algorithm is used to perform parameter estimation and model selection automatically. Besides, the Earth Mover's Distance (EMD) is used to measure the distributional similarity based on the Gaussian components [46]. The Support Vector Machine (SVM) is based on the statistical learning theory. It provides a new form of parameterization of functions. Indeed, it is robust method of storing large images as smaller but its computation is hard and does not work for subsequence indexing [47]. It has become popular to perform video classification using various image features, such as Histogram of Oriented Gradients (HOG) or Scale-invariant feature transform (SIFT). The simple classification algorithms by calculating the distance between feature vectors are suggested as the simple and fast classification method in this recognition phase.

15.3 Methodology

The face recognition framework is proposed as the basis of the recognition system. The entire face recognition practice is divided into different processes. Image acquisition, detection, recognition and matching consider as the main steps in the proposed face recognition framework which will constantly occur in any recognition system regardless of the type and modality of the entire system. Besides, some pre-processes are defined. The input and output of each preprocess in our terminology is tabulated in Table 15.3 and the sequence of each preprocess is depicted on Fig. 15.4.

Table 15.3 Pre-processes in automated face recognition

No.	Pre-processes		
	Appellation	Input	Output
1	Localization	Captured image	Pattern
2	Normalization	Face/Face features image	Normalized face images
3	Segmentation	Captured image or feature face image	Pattern
4	Features extraction	Processed image	Face features image
5	Classification	Processed image	Authentication

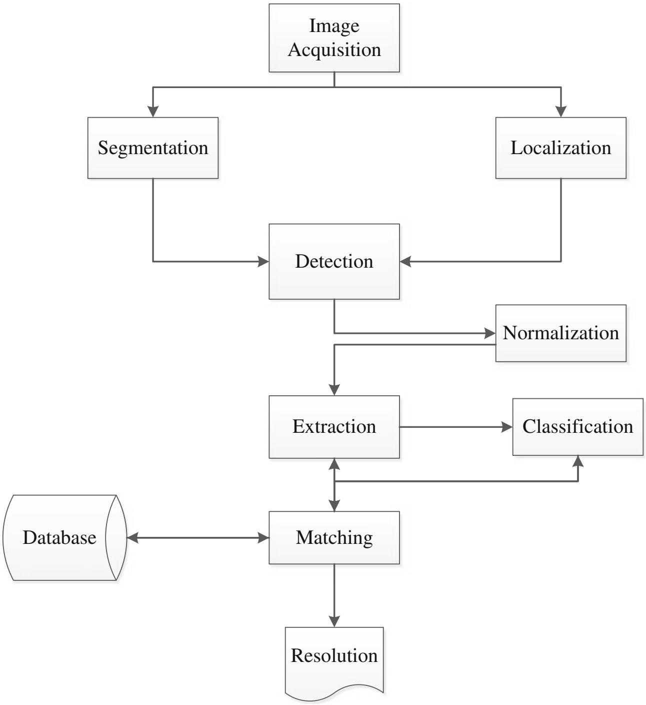


Fig. 15.4 Face recognition framework

15.3.1 Image Preprocessing Using for Detection

In this survey, it is decided that a multi-resolution image preprocessing technique to be used for preprocessing the image input and gathering more information about that. The multi-resolution technique makes possible to adapt the preciseness of the detection result to the need of the following recognition process. For this reason, segmentation of the image and localization the facial features are considered as the two major required processes for face identification. In the general case those have the advantage to be sparse in terms of computational resources.

15.3.2 Color Conversion and Segmentation

In the segmentation method, at the beginning, the facial image pixels will be retrieved and the background pixels will be ignored. The edge color thresholding based on the skin color tones has been selected as a key element for separating background information from facial in the image. Consequently, in the first step after capturing the input color image with the standard webcam image, the original image color space is conversed from RGB to HSI color space. Basically after accomplishing this process, the region of interest which is head model location will be estimated. On the other hand, the faces will be detected.

15.3.3 Localization

As localizing the two eyes and mouth region is plenty and beneficial for detecting one face in an image, not only the coordination of local appearance features of the eyes and mouth will be computed but also their correlation will be computed for enhancing the face identification. Furthermore, the localization will be done for gathering more information about the facial features coordination's and correlation of them in facial image before face authentication. Haar like Cascade decomposition technique will be employed for enhancing the accuracy of robust automatic face identification via extracting feature vectors from the basic parts of a face which are two eyes and mouth.

15.3.4 Normalization

Once the identifying the facial image's ROI and localizing points of interest on the face, the geometry normalization will be applied on image in order to deal with the pose change issue under the condition of one sample per person. The detected face image will be aligned, rotated on horizontal line based on two eye centers and mouth. Moreover it will be rescaled, then a cropping step will be applied within a 128*128 cropping window, keeping the face image and abandoning ear, hair and jaw line external features. Cropping an image will reduce the size of the data and iterations required to scan the full image by dynamic time wrapping.

15.3.5 Face Recognition

In this study, the appearance (photometric cues) of facial component information is decided to be used for human facial authentication. For authenticating the detected faces, first of all the feature extraction will be done and subsequently the classification will be pursued.

15.3.6 Feature Extraction

The holistic approach is decided to be applied for the proposed automatic face recognition framework. The main idea of recognizing the face images is decomposing images into the smallest set of characteristics feature images, which may be thought of as the principal components of the original images. Indeed, warping the face benefits the holistic methods because most of holistic methods stack the face pixels into one dimensional vector. After warping, face information can be incorporated into authentication. In this study, the image will be divided into two equal halves that carry important discriminating information and centered independently at the component centers. On the other hand, the image information will be partitioned and afterward recognition process compares the corresponding facial components. Applying the image pre-processing techniques which already were employed aid the feature vectors to be extracted from the face more precisely. Indeed, the feature extraction attempts to benefit the final classification by averaging the wrapped feature. Consequently, the wrapped half face will be averaged and each average half face will be projected onto its corresponding Eigenspace for recognition.

15.3.7 Classification

The key to classification will be the similarity or distance function between selected features. In this project, PCA will be used as the feature extraction algorithm that guarantees minimal loss of information and also selecting the features with most discriminative power. Then, final component classifier will be applied on wrapped facial feature and the final decision will be made by majority voting or linearly weighted summing. In view of the difficulty of estimating the parameters of sophisticated classifiers, the simple Euclidian distance are decided to be used (Fig. 15.5).

15.4 Proposed Method

15.4.1 Color Conversion

The face detection made by a decision function needs to be filtered for removing the non-facial image parts with the help of HSI color spaces. cvCvtColor function in the OpenCV library aid to convert the image from the RGB color space to HSI color

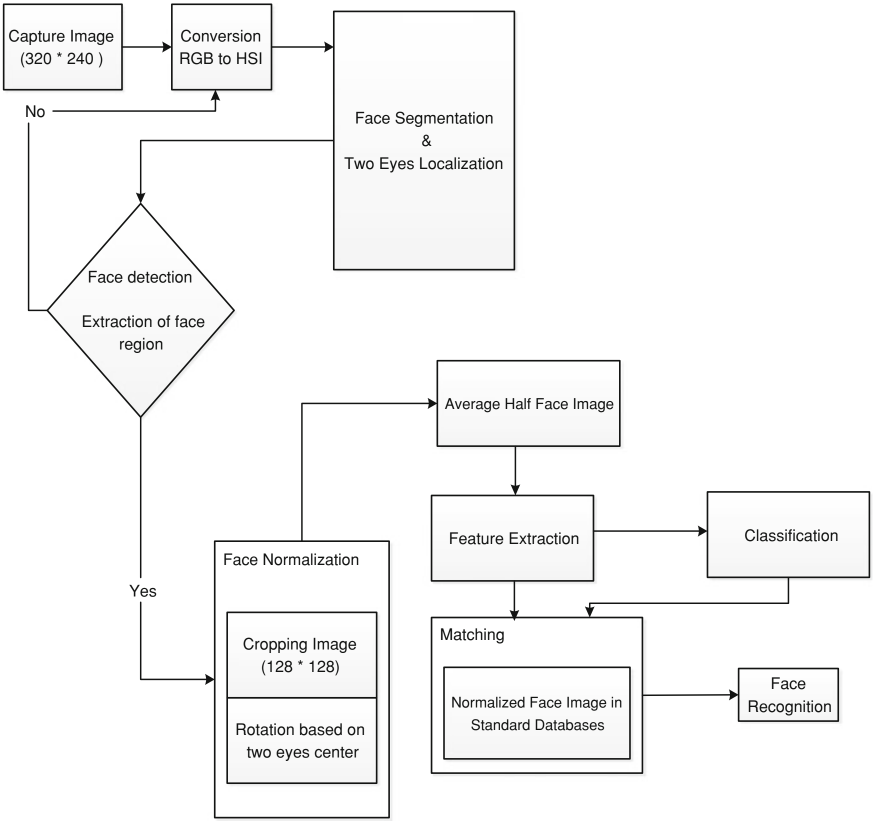


Fig. 15.5 Face recognition detailed block diagram

space. In OpenCV, value range for ‘hue’, ‘saturation’ and ‘value’ are respectively [0–255], [0–255] and [0–255]. After color conversion to HSI, The displayed output will be based on three color channels that represents the color, the amount to which that respective color is mixed with white and the amount to which that respective color is mixed with black.

$$\text{HSV Image} = \text{cvtColor(matOrginal,HSV,COLOR\_BGR2HSV)}$$
$$\text{Where (COLOR\_BGR2HSV} = 40) \tag{15.1}$$

As shown in Fig. 15.6, the conversion of RGB to HSI color space will aid to separate color components from intensity for enhancing robustness to lighting changes and removing shadows. As the result, only the intensity component will be altered and the color components left alone. Moreover, This HSI filter fairly represents the skin tones.

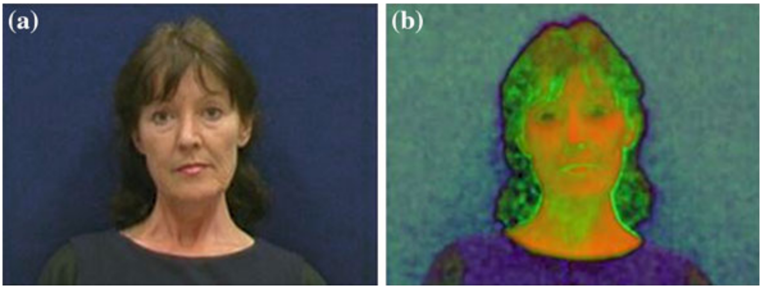


Fig. 15.6 Color space conversion after using HSI filter. **a** Original color image with RGB color space. **b** Color image after conversion to HSI color space

15.4.2 Segmentation

After HSI image is filtered based on threshold of the color spaces values, constraining the image pixels for acquiring the facial image is done during the segmentation process by subtracting the slice of image pixels from the entire pixels which are not required to be analyzed. Image color segmentation assists to partition an image to non-overlapping areas. During the segmentation process, the color intensity will be analyzed instead of chrominance and each segmented area is defined as the homogeneous connected pixels. Three phases are accomplished for achieving the output of the segmentation process including: (1) Acquiring the binarized masked image, (2) Morphological operation, (3) Acquiring the color masked image.

For doing color segmentation by image thresholding and using the Masks on the image, it is required that color image to be binarized by converting to a binary image where the Region of interest (ROI) is white (pixels with value 1) and all others are blacked out (pixels with value 0). The color algorithm filters color of interest base on a minimum and maximum threshold of color spaces. The concept of this algorithm is selecting pixel values which are greater than the fixed range of predefined threshold and reject the rest of them. On the other hand, for creating the mask image by the black color, each image pixel *P* will be segmented as ROI when P(x, y) is between Max Threshold and Min Threshold. Indeed, the decision function forms the skin map by checking the conditions for the existed skin area based on the following threshold specifications:

$$\begin{cases} 0 < \text{Hue} < 20 \\ 48 < \text{Saturation} < 255 \\ 80 < \text{Intensity} < 255 \end{cases} \tag{15.2}$$

InRange function in OpenCV checks that array elements lie between scalars with inclusive lower boundary and exclusive upper boundary and goes through the array elements to check whether elements are in the specified given range or not. If so, it is selected and feed result to last parameter as output image otherwise rejected. It is required that lowest and highest color of range to be given to cv::Scalar function for pixels extraction.

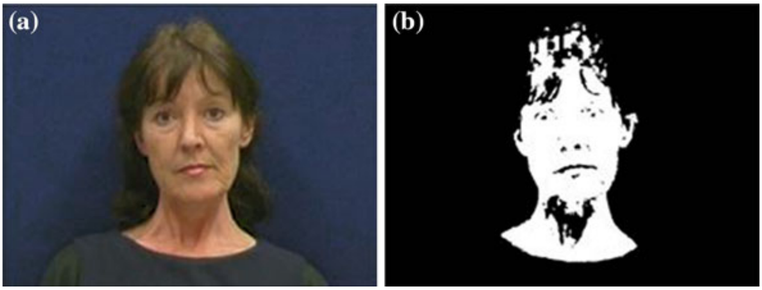


Fig. 15.7 Detecting face template after bianrizing the image. **a** Original image. **b** Binary image

$$\text{Binarized masked image} = \begin{bmatrix} \text{InRange (HSV, Scalar(minH, minS, minV),} \\ \text{Scalar (maxH, maxS, maxV), threshold)} \end{bmatrix} \quad (15.3)$$

At the end, the output of this step is the face template based on the skin mask that black pixels will correspond to background of picture while skin areas are marked white. As the result, the output will be binary image that represents the segmented area. Indeed, masking displays a region of interest (ROI) which is face template and the next processes will be applied on that region which is selected as the bunch of pixels and shape and size of the disjointed image for simplifying the facial identification computation by removing the background areas as shown in Fig. 15.7.

The outline of face template in binary image is found out by adjusting the threshold values. The binary image in OpenCV uses getStructuringElement function that takes element shape squares, the dimension and a point. This function returns a structuring element of given dimension. While the variance between ROI and non ROI detected regions grows, the change becomes bigger and the edge becomes stronger but the last output has too much noise that must be removed. Also, the small holes on the image must be filled up in order to enhance the integrity of segmented area. For resolving above mentioned issues, the morphological operations which rely only on the relative ordering of pixel values are used. Erosion and Dilation are two morphology operators which are applied.

The connected region from the binarized facial image is extracted by color, edge density and illumination cues. The dark colors and high density of edges are quite vital for detecting the face area. The dilation operator is applied twice on the binarized image for enhancing the segmented region detection through connecting the small regions and integrating them together. It connects areas that are separated by spaces smaller than the structuring element and adds pixels to the perimeter of each image object. Based on the experimental result, Dilate with larger element makes sure object is nicely visible. For this reason, in this study the binary image is divided into the neighbored 8*8 pixel set in order to consolidate each pixel set by calculating the weight of zeroes and ones, then it replaces the current pixel with the maximum pixel value found in the defined pixel set.

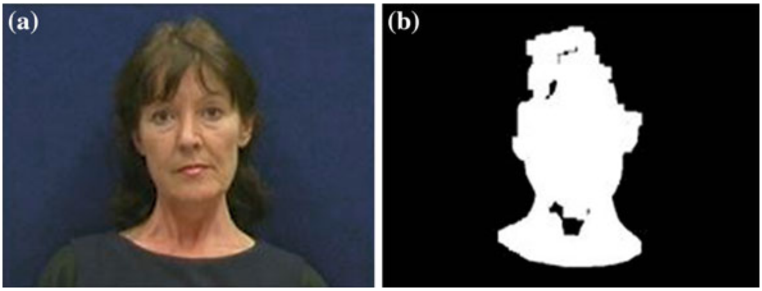


Fig. 15.8 Effect of dilation filter on original image. **a** Original image. **b** Binary image

$$\left\{\begin{array}{c} \text{Mat dilateElement} = \\ \text{getStructuringElement (MORPH\_RECT, Size(8, 8))} \\ \text{dilate (thresholdHSV, ThresholdHSV, Mat dilateElement)} \end{array}\right. \tag{15.4}$$

After applying this technique, the entire face pattern will be formed as shown in Fig. 15.8. As the result, the effects of a dilation using this structuring element on a binary image are: (1) Enlarging the sizes of face template by extending operation regions brightness (white pixels) within an image. (2) Filling up the holes and broken areas by connecting areas that are separated by spaces smaller than the size of the structuring element.

After Dilation filtering, an Erosion filter is applied twice to get rid of the unwanted connected regions. It makes the bright areas of the image smaller or not existence and the dark zones will get bigger. In this study, Erosion sets to split the binary image in to the neighbored 3*3 pixel set in order to consolidate each pixel set, it will calculate the weight of zeroes and ones then replaces the current pixel with the minimum pixel value found in the defined pixel set.

$$\left\{\begin{array}{c} \text{Mat erodeElement} = \\ \text{getStructuringElement(MORPH}_{\text{RECT}}\text{, Size(3, 3))} \\ \text{erode (thresholdHSV, ThresholdHSV, Mat erodeElement)} \end{array}\right. \tag{15.5}$$

The effect of Erosion on a binary image is to erode away the boundaries of regions of white pixels in binary image. Thus areas of foreground pixels shrink in size, and holes within those areas become larger. Indeed, it produces contrasting results when applied to binary images (As shown in Fig. 15.9).

Each slice of binary image after applying those two filters is a rectangle matrix numbers. A binary matrix comprises white pixels for face template illustration and black pixels for background illustration. For retrieving back original color pixels of ROI in the image (the masked image), the multiplying of two image matrices (binary mask and original image) is done by the bitwise AND operation. It is notable, multiplication can be done appropriately because the binary image does not effect on

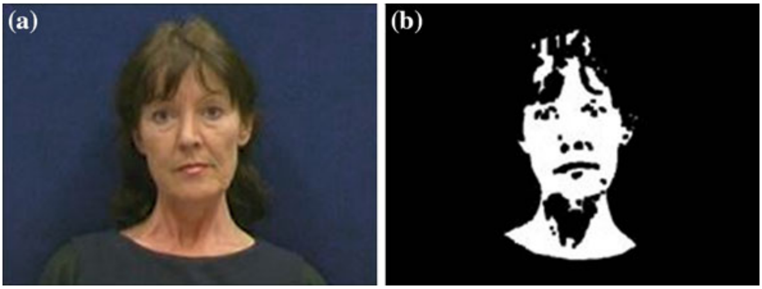
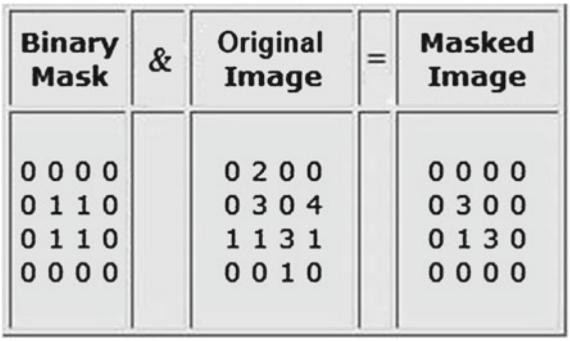


Fig. 15.9 Effect of erosion filter on original image. **a** Original image. **b** Binary image

Fig. 15.10 Matrices multiplication for attaining masked image



Binary Mask	&	Original Image	=	Masked Image
0 0 0 0		0 2 0 0		0 0 0 0
0 1 1 0		0 3 0 4		0 3 0 0
0 1 1 0		1 1 3 1		0 1 3 0
0 0 0 0		0 0 1 0		0 0 0 0

the original matrix size, number of slices and voxels in each direction. Consequently, after multiplying an original matric by a binary masked matrix, any pixel values in the original image will be retained if it is multiplied by 1 but the pixel values will be zeroed out when multiplied by 0 (Fig. 15.10).

The last output of segmentation process is the masked image where the skin color pixels of face region are detected and also all the facial features are presented in the segmented image. All the color pixels of face area cannot be precisely identified after the color segmentation because the head shape will be segmented based on the skin color range. Consequently, foreground's pixels color space ranges on the facial image that are not analogous to the skin color will be ignored and the background's pixels color space ranges that are similar to the skin color will be considered as the human skin color and will not be excluded from the original image pixels (As shown in Fig. 15.11).

15.4.3 *Eyes and Mouth Identification and Localization*

After undertaking segmentation, extracting the necessary information from the facial image for presenting facial state is important. Facial features identification and local-

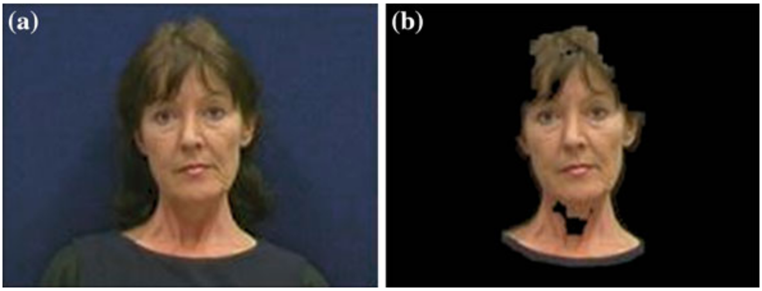


Fig. 15.11 Masked segmentation result. **a** Original image. **b** Masked image

ization are the two most beneficial methods for grabbing required information from face. Consequently, for applying the localization on the facial image in order to grab the required information, the following steps have been done: (1) Defining the Sub_ROIs. (2) Localizing the interest points. (3) Finding the mathematic associations among the selected interest points.

In order to reach this objective, the Haar-like Cascade have applied. Haar descriptor is used for understanding the contrasts difference between several rectangular regions in image. This method uses simple Haar-like features that encode the existence of oriented contrasts exhibited by a human face in the image and cascade of boosted classifier as a statistical model that is built iteratively as a weighted sum of week classifiers. The implementation of Harr-like algorithm and a face detection Harr Cascade is available as part of Open CV Computer Vision Library.

The two eyes and mouth are the most salient human facial features for frontal faces detection and using them will be beneficial for face preprocessing. As the masked images will be processed by Haar-like Cascade method, detecting two eyes and mouth will not have many false positives. Haar cascade algorithm generates data collection via 24*24 pixels rectangle choices and according to collected data, recognizes the eyes and mouth Sub-ROIs in facial image. The function detectMultiscale in OpenCV perform detecting the eyes and mouth in the segmented image by simply checking whether the suggested eye and mouth coordinates really fall in the masked facial region or not. The output of the detectMultiScale function will be a vector of the rectangular type object. Indeed, it returns the detected objects as a list of rectangles and the vector facial features will save the returned data. Moreover, the scale factor OpenCV uses for increasing the image size at each image scale and the minimum number of neighbors should each candidate rectangle to group together is considered as three. Figure 15.12 displays the rectangular Sub-ROIs given by the face detector.

The classifier is trained on images of fixed size and the trained classifiers using Haar-like features are provided in HaarTraining in OpenCv. A cascade of boosted classifiers working with haar-like features is trained with a few samples of particular facial features. The training is existed as an .xml file that contains the definitions of classifier. The following cascades are used to localize facial features (eyes and mouth).

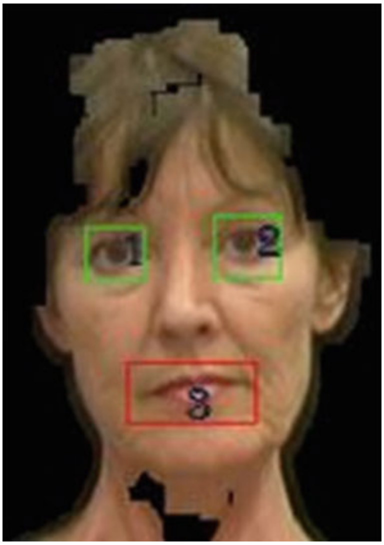


Fig. 15.12 Eyes and mouth Sub-ROI detection (*1*) left eye Sub_ROI (*2*) right eye Sub_ROI (*3*) mouth Sub_ROI

$$\left\{\begin{array}{l}\text{CascadeClassifier haar\_cascade\_eye.xml}\\ \text{CascadeClassifier harr\_cascade\_mouth.xml}\end{array}\right. \tag{15.6}$$

From the above experiment, it is exposed that the cascade of boosted classifiers based on Haar-like features gives small identification accuracy rate. As the result, regionalizing the image, the likely area, where a facial feature might exist, should be estimated. Consequently, for enhancing the accuracy of the classifier training attribute and reducing false facial feature localization, region of interest is tried to be limited by estimating the facial features locations. The facial feature cascades detect facial features in the matrix image but the best technique to remove extra feature detection is to further regionalize the region for facial feature detection. For enhancing the accuracy, eyes cascade is done in the upper horizontal half of the image, and the mouth cascade is done in the lowest horizontal 1/3 of the image (Figs. 15.13, 15.14, 15.15 and 15.16).

After reducing the area analyzed, not only localization accuracy will be enhanced since less area exists to produce false positives but also efficiency will be increased since fewer features need to be computed. This technique has been tested with successful results, because the detected eyes and mouth coordinates that do not fall in their respective defined region will be discarded but certainly it is not perfect ones because rotated and tilted faces will not be always detected or may misrepresent the actual location of the features.

Points of interest **(POI)** are the consistent selected pixels in the most descriptive areas of the given multiple facial images. Those aid to extract indispensable information about the facial image under different lighting and scale. That information not only should assist the face recognition process but also can be used for face geom-

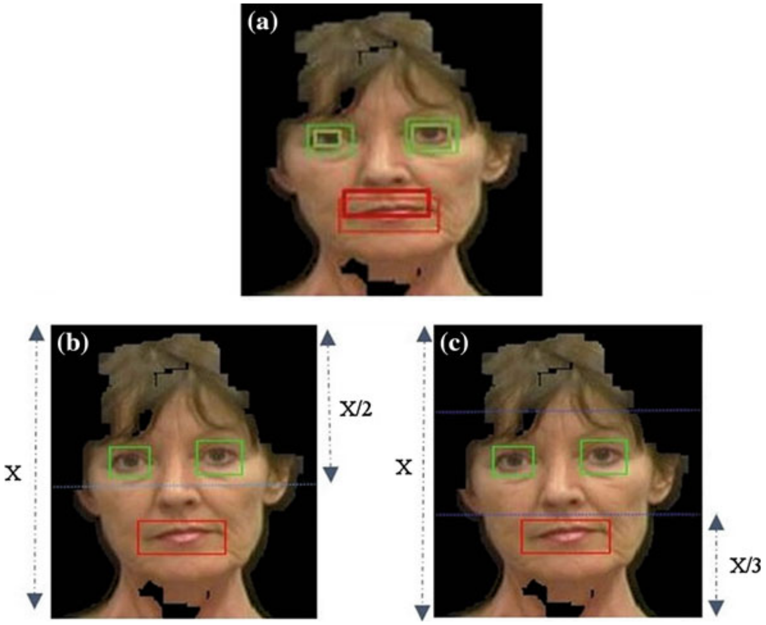
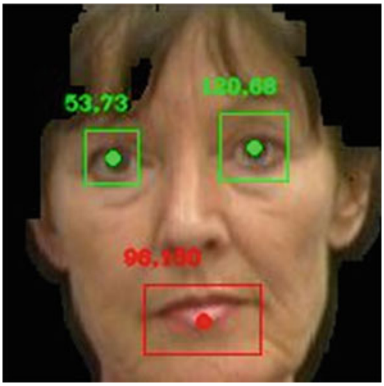


Fig. 15.13 Detection based on regionalization. **a** Detection before regionalizing. **b** Eye detection after regionalizing. **c** Mouth detection after regionalizing

Fig. 15.14 Interest points localization on facial image



etry normalization such as scaling, positioning, rotating and etc. In order to select the location of interest points and quantities of them, the characteristics of suitable points of interest must be well-defined. The consistent interest point's positions provide distinctive and unique information about each individual facial image that can be used during the recognition. Moreover, selecting the point of interest which has some association with the permanent facial features is the fairly beneficial approach

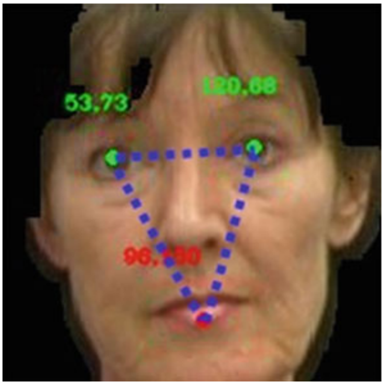


Fig. 15.15 Facial triangles on face

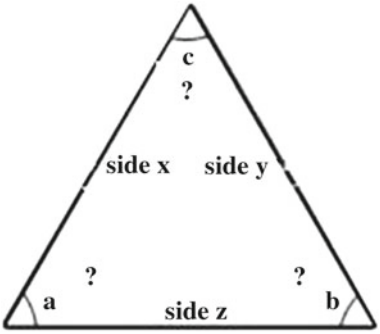


Fig. 15.16 Three sides and non-included angle given in the *triangle*

because not only facial features region can easily be obtained but also detecting them can be done simply by different detection methods.

In this study, three POIs belong to three features (Left eye, right eye and mouth) are localized. As those features have frontal view and do not have any curves. Therefore, detection will be done simpler. The eye and mouth detection is performed using three different Sub_ROIs, this supports tracing of three individual POIs in each Sub_ROI. As each rectangular Sub_ROI for each feature has the fixed specific height and width, center of them is calculated as point of interest.

$$\text{Interest Point} = ([\text{Width Sub\_ROI}/2, \text{Height Sub\_ROI}/2]) \tag{15.7}$$

Consequently, the interest point position in each facial feature Cascade is grabbed by finding the center of them which is approximately nearby the actual midpoint of each facial feature (Fig. 15.14).

As in Euclidean geometry any three points, when non-collinear, determine a unique triangle and a unique plane. With localizing those three interest points, one facial triangular can be realized on the face (Fig. 15.15).

Each triangle can be classified based on their specifications according to the relative lengths of their sides; the facial triangular specifications and perimeters are unique for each human face and dependent to position and distance between each two interest points. The distance between two interest points will be calculated based on the following Euclidean distance formulas (where Point1 = (x1, y1) and Point2 = (x2, y2)):

$$\mathrm{D} = \sqrt{(\mathrm{x}^2 - \mathrm{x}1)^2 + (\mathrm{y}^2 - \mathrm{y}1)^2} \tag{15.8}$$

As there is correlation between the sides of triangle and its interior angles, by changing the length of each edge, the interior angles will have different values. As the result angles, indicates the association of the three interest points on the face, and for this reason they are selected as the attribute on facial triangular (Fig. 15.16).

The angles in the triangle can be calculated from its sides as presented in the below mathematic formula:

$$\text{Angle (radian)} = \begin{cases} \mathrm{a} = \mathrm{Arc}\cos(\mathrm{x}^2 + z^2 - y^2)/2xz \\ \mathrm{b} = \mathrm{Arc}\cos(y^2 + z^2 - x^2)/2xy \\ \mathrm{c} = \mathrm{Arc}\cos(x^2 + y^2 - z^2)/2xy \end{cases} \tag{15.9}$$

The other benefit of calculating angles will be described more during applying the geometry normalization on the image. Since the return value of acos function (equal to Arccos in mathematical notation) for calculating the angle measures in openCV math library is radian the result of above calculation will be converted to degree as below:

$$\text{Angle (degree)} = a * \frac{180}{\pi} \quad \text{where } \pi \sim 3.14 \text{ and a} = \text{Angle(radian)}. \tag{15.10}$$

15.4.4 Average Half Face

The facial features identification is considered as one of the most challenging task because it cannot always be acquired reliably due to the quality of images, illumination, and some other disturbing factors. Furthermore, it takes a lot of computations to identify accurate facial features. The main aim to use the average-half-face for face recognition is assisting the computer to analyze a similarity measure between faces using images that have non-uniform illumination. Moreover, the calculating and using of the half average face in recognition process restricts the region of image for facial profiles extraction. The average half face-template is constructed based on the average full face-template, for reducing the symmetry redundancy of density in the full face-template. The resulting average-half-face is then used as the input for face recognition algorithms. The average-half-face is constructed based on full frontal face image in four steps as follow. (1) Rotation: The face image must be rotated for face alignment. (2) Half face calculation: The face image must be

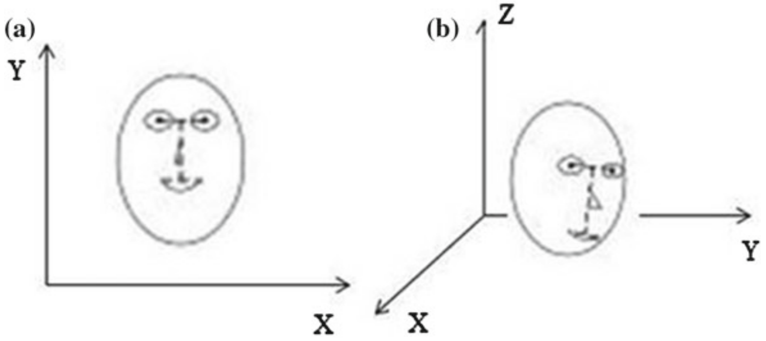


Fig. 15.17 Plotting face on 2D and 3D atmosphere. **a** Plotting face on 2D axis. **b** Plotting face on 3D axis

centered and divided in two halves. (3) Half face flipping: One half of the face image must be flipped (reversing the columns of one of the halves). (4) Average half face calculation: The two halves must be averaged together.

The challenge of visual recognition is distinguishing similar visual forms despite substantial changes in appearance arising from changes in position. The invariance to orientation and view angle are necessary for recognizing the faces because the images that are useful for face recognition always present the user from the similar angles and orientations. As all the facial images are not aligned in horizontal and vertical axes; rotating the image for standardizing the position of that before performing actual face recognition training/projection is compulsory. Moreover, the face alignment will enhance the results of this Average half face calculation. So, the initial attempts is computing the face condition for assuring that it is aligned in position, and if it is not to be aligned, the rotation angle will be used in order to normalize the input to face measurement units. In this study, it is considered that the face is irrespective to z (stacked) coordinate. In other words z coordinate is kept constant and face varies only in x (horizontal) and y (vertical) coordinates during the computation. As shown in Fig. 15.17, plotting face on two dimensional axes is irrespective to z coordinate.

An Affine Transformation represents a relation between two images. For rotating an image the linear transformation will be took place which express in the form of a matrix multiplication. The information about the relation between 2 images (X $\begin{bmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ z & \cdots & x \end{bmatrix}$ (original image), T $\begin{bmatrix} z & \cdots & a \\ \vdots & \ddots & \vdots \\ x & \cdots & b \end{bmatrix}$ (Transformed image)) can be obtained roughly, through the rotation Matrix $X \begin{bmatrix} p & \cdots & q \\ \vdots & \ddots & \vdots \\ r & \cdots & v \end{bmatrix}$ which is a geometric relation

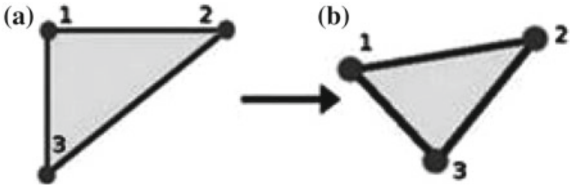


Fig. 15.18 Three facial points transformation. **a** Original *triangle*. **b** transformed *triangle*

between points that need to be applied on matrix X $\begin{bmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ z & \cdots & x \end{bmatrix}$ to obtain matrix T. $\begin{bmatrix} z & \cdots & a \\ \vdots & \ddots & \vdots \\ x & \cdots & b \end{bmatrix}$. Consequently for calculating the transformed image it is compulsory to find information for rotation Matrix M.

$$\mathrm{X}\begin{bmatrix} a & \cdots & b \\ \vdots & \ddots & \vdots \\ z & \cdots & x \end{bmatrix} * \mathrm{M}\begin{vmatrix} p = ? & \cdots & q = ? \\ \vdots & \ddots & \vdots \\ r = ? & \cdots & v = ? \end{vmatrix} = \mathrm{T}\begin{bmatrix} z & \cdots & a \\ \vdots & \vdots & \vdots \\ x & \cdots & b \end{bmatrix} \tag{15.11}$$

The matrix M $\begin{bmatrix} p & \cdots & q \\ \vdots & \ddots & \vdots \\ r & \cdots & v \end{bmatrix}$ is analyzed based on relation of three points with scale factor one on facial image. As shown in Fig. 15.4–15.6 the three points, including points 1, 2 and 3 which form a triangle on an original facial image must be mapped into rotated facial image. Thus, if the Affine Transformation with these three points is exposed, then this found relation can be applied on the entire image pixels. The three points which is considered to be used for transformation are center of eyes (already calculated during the localization) and the center of facial image (Fig.15.18).

The transform will be obtained from the relation between above three specific points on facial image by the getRotationMatrix2D function in OpenCV. In this study, the M matrix (rotation matrix) will be calculated based on two specifications; face center and amount that angle must be rotated. For calculating an Affine Transform on the image, the relationship of pixels between right and left eye centers and the center of facial image aids to calculate the amount of angles needed for aligning the face. The centralized face can be easily calculates by splitting the difference of height and width of the two centers of eyes into two. As the result if we consider c as the center of face, it will be calculated by following formula: $c = [x2 - x1/2, y2 - y1/2]$ where the coordination's of eyes center are. $\begin{cases} \text{Eye1} = (x1, y1) \\ \text{Eye2} = (x2, y2) \end{cases}$. When two lines intersect, the angle between them is defined as the angle through which one of the lines must be rotated to make it coincide with the other line. In OpenCV a positive angle is counter-clockwise and negative angle is clockwise. As shown in Fig. 15.19 the amount that

Fig. 15.19 Angle for rotating the eyes

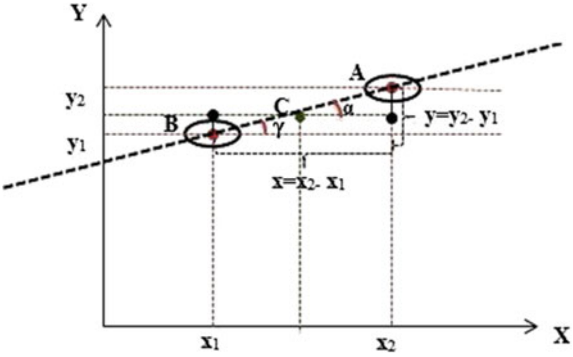
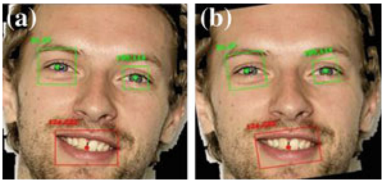


Fig. 15.20 Rotating the full face image. **a** Masked image. **b** Masked image after rotation



angle must be rotated is calculated by the below mathematic formula.

$$\gamma\,(\text{degree}) = (\text{Arc}\tan((x2 - x1)/(y2 - y1)) * \frac{180}{\Pi} \tag{15.12}$$

Based on transversal rule if two parallel lines are intersected by a transversal, then the corresponding angles are congruent. Consequently, $\gamma = \alpha$. After calculating an Affine Transform matrix of the image, warpAffine function is used for the purpose of rotating the wrapped image by applying the found rotation to the output of the Transformation. This rotation will be done with respect to the image center and the final output will be the aligned face. Consequently for rotation by an angle the original Matrix X will be rotated and transformed to Matrix T through the angles based on mathematic formula provided below.

$$\text{X}\begin{bmatrix} a \cdots b \\ \vdots \ddots \vdots \\ z \cdots x \end{bmatrix} * \text{M}\begin{bmatrix} p = \cos\alpha & \cdots & q = \sin\alpha \\ \vdots & \ddots & \vdots \\ r = -\text{Sin}\alpha & \cdots & v = -\text{Cos}\alpha \end{bmatrix} = \text{T}\begin{bmatrix} z \cdots a \\ \vdots \;\vdots\; \vdots \\ x \cdots b \end{bmatrix} \tag{15.13}$$

Applying rotation process during average half face calculation is consolidating and keeping the symmetry of pixels values in each ROI and Sub_ROI. As the result after rotating the image and splitting it from middle in to two equal portions, the pixels values are equalized in each half face area (Fig.15.20).

Once the interest points in each of the Sub_ROI are determined and the face is aligned. It needs to split the entire face into two equal halves. For this reason, the

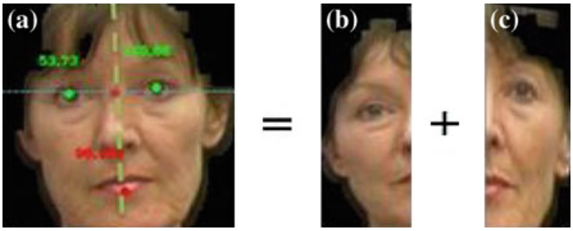


Fig. 15.21 Splitting the full face image in two halves. **a** Full face masked image. **b** *Right* half face image. **c** *Left* half face image

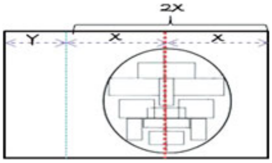


Fig. 15.22 Cropping image for having symmetric portions

line joining two points and the vertical axis must be passed through the middle of the face. The line joining includes two points which one of them is the coordinates of the center position between the two eyes and the other is the center of the mouth. Consequently, for halving the facial image the following steps have been pursued:

(1) The coordinates of the center position between the two eyes in the eye map to be computed (cEye $= \left(\frac{x1+x2}{2}, \frac{y1+y2}{2}\right)$ where $\begin{cases} \text{Eye1} = (x1, y1) \\ \text{Eye2} = (x2, y2) \end{cases}$).
(2) The coordinates of mouth center position to be computed (Mouth = (x3,y3)).
(3) The line of pixels that join those two points will be on the line with the following equation:

$$\mathrm{Y} = (\left[\frac{\mathrm{y3} - \frac{\mathrm{y1+y2}}{2}}{\mathrm{x3} - \frac{\mathrm{x1+x2}}{2}} * \left(\mathrm{X} - \frac{\mathrm{x1\ +\ x2}}{2}, \frac{\mathrm{y1+y2}}{2}\right)\right) + \frac{\mathrm{y1+y2}}{2} \qquad (15.14)$$

(4) Copying all the pixels from each side of the seamed line and store the pixels data in two different new array matrices, one of them will be called ‘Right Half’ array matrix and the other one ‘Left Half’ array matrix (Fig. 15.21).

Before calculating the average half face, the attained images after splitting into two halves, must be symmetric. On the other hand, the weight of their pixels value must be in balance. There is the probability that the face does not locate exactly at the center of image canvas. Consequently, the calculated half images based on the defined line will be asymmetric because the weights of array matrix pixels value will not be equivalent. For this reason, it is required to crop the image from side (left or right) where double of that is bigger than size of image canvas. As the result, after cropping, the face will be exactly localized at the center of the image canvas and it will be ready for averaging calculation (as shown in Fig. 15.22).

Before creating normalized average half face image, the left half face must be mirrored and then the two halves must be morphed together. The OpenCV offers

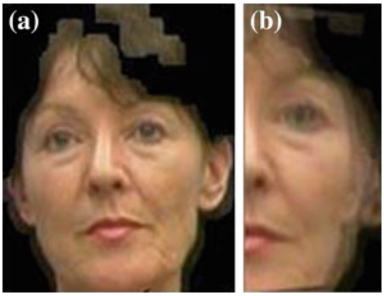


Fig. 15.23 Full face and average half face results. **a** Full face masked image. **b** Average half face image

a function to flip the source array image. Flipping a two dimensional array around vertical direction have been done through flip function and the destination array to be kept with the same size and type as source image. The full face is mixture of the symmetrical left face and the right one. As the result, the full face can be separated into the left face and the right one at the axis center of symmetry and in the following the average value of every corresponding pixel of face area to be computed. The pixel value in the k_{th} $(k = 1, 2, \ldots, n)$ position of right face area is $R_k(i, j)$ and left face area is $L_k(i, j)$, thus, the average face template can be calculated as follows: $T_k(i, j) = 1/2L_K(i, j) + 1/2R_K(i, j)$. The final result of average half face image is shown in Fig. 15.23.

15.4.5 Feature Extraction and Classification

At this time, extraction of PCA features will be presented from each average half face. In the following, the matching of features of the unknown (test feature vector) and the known (training feature vector) face images will be done for recognizing the face. The various steps are used to calculate Eigenfaces. Those steps are elaborated as below:

(1) Prepare the data (2) Subtract the mean (3) Calculate the co-variance matrix (4) Calculate the eigenvectors and eigenvalues of the covariance matrix (5) Calculate Eigenfaces (6) Classifying the faces.

15.4.5.1 Prepare the Data

The two dimensional facial image will be represented as one dimensional vector by concatenating each row or column into a long thin vector. On the other hand, for the partitioned region, significant information will be extracted and converted into a 1-D vector sequence to be aligned with the reference sequence from template image. Then, the mean data vector of images will be computed for set of sampled images (M vectors of size N) and the images training set becomes $\Gamma_{1,}\ \Gamma_{2,\ldots,}\ \Gamma_{m.}$

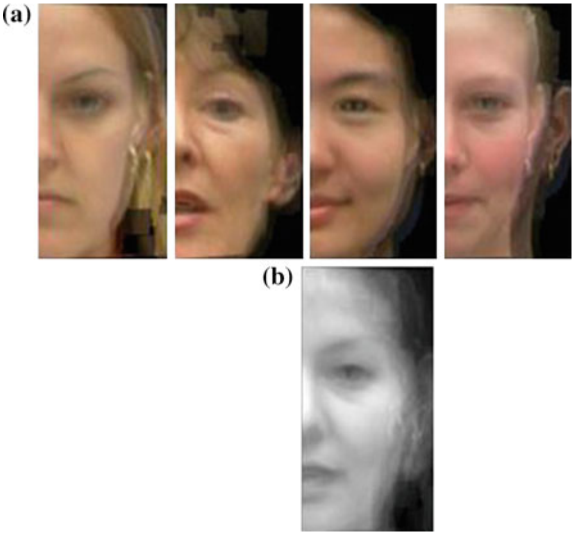


Fig. 15.24 Mean average half face made from the four average. **a** Set of original average half faces. **b** Mean average half face

15.4.5.2 Subtract the Mean

The mean data vector A has to be calculated based on the following formula $A = (1/M\Sigma_{n=1}^{M}\Gamma n)$. Then the mean result will be subtracted from the data vector of each person Γ_i and the variances between the data vectors and their average are calculated as $\Phi = \Gamma_{i-}A$. Eventually, outcome will be stored in the variable Φ (Fig.15.24).

15.4.5.3 Calculate the Co-variance Matrix

To reduce the dimension of matrices, the covariance matrix Ai will be constructed and computed as follow for each new image vector $Ai = \Phi T \Phi$.

15.4.5.4 Calculate the Eigenvectors and Eigenvalues of the Covariance Matrix

In this step, the image will be represented as a set of eigenvectors U_i (special face images) and the corresponding eigenvalues λ_i (blending ratios). Those values will be calculated from the covariance square matrix A_i.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}$$

The roots of the polynomial equation defined by $|\mathbf{A} - \lambda\mathbf{I}| = 0$ where **I** is an identity matrix. If $\mathrm{I} = 2$, then the eigenvalues are the roots of

$$\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = \left|\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix}\right|$$

$$\begin{aligned} &= (\mathbf{a}_{11} - \lambda)(\mathbf{a}_{22} - \lambda) - \mathbf{a}_{21}\mathbf{a}_{12} \\ &= \lambda^2 - \lambda(\mathbf{a}_{11} + \mathbf{a}_{22}) - (\mathbf{a}_{21}\mathbf{a}_{12} - \mathbf{a}_{11}\mathbf{a}_{22}) = 0 \end{aligned} \tag{15.15}$$

After using the quadratic formula, λ will be obtained based on the below formula:

$$\lambda = \frac{\mathbf{a}_{11} + \mathbf{a}_{22} \pm \sqrt{(\mathbf{a}_{11} + \mathbf{a}_{22})^2 + 4(\mathbf{a}_{21}\mathbf{a}_{12} - \mathbf{a}_{11}\mathbf{a}_{22})}}{2} \tag{15.16}$$

In general, the eigenvalues are the p roots of $c_1\lambda^p + c_2\lambda^{p-1} + c_3\lambda^{p-2} + \cdots + c_p\lambda + c_{p+1} = 0$ where c_j for $i = 1, \ldots, p + 1$ denote constants. Finally, when **A** is a symmetric matrix, the eigenvalues are real numbers and can be ordered from largest to smallest as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ where λ_1 is the largest. The actual component values in the eigenvalues will be computed and displayed as following code snippets.

$$\textit{Mat eigenvalues} = \textit{Recognizer} \rightarrow \textit{getMat}(\text{``}\textit{eigenvalues}\text{''}); \tag{15.17}$$

Each eigenvalue of **A** has a corresponding nonzero vector **b** called an eigenvector that satisfies $\mathbf{Ab} = \lambda\mathbf{b}$. To view the eigenvectors, the columns of data must be extracted. As data in OpenCV and C++ is stored in matrices using row-major order, for extracting a column, the Mat::clone function is used to ensure the data will be continuous, otherwise reshaping the data to a rectangle will be impossible. The actual component values in the eigenvectors will be calculated by the below code snippets.

$$\textit{Mat eigencevtors} = \textit{Recognizer} \rightarrow \textit{getMat}(\text{``}\textit{eigenvectors}\text{''}). \tag{15.18}$$

15.4.5.5 Calculate Eigenfaces

Once a continuous column Mat is obtained, the corresponding Eigenfaces F_i will be calculated based on the eigenvectors U_i, as follow: $F_{i=}[\Phi]U_i$. As shown in Fig. 15.25, for four people with one half faces for each, there will be four Eigenfaces. Indeed, for each facial image in the training set, one Eigenface will be generated. It is notable, as the Eigenface method is not scale invariant to provide the scale invariance; the database was resized once with the information gathered during the face detection process through eye width and face width.

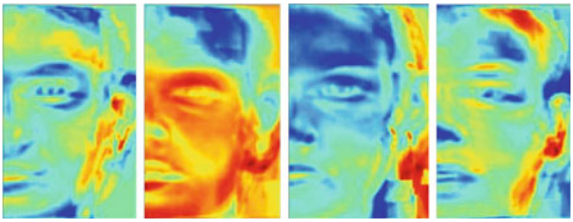


Fig. 15.25 Eigenfaces of four average half faces

15.4.5.6 Classifying the Faces

The classification step is considered as the last step for authenticating the human face; this step is applied on the trained Eigenfaces components which were already calculated by PCA machine-learning algorithm and stored in the database. The applied recognition method that reconstruct the facial image using the eigenvectors and eigenvalues, and compare this reconstructed image with the input image. The OpenCV's FaceRecognizer class generates a reconstructed face from any input image, by using the subspaceProject function to project onto the eigenspace and the subspaceReconstruct function to go back from eigenspace to image space. After combination of the trained eigenvectors with the eigenvalues from a similar test image the images that are somewhat similar to the training set will be selected as the verified one. The OpenCV's FaceRecognizer class aids to validate the two dimensional facial images simply by calling the predict function on a facial image as follows:

$$int\ identity = model \longrightarrow predict(processedFace); \tag{15.19}$$

This identity value is the label number that will be dedicated to each collected average half facial image during the training. The main issue of this technique is always predicting the most likely human face in database; even if the input facial image belongs to an unknown person. On the other hand, the verification of the human face easily will be confused with an intruder. Consequently, it is quite hard to trust the result due to recognition confusion. As confusion of recognition is more harmful than non-recognition, it is decided to practice a severe acceptance threshold in order to reject intruders and gain the more precise verification result. Indeed, establishing a threshold of Euclidean distance for confidence metric is used to judge how much the result of the prediction is reliable and in the following aids the unknown faces verification to be rejected. The acceptance threshold uses the confidence metric which is based on the distance in Eigen-subspace. The resulting weights form the weight vector ω_K^T is $\omega_K = \omega_K^T(\Gamma_\mathrm{k} - \mathrm{A})$ where k $= 1, 2, 3, 4, \ldots,$ n and $\omega_K^T = [\omega_1\omega_2\omega_{3\ldots}\omega_\mathrm{m}]$. The Euclidean distance between two weight vectors d $(\omega_\mathrm{i}, \omega_\mathrm{j})$ provides a measure of similarity between the corresponding images i and j. Moreover, it is notable that the predefined acceptance threshold is constant for all tests. Based on the experimental result, it is comprehended, if the acceptance threshold is defined too high then the recognition result will be the face of person is belong to the unknown person and if it is too low, the recognition of individuals in different pose and facial expression will not be recognized properly. This experiment shows the acceptance threshold value must be presented carefully because it can have bad or good effects on the recognition rate. In this study, the predefined confidence threshold value is implied with the value of 0.3 where the value can be between zero and one. This value is selected after 20 tests on 10 images for showing two images similarity.

15.5 Experimental Result

This section assesses the impact of the proposed model on face recognition process via a color image database. The experiments are done based on the evaluation using the heterogeneous face database for testing of system recognition performance. The set of images are used for training purpose and set of testing images are used for gallery probe matching and evaluation. The panoramic faces obtained from VidTIMIT database [48] is used as input to a recognition system. The negative and positive samples are used for evaluating the correct recognition rate with different pose and expression. Following steps have been done for evaluating the proposed framework:

(1) Understanding the ratio of training to test images numbers needed.
(2) Changing the training and test image combination.
(3) Calculating false positive rate and sensibility.
(4) Comparison of proposed method with the similar recognition frameworks.

15.5.1 Understanding the Ratio of Training to Test Images Numbers Needed

In order to appraise the proposed framework behavior and performance, first of all, it is required to know how many test images are needed for evaluating the proposed method on the small scale. As the number of ten ordered subjects can belong to one video frames in VidTIMIT database, the selected images will be unordered from twenty different subjects. In the following, the effect of different number of training and test image combinations are tested. Table 15.4 displays the performance of different tests based on the achieved success rate on set of images. The success rate of entire framework is calculated based on the ratio of number of images correctly recognized to total number of images in percentage.

Based on above experiment, choosing the five trained images per individual, adding them in the generalized test database and using five testing images for each individual indicates the better results. Therefore, the rest of the tests are conducted by using five training and five test images per person. Indeed, as it is decided to test

Table 15.4 The success rates using different number of training and test images

Total No. of images	No. of training images per subject	No. of test images per subject	Success rate (%)
20	2	8	65.00
30	3	7	70.00
40	4	6	80.00
50	5	5	90.00

Table 15.5 The success rates using variant facial pose

Test No.	Face position with different facial expression	Images tags used for training	Images tags used for testing	Success rate (%)
1	Frontal view	1, 3, 5, 7, 9	2, 3, 5, 8, 10	99.00
2		2, 4, 6, 8, 10	1, 3, 5, 7, 9	98.00
3		1, 2, 3, 4, 5	6, 7, 8, 9, 10	96.00
4	Pose variation	11, 12, 13, 14, 15	12, 116, 17, 18, 15	18.00
5		12, 116, 17, 18, 15	11, 12, 13, 14, 15	15.00

the proposed framework on the small scale, maximum of five images per individual for 20 people are decided to be evaluated.

15.5.2 Changing the Training and Test Image Combination

In the next part of the experiment, the effect of changing the images of each individual used in the training and testing stage in a rotating manner has been studied. Indeed, the recognition framework success rate is tested based on comparing the basis images which existed in the database with the images that are not existed in the database for each person and tried to understand the impact of exchanging the test image and the trained image with each other on success rate. Moreover, faces with the variant poses are considered to be tested in order to understand the success rate of them in compare with faces with the frontal view. As shown in Table 15.5, it is observed that the success rate changes with respect to the utilized sets of training and testing images. Based on the achieved experimental result, variations in pose cause the success rate to be incredibly decreased to 15 %. The best achieved result belongs to face frontal view which is 99 % when total numbers of training samples are twenty five. Furthermore, exchanging the test image and training image does not impact too much on success rate of the proposed recognition framework.

Table 15.6 The contingency table

Predicted value	Observed Value		
		Positive result (%)	Negative result (%)
	True image	84	16
	False image	92	8

Table 15.7 Face recognition performance result

True acceptance rate (%)	False acceptance rate (%)
91.30	33.33

15.5.3 Calculating False Positive Rate and Sensibility

In this part, the contingency table (as shown in Table 15.6) is erected. This table aids to understand the true positive and true negative results of systems. Moreover, error rates of the system based on the type I and II errors are calculated.

In the following, the probability that the face to be recognized positively, given that test image is not in the database (false acceptance rate) and the face to be recognized positively, given that test image is in the database (the true acceptance rates or sensibility) are[1] calculated based[2] on below formulas:

(1) $False\,Acceptance\ Rate = \frac{False\ Nagative\ Result}{Negative\ Result}$

(2) $True\,Acceptance\ Rate = \frac{True\ Nagative\ Result}{Positive\ Result}$

The face recognition performance result is tabulated in Table 15.7 after doing image preprocessing.

15.5.4 Performance Comparison of Proposed Method with the Similar Recognition Framework

In this section, the robustness of our proposed recognition framework is verified in compared with the similar frameworks using PCA Algorithm. Indeed, the result of the proposed framework is compared with the experiments of full face recognition framework on 2003 [49]. Effectiveness of the face recognition methods is evaluated based on false rejection rate and false acceptance rate for the verification operation.

Heseltine et al. [49] claimed the false rejection rate of their framework using simple PCA algorithm without image preprocessing is 25.5 % and their false acceptance rate is above 40 %. After applying the combinations of color normalization, statistical

[1] Negative Result = False Image Negative Result + True Image Negative Result.

[2] Positive Result = True Image Positive Result + False Image Negative Result.

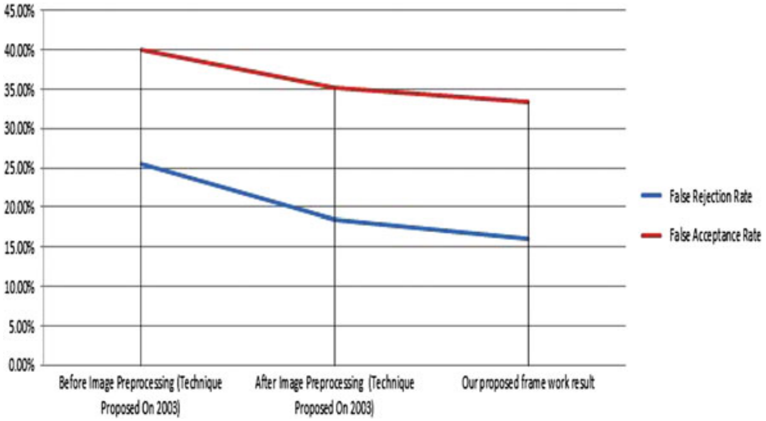


Fig. 15.26 Error rate and false acceptance rate comparison after and before image preprocessing

methods, convolution filters as the image pre-processing techniques their false rejection rate is decreased to 18.4 % and the false acceptance rate is decreased to 35.10 %. The Fig. 15.26 illustrates the comparison result among our recognition error rates and the appearance based approached offered after and before image preprocessing.

15.6 Conclusion

In this chapter a recognition framework is proposed for human face recognition. The comprehensive proposed frameworks, including different processes, techniques and algorithms that are used for human face authentication. The Multi image preprocessing techniques are applied on an image for decomposing an image into different sub bands. Those techniques and algorithms are used for regionalizing the image, eliminating unnecessary image pixels, taking needed information from facial image and identifying two dimensional head model and face on images. The color thresholding segmentation and Haar-like cascade algorithm are the techniques that assist to achieve above goal. The color thresholding has negative impact on performance of Haar-like Cascade algorithm detection, due to missing the color pixels of face. On the following the alignment is done by global and local transformations of the whole image and facial features, respectively in order to apply the required changes on image. Indeed, above techniques assisted to detect head model and normalize the facial image. Finally, these aligned cropped virtual view images after image preprocessing are used as training data for the PCA technique. The PCA algorithm is applied on preprocesses image to authenticate the preprocessed image based on Eigen distances calculation and comparisons. The most notable and significant key results after doing numerous testing are summarized as below:

(1) After assessing the proposed recognition framework. The false rejection rate is 16 %, false acceptance rate is 33.33 % and true acceptance rate is 91.30 %. These measurements imply that performance of the recognition system using PCA algorithm after image preprocessing enhanced in comparison with does not applying the image preprocessing methods.
(2) The trained average- half face enhances the performance of recognition in compare with the full face image. As shown in Fig. 15.26, the false rejection rate is decreased from 18.14 to 16 % and false acceptance rate is decreased from 35.10 to 33.33 %. These measurements illustrate the impact of the average half face on recognition rate is extremely constructive.
(3) The variations in pose cause a large reduction in verification accuracy in compare with the frontal view facial image. As demonstrates on Table 15.5, the recognition rate of the posed face is around 15 % but this rate for frontal view is increased to minimum 96 %. Indeed, this face recognition method is mostly reliable in the certain conditions that will be trained for.
(4) The half-face image reduces storage into half in compare to full image because the size of each image will be declined to half.

The major advantages of above recognition framework are due to slightly intimating the illumination issue by using averaging technique and enhancing accuracy of recognition by increasing the true acceptance rate of recognition. The main disadvantage of proposed framework is by reason of less identification and recognition precision for the posed faces. In spite of, achieving the acceptable performance result from the proposed recognition framework, to further improve; there are some possible system enhancements:

- Multiple face detectors can be trained, including poses other than frontal. The detectors can be based on particle filtering with a probabilistic tracker or color histogram with a deterministic tracker which aid to capture more pose variation.
- More facial segmentation can be applied for wrapping the face and segmenting the face based on the significant features in face. In the following human face can be verified based on applying the appraising method for recognizing the set of extracted facial features.
- Other distance metrics such as Mahalanobis and Manhattan distances can be used for classification.
- Different classification algorithms such as SVM or LDA can be applied in order to reduce the recognition time complexity and increase performance of verification system.

Acknowledgments We express our deepest appreciation to Universiti Teknologi Malaysia for their financial support and encouragement during the course of this study, colleagues for their invaluable view and tips and lastly to our family for their spiritual support and encouragements.

References

1. Rahman, N.A.B.A., Bafandehkar, M., Nazarbakhsh, B., Mohtar, N.H.B.: Ubiquitous Computing For Security Enhancement Of Vehicles, IEEE, pp. 113–118 (2011)
2. Delac, K., Grgic, M.: Face Recognition. I-TECH Education and Publishing, Vienna (2007)
3. Mann, S.: Intelligent Image Processing. John Wiley & Sons Inc., Toronto (2002)
4. Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. **3**(3), 1–16 (1991)
5. Marques, O.: Practical Image and Video Processing. John Wiley & Sons Inc., Florida (2011)
6. Li, S.Z., Zhang, L., Liao, S.C., Zhu, X.X., Chu, R.F., Ao, M., He, R.: A Near-Infrared Image Based Face Recognition System, pp. 1–6. Institute of Automation, Chinese Academy of Sciences, Beijing (2004)
7. Tan, X., Chen, S., Zhou, Z.-H., Zhang, F.: Face Recognition from a Single Image per Person. Nanjing University of Aeronautics and Astronautics, Nanjing (2010)
8. Lu, X.: Image analysis for face recognition. Michigan State University, pp. 1–37 (2012)
9. Patra, A.: Development of efficient methods for face recognition and multimodal biometry. Indian Institute Of Technology Madras, pp. 1–176 (2006)
10. Sandhu, P.S., Kaur, I., Verma, A., Jindal, S., Singh, S.: Biometric methods and implementation of algorithms. Int. J. Electr. Electron. Eng. **3**(8), 492–497 (2009)
11. Olivares-Mercado, J., Aguilar-Torres, G., Toscano-Medina, K., Nakano-Miyatake, M., Perez-Meana, H.: GMM vs SVM for face recognition and face verification. In: Corcoran, P.M. (ed.) Reviews, Refinements and New Ideas in Face Recognition, pp. 1–338. InTech, Rijeka (2011)
12. Tan, X., Chen, S., Zhou, Z.-H., Zhang, F.: Face Recognition from a Single Image per Person. Institution of Automation, Chinese Academy of Sciences, Beijing
13. Gupta, A., Dewangan, V., Ravi Prasad, V.V.: Facial Recognition, Infosys, pp. 1–12 (2011)
14. Burger, W., Burge, M.J.: Principles of Digital Image Processing. Springer, London (2009)
15. San Martin, C., Carrillo, R.: Recent Advances on Face Recognition Using Thermal Infrared Images. Springer, London (2009)
16. Singh, S.K., Chauhan, D.S., Vatsa, M., Singh, R.: A robust skin color based face detection algorithm. Tamkang J. Sci. Eng. **6**(4), 227–235 (2003)
17. Cheddada, A., Mohamadb, D., Abd Manaf, A.: Exploiting Voronoi diagram properties in face segmentation and feature extraction. Pattern Recogn. **41**(12), 3842–3859 (2008)
18. Skarbek, W., Koschan, A.: Colour Image Segmentation. Technische Universitat Berlin, Berlin (1994)
19. Mohammad S.I., Azam T.: Skin color segmentation in YCBCR color space with adaptive fuzzy neural network. Image Graph. Signal Process. **4**, 35–41 (2012)
20. Corcoran, P.M.: Reviews, Refinements and New Ideas In Face Recognition. InTech, Rijeka (2011)
21. Phung, S.L., Bouzerdoum, A.: Skin segmentation using color pixel specification: analyse and comparison. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 146–154 (2005)
22. Maini, R., Aggarwal, H., Study and comparison of various image edge detection techniques. Int. J. Image Process. **3**(1), 1–12 (2012)
23. Curran, K., Li, X., McCaughley, N.: The use of neural networks in real-time face detection. J. Comput. Sci. **1**(1), 47–62 (2005)
24. Wong, K.-W., Lam, K.-M., Siu, W.-C.: An efficient algorithm for human face detection and facial feature under different condition. Pattern Recogn. **34**, 1993–2005 (2001). (Pergamon)
25. Jeng, S.-H., Liao, H.Y.M., Chin C.H., Ming Y.C., Yao T.: Facial feature detection using geometrical face model: an efficient approach. Elsevier Sci. **31**(3), 273–282 (1998)
26. Barequet, G., Dickerson, M., Eppstein, D., Hodorkovsky, D.: On 2-site Voronoi diagrams under geometric distance functions. J. Comput. Sci. Technol. **28**(2), 267–277 (2013)
27. Benson, D.J.: Computational Methods in Lagrangian and Eulerian Hydrocodes. University of California, San Diego (2003)
28. Rosen, D.: Parametric modeling. 9 7 2013. http://www.srl.gatech.edu/education/ME6175/notes/ParamModel/Para. Accessed 20 July 2013

29. Salah, A.A., Akarun, L.: 3D Facial Feature Localization for Registration. Bogazigi University, Istanbul (2012)
30. Phillip I.W., John F.: Facial feature detection using haar classifiers. J. Comput. Sci. Coll. **21**, 127–133 (2006)
31. I. M'. es: Face Recognition Algorithms. Universidad del Pais Vasco, pp. 1–78 (2010)
32. Costache, G., Mangapuram, S., Drimbarean, A., Bigioi, P., Corcoran, P.: Real-time video face recognition for embedded devices. In: New Approaches to Characterization and Recognition of Faces, pp. 115–130. InTech, Rijeka (2012)
33. Jyoti S.B., Sapkal, S., Comparative study of face recognition techniques. Int. J. Comput. Appl. **ETCSIT**(1), 12–17 (2012)
34. Degtyarev, N., Seredin, O.: Comparative Testing of Face Detection Algorithms. Tula State University, Tula (2013)
35. Gnanaprakasam, C., Sumathi, S., Rani Hema Malini, R.: Average-Half-Face in 2D and 3D Using Wavelets for Face Recognition, WSEAS International Conference on Signal Processing, pp. 107–113 (2013)
36. Chawla, N.V., Bowyer, K.W.: Designing Multiple Classifier Systems for Face Recognition, pp. 407–416. Springer, Berlin (2005)
37. Bhadu, A., Kumar, V., Hardayal S.S., Rajbala T.: An improved method of feature extraction technique for facial expression recognition using Adaboost neural network. Int. J. Electron. Comput. Sci. Eng. **1**(3), 1–7 (1956)
38. Aguilar, G., Olivares, J., Sánchez, G., Pérez, H., Escamilla, E.: Face Recognition Using Frequency Domain Feature Extraction Methods. Instituto Politécnico Nacional, SEPI Culhuacan, México (2013)
39. Harguess, J., Aggarwal, J.K.: A Case for the Average-Half-Face in 2D and 3D for Face Recognition. Austin (2012)
40. Tan, X.: Face Recognition from a Single Image per Person. Nanjing University of Aeronautics and Astronautics, Nanjing (2010)
41. Zhao, W., Chellappa, R., Phillips, P.J.: Subspace Linear Discriminant Analysis for Face Recognition. University of Maryland, Maryland (1999)
42. He, X., Niyogi, P.: Locality Preserving Projections. The University of Chicago, Chicago (2010)
43. Brunelli, R., Poggio, T.: Face recognition: feature versus template. IEEE Trans. Pattern Anal. NAS Mach. Intell. **15**(10), 1042–1052 (1993)
44. Mohamad, F.S., Manaf, A.A., Chuprat, S.: Histogram-Based Fruit Ripeness Identification Using Nearest-Neighbor Distance, FITC, pp. 1–4 (2010)
45. Olivares-Mercado, J., Aguilar-Torres, G., Toscano-Medina, K., Nakano-Miyatake, M., Perez-Meana, H.: GMM vs SVM for Face Recognition and Face Verification. National Polytechnic Institute, Mexico (2009)
46. Wu, Y., Chan, K.L., Huang, Y.: Image Texture Classification Based on Finite Gaussian Mixture Models. Nanyang Technological University, Singapour (2013)
47. Lucey, S., Ashraf, A.B., Cohn, J.F.: Investigating Spontaneous Facial Action Recognition Through AAM Representations of the Face. Carnegie Mellon University, Pennsylvania (2013)
48. Sanderson, C.: Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag, Saarbruecken (2008)
49. Heseltine, T., Pears, N., Austin, J., Chen, Z.: Face recognition: a comparison of appearance-based approaches. In: VIIth Digital Image Computing: Techniques and Application, pp. 1–10 (2003)

Chapter 16
Biometric and Traditional Mobile Authentication Techniques: Overviews and Open Issues

Reham Amin, Tarek Gaber, Ghada ElTaweel and Aboul Ella Hassanien

Abstract Currently, mobile smartphone devices contain a critical and sensitive data. In addition, they provide access to other data, on cloud for example, and to services somewhere on the Internet. Mobile authentication aims to protect against unauthorized access. The current operating systems of mobile smart phones offer different authentication mechanisms. Nonetheless, in some situations, these mechanisms are vulnerable and in other situations, they are not user friendly enough, thus not widely adopted. In this chapter, we will give an overview of the current mobile authentication mechanisms: traditional and biometric, and their most commonly used techniques in the mobile authentication environment. In addition, the pro and cons of these techniques will be highlighted. Moreover, a comparison among these techniques will be conducted. The chapter also discuss the other techniques which could much suitable for the current environment of the mobile applications. Furthermore, it discuss a number of open issues of the mobile authentication which needs further research in the future to improve the adoption of the biometric authentication in the smartphones environment.

16.1 Introduction

Mobiles are considered the largest market portion. Mobile phone has become incredible device which can be used for various tasks including telephony, multi-networking, entertainment, business functions, computing and multimedia. In other words, mobile

R. Amin · T. Gaber (✉) · G. ElTaweel
Faculty of Computers and Informatics, Suez Canal University, Ismailia, Egypt
e-mail: tarekgaber@ci.suez.edu.eg

A. E. Hassanien
Faculty of Computers and Information Science, Cairo University, Cairo, Egypt

T. Gaber · A. E. Hassanien
Scientific Research Group in Egypt (SRGE), Cairo, Egypt

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_16, © Springer-Verlag Berlin Heidelberg 2014

devices are being used worldwide, not only for communication purposes, but also for personal affairs and for processing information obtained anywhere at any time. Tesng et al. in [1], have reported that more than 4 billion users are using mobile phones around the world. They also expected that by 2015, around 86 % of the world population would own at least one mobile phone.

Most of mobile phones are currently embedded with digital imaging and sensing softwares. Examples of these sensors include voice sensors (microphones), GPS sensors, optical/electrical/magnetic sensors, temperature sensors and acceleration sensors. Such sensing softwares have many applications including medical diagnostics, e.g. heart monitoring, temperature measurement, hearing and vision tests, thus helping in the improvement of the health care [2].

Due the high capabilities of mobile phones, they had confirmed themselves to be highly attractive aims for theft. This theft is usually not only because of the mobile cost but also because of gaining access to the owner's information. For instance, steal owner's identity to buy goods online at the owner expense or to explore new functionality are considered the prime motivations for theft [3].

How precious your phone is not only depending on its price or new technology added to it but also on the information saved on it. Examples of this information include some information related to your work, to your online banking, or your health case. Such information is considered much important than the mobile itself. Therefore, it is a crucial mission to protect the way to access the mobile. Mobile authentication is the first step to protect mobile's access. It could be seen as a gateway to get access to a mobile. The main aim of the authentication technique is addressing the question: "How user proves to device that he is who is claimed to be?".

There are two major categories of authentications: *traditional and biometric*. A summary of these two categories are shown in Fig. 16.1. The more easy to use the authentication method, the more attractive the method will be to the user. However, the user attractiveness is not only the accurate proof of the method's efficiency but also there are other factors to evaluate the authentication method.

In this chapter, we will give an overview of the current mobile authentication mechanisms: traditional and biometric, and their most commonly used techniques in the mobile authentication environment. It will also conduct a comparison between these common techniques and highlight some open issues to improve the usability and security of the authentication systems to suit the mobile constraints. The rest of the chapter is divided into four sections. Section 16.2 introduces the traditional authentication techniques while Sect. 16.3 presents overview of the biometric authentication system then reviews the most commonly used biometrics which is divided into physiological and behavioral biometric techniques. Section 16.4 presents the difference between the traditional and the biometric authentication techniques. Section 16.5 gives a comparison between the various authentication techniques. Section 16.6 presents the comparison between the explicit and implicit authentication techniques. Section 16.7 gives some open points for further search and finally Sect. 16.8 concludes the overall chapter.

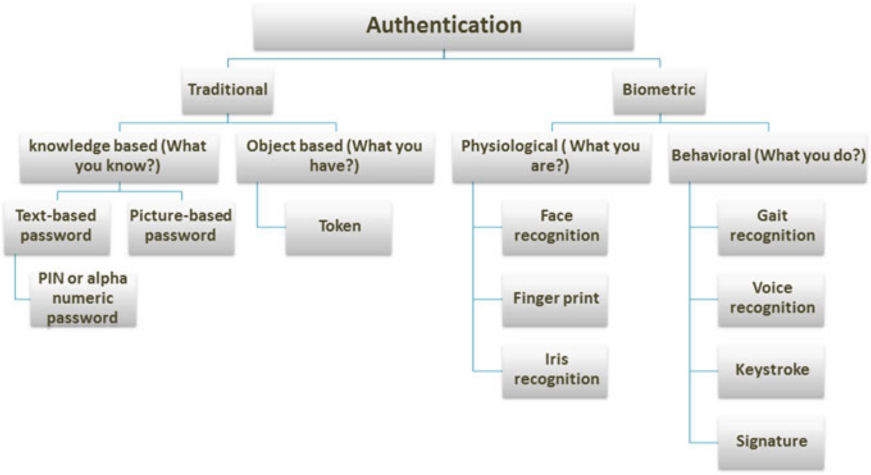


Fig. 16.1 Classification of authentication techniques

16.2 Traditional Authentication Methods

Traditionally, authentication methods are either knowledge-based or object-based authentication [4, 5]. Knowledge-based method depends on what user already knows, whereas object-based method depends on what user already has. In the next sections, an overview of these two classes is given.

16.2.1 Knowledge-Based Authentication

Knowledge-based techniques are the most common used authentication techniques. They are based on "What user knows?" to identify his/her. They include two classes: text-based and picture-based passwords [6].

Text-based password Text-based password includes Personal Identification Numbers (PINs), Personal Unblocking Key (PUK), and alpha numeric password. The most widely known is the PIN. Upon switch-on the mobile device, a user is asked to enter the correct 4–8 digit PIN [5, 7].

PIN is also used to protect the SIM card. After 3 failed attempts to enter the PIN, the SIM locks out and the PUK (PIN Unlock) is then requested. If the PUK is also falsely inputted for 10 times, the SIM becomes useless [8]. PIN achieved probability of as low as 10^{-n} to accept imposters falsely, where 10 is the range of numbers from 0 to 9 that could be used for PIN and n is the length of the PIN digits [9].

Another advance to the text-based password was to use the *alphanumeric password*. This makes password more difficult to guess and maximize the probability to accept imposters. This enhances the quality of PIN by using Non-Dictionary words to avoid risk of dictionary based attacks. However alphanumeric password could be violated by the brute-force approach (testing every combination of characters for every length of password). One solution to this problem was to use passwords that mixed between case/symbols or the Password ageing to change password regularly [7]. Nonetheless, this technique suffers from the following problems:

- The memory load to remember
- Shoulder surfing attack
- Reusing of the same password for multiple accounts
- Disturbing user with frequently entering
- Writing password down
- Need for additional customer service with the incorrect PUK code.

These problems push people to use weak passwords or irregularly change them or never use any password at all. In addition, password isn't actually representing its owner. This makes theft of owner's identity is much easier. Although the previously mentioned password problems, password authentication method has a good advantage over the other methods. It can be changed anytime [10].

Picture-based passwords The main problem with PIN was a memorization. To address this problem, Graphical Password was suggested. This method is based on the fact that people are easily remember images than strings. Graphical Password (GP) was originally introduced by Greg Blonder in 1996 [11]. There are three ways to implement it. It could be implemented by drawing curve connecting selected picture or by selecting some specific images or by pointing to points at some image.

Gao et al. [12] proposed authentication system by drawing a curve to connect specific degraded pictures. These pictures make a story for the user, thus enabling him to easily remember the password. For example, "mom and dad takes baby to doctor to get medicine" is password. Although the system is resistant against shoulder surfing, it takes time from user to login or differentiate the degraded pictures. According to user behavior: this method starts and ends with random pictures which could be forget by users at the first stage.

Khan et al. [6] presented a hybrid technique which combines recognition (select specific symbols) and recall (try to redraw selected symbol on screen). Firstly, a user has to enter username and password. Secondly, the user has to select at minimum 3 symbols using recognition. Then, the user draws these symbols on touch screen using recall by stylus or pen which is needed to get access. Then, a processing of the drawn symbol by normalization, noise removal, and other operations, e.g. feature extraction and hierarchial matching are performed. The feature extraction is used to extract (hyper stroke, bi-stroke and stroke) features while the hierarchical matching is used to check that the user has drawn objects by the same order during the selection stage. This method is illustrated at Fig. 16.2.

This technique makes hacking more difficult as the user not only enters a username/select symbol but also draws the object. However entering a username and a password are still related with the password problems mentioned above. In addition,

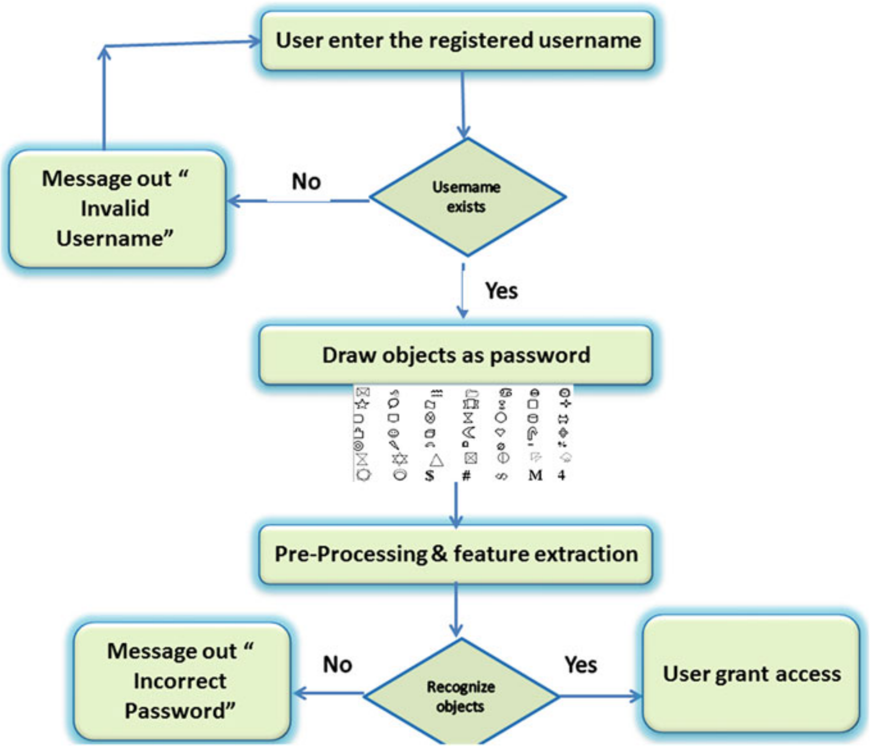

User enter the registered username
Username exists
No
Message out " Invalid Username"
Yes
Draw objects as password
Pre-Processing & feature extraction
Recognize objects
No
Message out " Incorrect Password"
Yes
User grant access

Fig. 16.2 Authentication phase for graphical password

it takes time to perform preprocessing on the drawn image and the other operations used to compare the drawn and stored objects. Moreover, when the user uses different hand for drawing, symbol differs, this is another problem.

16.2.2 Object-Based Authentication

Using passwords is the easiest authentication methods to access all mobile's data. However, as shown above, it is subject a number of problems. To avoid these problems, object-based authentication techniques were developed as a second factor for authentication besides password. These techniques include tokens-based authentication. The tokens are physical devices storing passwords. Examples of tokens include driver license and remote garage door opener [13].

Using token for authentication means that a user deals with some hardware to carry out the authentication process. This hardware contains software programs that implement a One-Time Password (OTP) algorithm to provide changed-over-time

PIN (random password) which is synchronized with a server [14]. Seed value of the PIN and a timestamp are given to a token algorithm to make predicting the random password more difficult to attackers [15].

For example, in order for a user access to his bank account through his credit card (token), he must first enter his username and corresponding password [6].

In mobile's environment, a SIM card is considered as an authentication token for a network subscribers. Every SIM consists of two unique identifiers: *IMSI* and *Ki*. *IMSI* (International Mobile Subscriber Identity) is 15 digits that uniquely identify mobile subscriber. *Ki* (Individual subscriber authentication Key) is a random number of 128-bits which is cryptographic key to be used to generate session keys. These identifiers are used to uniquely identify a legitimate user [8]. The secrecy of both *IMSI* and *Ki* provides authentication of the user's data. SIM-based authentication enables a user to subscribe with a network. Nonetheless, it doesn't check whether the mobile's user is the registered subscriber or not [16].

In general, tokens are more efficient than passwords as with them it is very difficult for an attacker to guess or remember tokens. However, the tokens-based authentication suffer from the following problems/limitations [15, 16]:

1. Additional cost comes from manufacturing/maintenance and installation/deploying for both the hardware and software.
2. Need for high computation in the poor constrained mobiles.
3. Effort to manage.
4. Possibility to be lost or stolen.
5. User has to hold or wear the token.

Clarke et al. [16] proposed a system which accommodates the above problems by developing tokens that authenticate users through using the wireless connection provided by mobiles. So that, tokens do not require passwords to be physically stored at a server synchronized with the token. These tokens could be worn like jewelry. However, at most cases mobiles are used as a stand-alone OTP Token [17, 18].

16.3 Biometric Authentication Techniques

The word biometrics comes from two ancient Greek words: bios = "life" and metron = "measure" [6]. In Paris, in the 19th century, Alphonse Bertillon, who worked as a chief of the criminal identification division of the police department, practiced the usage of body properties (biometrics) (e.g. fingers, height, or feet) to identify criminals. He also discovered the uniqueness of human fingerprints. Soon after this discovery, police started to save criminals fingerprints using card files. Later, police started to lift fingerprints from crime scenes and compare it with the stored ones to know criminals identities. Since then, biometrics has become a subject of interest in many areas for personal recognition such as: authentication in sensitive jobs such as people working at national security organization [19]. In the following sections, an overview of biometric authentication and its types are presented.

16.3.1 Biometric Authentication and Its Types

As explained in Sect. 16.2, the traditional techniques do not actually represent users. On the other hand, biometric authentication techniques depends on the users' unique features to identify the users [20]. The biometric authentication is a process in which a user is recognized automatically based on a feature vector extracted from his physiological or behavioral characteristics. Based on this, biometric methods are typically categorized into two types: physiological and behavioral. Physiological biometrics depends on physical attributes of a person such as what user already has (e.g. face, fingerprint or hand). Generally, it is based on the fact that these person's attributes do not change over time. Conversely, behavioral biometrics depends on an associated behavior of a person such that what user does (e.g. how a person writes or speaks) [16]. This behavior is recorded in a period of time while the person is doing his job from his temporal trait [21].

The main difference between the physiological and the behavioral biometrics is that the latter is more difficult to detect and emulate because it depends on an interaction of users with their own devices to extract specific and accurate habits. A detailed information about physiological and behavioral biometrics is given below.

16.3.2 Components of a Typical Biometric System

A typical biometric authentication system consists of five modules/components which are shown in Fig. 16.3.

- Sensor module: It is used to capture user's raw biometric data. An example is camera used to take a picture of human face.
- Feature extraction module: It is used to process the acquired biometric data to extract a set of features. For example, features on the surface of a face, such as the contour of the eye sockets, nose, and chin can be extracted.
- Matcher module: It is used to compute matching scores of comparing the extracted features against the stored ones.
- System database module: It is used to store the biometric templates of features the enrolled users [21].
- Decision-making module: It is used to either determine the user's identity or confirm the users claimed identity [22].

16.3.3 How Biometric Authentication Works?

Biometric system depends on comparing the recent feature set against the set stored in a database. It works in two stages: enrollment and recognition (verification or identification) [23].

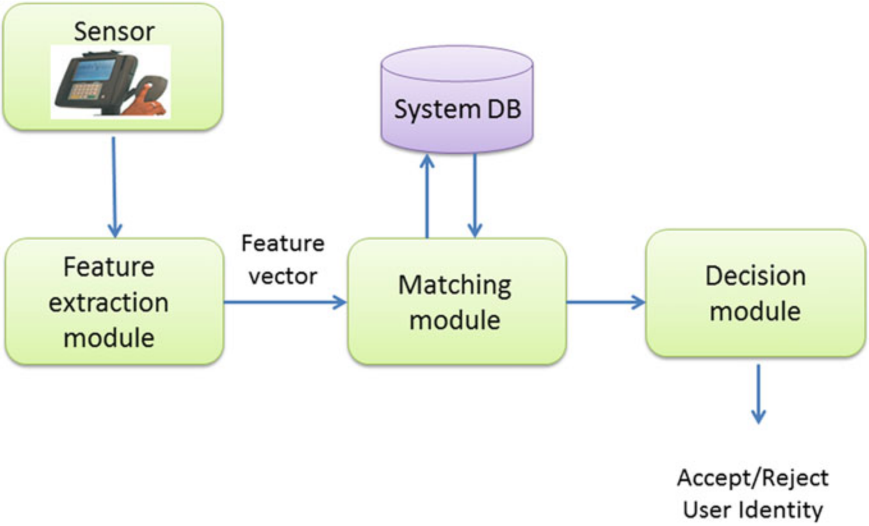


Fig. 16.3 Biometric modules

In the enrollment stage, a set of feature is extracted from the raw biometric template and then is stored in a database [24]. Along with some biographic information (e.g. name or PIN) which describing the user can be possibly stored with the feature set. The user's template can be extracted from either a single biometric or multiple samples. Thus, multiple samples of an individual's face, captured from different poses with respect to the camera generate the user's template.

In the recognition stage, an individual is verified whether he/she was really enrolled in the system. Depending on the application context, the recognition process can be achieved either in identification mode or in verification mode. With the identification mode, as seen in Fig. 16.4), a user does not claim his/her identity but the system searches all stored templates for all enrolled users in the database for a positive match. This means that the identification mode conducts a one-to-many comparison between the given user's template and the stored templates in the database [21].

In the verification mode, as shown in Fig. 16.5, a user first claims an identity, usually by entering a PIN, and then the system confirms whether this user is the one who has just claimed the identity corresponding to the PIN [21, 25]. Unlike the identification mode, the verification one performs a one-to-one comparison between a live biometric template (just built by the system) and the retrieved one from the database [24].

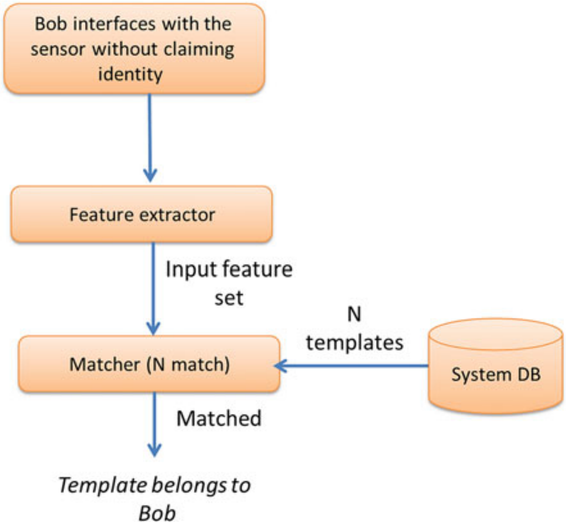


Fig. 16.4 Identification stage

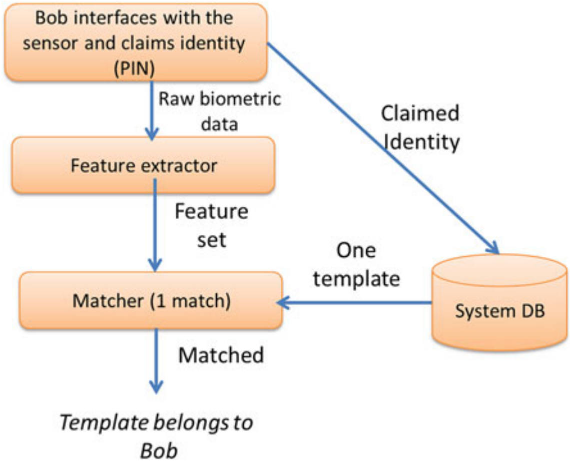


Fig. 16.5 Verification stage

16.3.4 Performance Evaluation of Biometric Authentication

The most widely used method to evaluate the performance of biometric systems include False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Rate (EER) (defined in ISO/IEC FDIS 19795-1). The FAR means the probability of accepting an impostor falsely, while the FRR means the probability of rejecting a rightful owner falsely. These probabilities depend on a predefined threshold which determines when the system accepts or rejects a user. However, the value of this threshold could affect the overall result of accepting or rejecting users. In case, a low threshold value is used, the system could output a high FAR value. In case, a high threshold value is used, the system could result in a high FRR value. Generally speaking, decreasing the FRR increases the FAR and vice versa. To achieve a tradeoff between the two cases, EER (Equal Error Rate) is employed to get the intersection of the values of FRR and FAR [26].

16.3.5 Physiological Biometrics

The first type of the biometric authentication approaches is the physiological which depends on what a user already owns. There are many physiological techniques including face, fingerprint, iris hand vascular, palm-print, and hand or ear geometry recognition. Due to usability and hardware constraints, not all of these biometric methods are suited for mobiles. Such methods include hand vascular, palm-print, and hand or ear geometry recognition [5, 27]. Below, we will give a review of most common biometric techniques used in mobile authentication.

Face Recognition Face recognition method is the one in which a human face is captured using a mobile's camera then this face is used to authenticate this human to the mobile. The face recognition authentication makes use one of two ways: (1) shape and location of facial properties such as the eyes, nose, lips and their spatial relationships, or (2) the overall face image [21]. Tao et al. [9] have developed a biometric system using a user's mobile camera which takes 2D face image to ensure the existence of user. This authentication system consists of five modules: *face detection, face registration, illumination normalization, face authentication and information fusion*.

The mobile camera first takes a sequence of images for the user and then processes them at the mobile's processor. Face detection then specifies the location of face in the self-taken photos. Face registration then identifies the face by localizing face attributes which are also saved in the database. Illumination normalization is a preprocessing step which is needed to eliminate an illumination causing a variability of the face images. This is done by noise removal techniques. Face verification is then invoked to match the most recent captured image with the one stored in the database. Information fusions is finally called by using different frames (calculating the Mahalanobis distance) to improve the system reliability and performance. A summary of this system is illustrated in Fig. 16.6.

The authentication-based face recognition confirms the physical presence of the user. In addition, it reduces the cost of getting tokens since the mobile's camera is already embedded in the mobile and can be used to capture the human face. However, this method suffers from a number of limitations. The human face changes over time or may get injured. Also, image capturing is subject to different lightning changes [28].

Fingerprint Fingerprint authentication is a method in which a user's fingerprint is scanned by a mobile's fingerprint sensor to check an identity of a mobile's user. This determines features such as pressure, the 3D shape of the contacted finger, ridges and

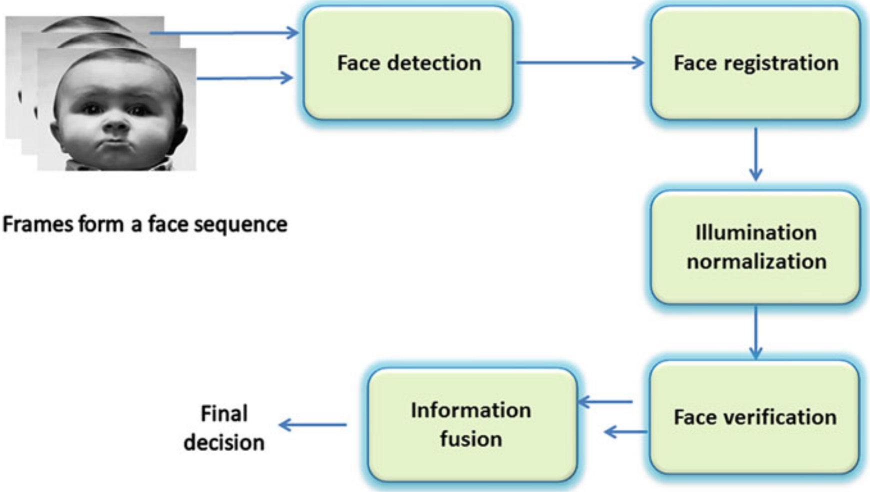


Fig. 16.6 Face recognition system [9]

valleys on the fingertip and other features [29]. Khan et al. [30] proposed two factor authentication system including fingerprint as one factor. This system identifies the user by sensing the ridges and furrows on the user's fingers. This scheme is composed of four phases: registration, login, authentication, and password change phase. In the registration phase, a user submits two types of information: his ID and password, and his fingerprint by the sensor included in the mobile. In the login phase, the user opens the login window to enter his ID and password and then imprints his fingerprint by the sensor. The mobile then verifies the user's fingerprint. If it is valid, the mobile sends the entered ID and password to a remote server. In the authentication phase, the remote server validate the message received. In the password change phase, when the user wants to change his old password, the user has to firstly login with his old password and his fingerprint. Then, the user is allowed to change his password.

As reported in [19], attacking applications with biometric-based authentications shows a much smaller risk comparing with attacking applications with password-based authentications. However, this method is subject to a number of problems to extract the accurate fingerprint information. The human fingerprint is affected by genetic factors, hurtled fingers, and aging. Also the finger reader can not differentiate between the live and the severed finger [31].

Iris recognition This technique depends on scanning a human iris[1] by a separate camera or a mobile's camera [21, 27]. Lee et al. [32] have developed an automated iris recognition system. As shown in Fig. 16.7, this system composes of seven components: Image Acquisition, Segmentation, Normalization, Feature Extraction and Encode, and Similarity Matching Templates.

[1] The iris is the annular part of the eye bounded by the sclera and the pupil.

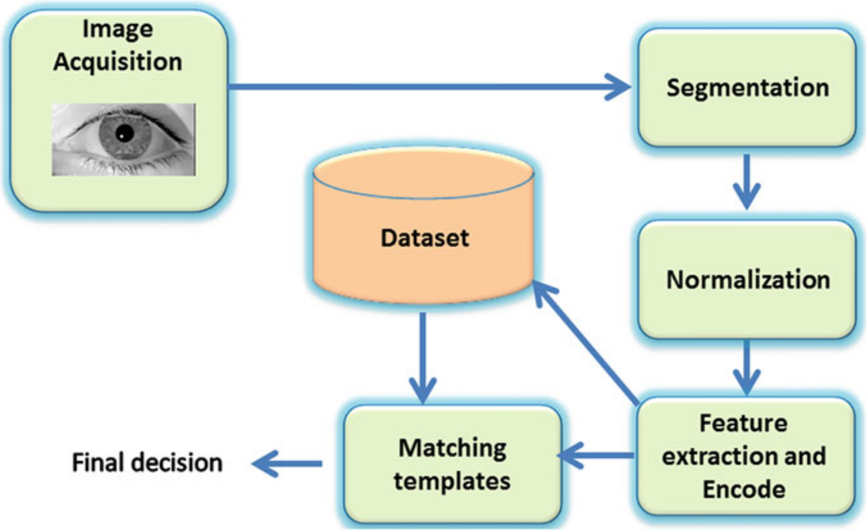


Fig. 16.7 Iris authentication system [32]

In the image acquisition, this system captures a sequence of images from a video frame for the same person using two different cameras at different positions. In segmentation phase, the iris region is isolated from eye images and the image with the best quality is chosen. In normalization phase, 2D representation of the iris pattern is constructed and the noise is removed by masking filters. In Feature Extraction and encode phase, both the edge and line features are extracted from the iris image. Edge features are compared with the intensities of eye regions and the region across the upper cheeks. Line features are compared with the intensities of eye regions and the nose. These features are filtered by classifiers (e.g. Wavelet Transform, Laplacian-of-Gaussian filter, Discrete Cosine Transform, etc.). Then, the iris image is encoded into binary format. In the similarity matching templates, Hamming Distance (HD) is computed between the two iris templates (the existing one and the recent generated one) to decide whether the iris pattern belongs to the same person or not.

Park et al. [33] proposed iris-based authentication system taking an iris image even if a user is wearing glasses by turning on/off the dual (left and right) infra-red (IR) illumination iteratively. Then, the system detects the occluded areas such as the eyelid, eyelash, and corneal specular reflections (SRs) which happen on surface of glasses. To detect the boundaries of the pupil and the iris, the Adaboost classifier was used. This classifier uses a one-step greedy strategy for a sequential learning method. Then, the iris code bits were extracted from the detected areas. The detected iris image was normalized and divided into rectangular polar coordinates (8 tracks and 256 sectors). Finally, the extracted iris code bits were compared to the enrolled template using the hamming distance (HD). If the calculated HD is higher than the specified threshold, the user was accepted. Otherwise user was rejected.

In contrast to face recognition which can be changed over time, the iris is stable. Also, compared to the fingerprint using a sensor which cannot differentiate between the live and the severed finger, the iris's sensor could ensure the live eye as it can measure the depressions and dilations of the pupil [19]. However, this technique takes a quite long time to authenticate a user. In addition, the eye's alignment and any eye's hurt affect the accuracy of users' authentication [16].

In general, we can conclude that the physical biometric authentication techniques suffer from the following problems [14, 29]:

- They require additional hardware (i.e. camera, finger print reader, etc.) which may be already available in mobiles.
- There is a cost for maintenance and the authentication failure.
- There is a high computation done in the poor constrained mobiles.
- A number of biometric identifiers are prone to wear and tear, accidental injuries, and pathophysiological development (accidents, manual work, etc.).
- They are not adaptable to people with disabilities (i.e. blind user can't use face recognition but may use voice biometric).

16.3.6 Behavioral Biometrics

In addition to the authentication techniques which is based on what a user has or knows. There are other techniques which identify users based on what they are usually doing in their own lives. Such techniques are known as *behavioral biometrics*. There are a number of behaviors which can be used to authenticate users. This includes gait, signature, keystroke, etc. Authentication techniques, based on these behaviors, are highlighted below.

Gait Recognition Gait is a behavioral biometric that uses a sequence of video images of a walking person to measure several movements to identify a mobile's user. Typically, shown in Fig. 16.8, gait recognition system consists of five stages: video capture, silhouette segmentation, contour detection, feature extraction and classification. Firstly, a video of a walking person is captured by a camera. Secondly, using some segmentation and motion detection methods, the person is segmented from the surrounding area. Thirdly, the contours of the person are detected to specify the outer boundary of human body [34]. Fourthly, gait features are then extracted. Finally, a classifier is used to identify a person. In the classification, the similarity between the extracted gait feature and the stored ones is computed to identify the walking person.

Derawi et al. [20] have proposed the gait authentication for mobile's user. They have used the low embedded accelerometers found at Google G1 phone. In this system, each volunteer placed a mobile device at his hip. When he walks while wearing his normal shoes, gait data is collected at each four walks (2 walking down and 2 walking back). To identify the person, background segmentation was used to isolate the person from the surrounding background. The first walk was stored as a reference template in a database whereas the others walks were used for extracting

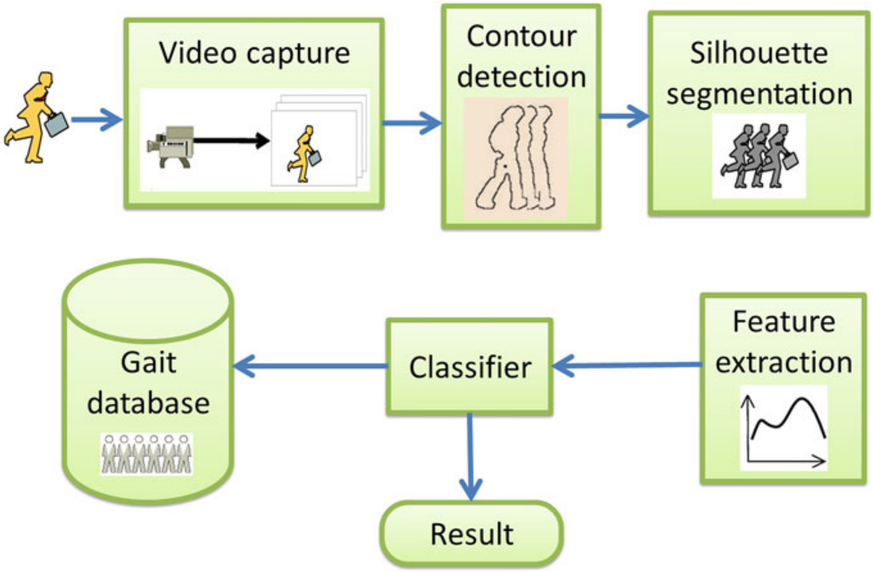


Fig. 16.8 Gait recognition [34]

feature vectors. Features are extracted from each walk, e.g. the acceleration of gait. To identity the mobile's user, Dynamic Time Warping (DTW) is used to compare the extracted feature vectors with the reference ones (enrolled templates). If DTW found a match, then the user is granted the access to the mobile. Otherwise he is rejected.

Compared to other methods, the gait-based authentication enjoys a number of advantages. The gait data is a unique identifier for each person and cannot be shared. Also, none could fake the other's gait. Images building this data can be taken at a distance and it does not require users' involvement [35]. However gait data differs at some cases such as: walking for a long time, injury, weight or footwear changes.

Voice Recognition The voice recognition is an authentication technique in which a user say his password to authenticate himself to his mobile. This technique uses different acoustic features of individuals to authenticate the user. These acoustic patterns reflect both learned behavioral patterns (i.e. voice pitch, speaking style) and anatomy (i.e. shape and size of throat and mouth) [27]. The voice recognition system may be either text-dependent (user speaks a predetermined phrase) or text-independent (user speaks what he/she wants).

Riva et al. [36] developed a progressive authentication system composed of three factors, *face, voice* and *PIN*, to authenticate a user. There are three protection levels: public (access all public applications), private (access both public and private applications) and confidential level (access to all applications). These levels are used to protect important applications against unauthorized use, while providing a way to use the less sensitive applications. The voice recognition of this system depends on Speaker Sense Algorithm. This algorithm uses Gaussian Mixture Model (GMM)

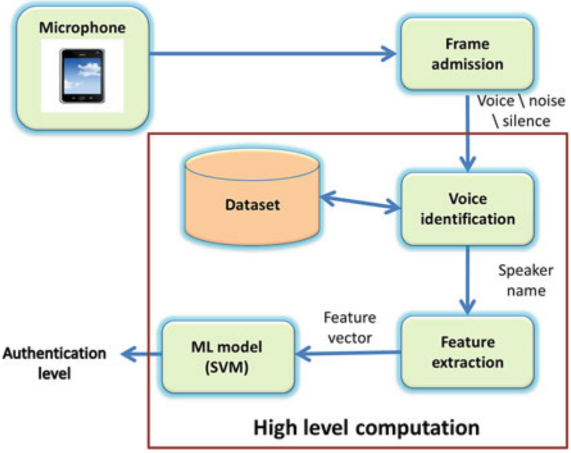


Fig. 16.9 Voice recognition authentication system

classifier to train system by audial recording for 2 min during a phone call. Then, the user's voice is recorded every 20 ms to continuously validate user.

There are two phases for voice recognition: *Frame Admission* and *Voice Identification*. In the frame admission, the recorded sound is analyzed to identify voice, noise and silence frames. In Voice Identification, voice frames are used to recognize the speaker. Voice Identification and high-level processing are done into the cloud (Windows Azure) or a remote device. Then, using the attached mobile's sensors, the system extracts features (e.g. Time elapsed since the last the phone was on the table, or pocket) and then produces a feature vector. This vector is redirected to a machine learning (ML) model to associate a label to the vector. This label maps the user to one of the three protection levels (see above). A summary of this system is demonstrated in Fig. 16.9.

The voice-based technique might be preferable because any mobile already contain a microphone, so no an additional cost is imposed on the users [19]. However, human voice is sensitive to various factors like: aging, noise, medical conditions (such as a common cold) and emotional state, etc. [21]. Such factors affect the accuracy of the authentication results.

Keystroke This technique was developed to enhance the text-based authentication one. The keystroke technique is based on extracting keystroke features (e.g. the time of key holding or intervals between two keystrokes) when a user enters his/her PIN. A typical a keystroke system [37] is composed of fours modules : *data acquisition, feature extraction, matching and scoring*. Firstly, Keystroke dynamics data is collected when a user presses keys. This data is then examined to extract some features forming a biometric pattern. This biometric pattern is then compared to biometric templates enrolled during a training stage. Such comparison produces either a distance or score describing the similarity between the learned pattern and the stored templates. This similarity must exceed a threshold value. If similarity is less than

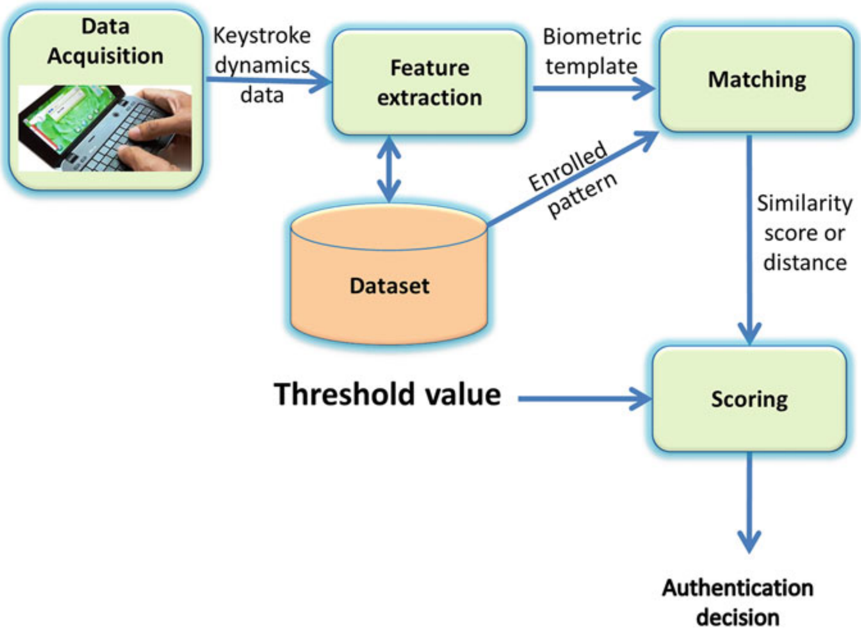


Fig. 16.10 Keystroke system [38]

this threshold, the pattern is rejected. Otherwise, the pattern is accepted and the user is granted the access to the system. This system is described in Fig. 16.10.

Chang et al. [39] proposed an authentication solution using keystroke dynamics besides the pressure feature. This is recorded using graphical-based password for touch screen mobile to enlarge the password space size. While mobile's owner is selecting 3–6 thumbnails into an image in some sequence representing the graphical password, the system is extracting and comparing the keystroke features. These features include pressure and time features such as Down-Up (DU) time, Up-Down (UD) time, Down-Down (DD) time and Up-Up (UU) time. To enroll a user's template in a database, five training samples were needed from each user. To verify the user's identity, a statistical classifier was used. This classifier compares the most recent template with the registered one in the enrollment phase. If they do not match, the system rejects the user's login request. Otherwise, the user is granted an access to the system.

The keystroke technique overcomes the shoulder surfing attack and is done implicitly without disturbing the user. In addition, compared to other biometrics, the keystroke does not add an additional hardware. However, its usability is not good on touchscreen mobiles. This is because the hold time is nothing compared to the normal sized keyboards [40]. This is affected by different keypad sizes or layouts of mobile devices in QWERTY keyboards.

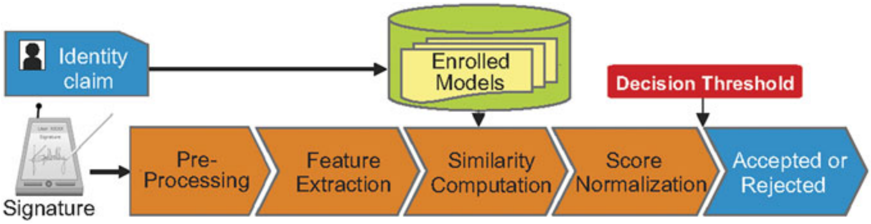


Fig. 16.11 Signature recognition [43]

Signature recognition In this technique, a user signs on a touch-screen of a mobile, then a system analyzes how the user types on the screen [41]. This is done by extracting some features like: time, speed, acceleration, pressure, and direction [42]. A typical signature recognition system is composed of five modules: *pre-processing, feature extraction, enrollment, similarity computation and score normalization*. This system first acquires digitized signals obtained from a touchscreen or a pen movement on a tablet while its user is holding the device. In the preprocessing step, the missing parts of the acquired signals are completed [42].

In the feature extraction, a feature vector, consisting of signature duration or average speed, is generated from each acquired signal. These features are either enrolled as templates or used by a statistical model representing the generated signatures [43]. The similarity computation module matches the claimed identity to the enrolled templates by computing a similarity score. This is done using distance-based classifiers (e.g. Euclidean distance or Mahalanobis distance) or statistical models (e.g. Dynamic Time Warping (DTW) or Hidden Markov Models (HMM)) [43]. To grant an access to the claimed user, a score normalization is used to normalize similarity scores to a given range of values and the compares the produced score with a pre-defined threshold value. This normalization is useful when multiple algorithms are used in a system and scores must be fused for a final decision [43]. A summary of a typical signature system is shown in Fig. 16.11.

Compared to other authentication methods, the signature technique is considered the most common used one in many verification tasks [44]. It has a high user's acceptance [43]. Unlike sensors and cameras used in fingerprint and face recognition, a mobile devices do not require any additional acquisition hardware. However, the signature technique seems different when signing on smart phones or signature pads or pen-based tablets. Signature on the touchscreen devices is less qualified because information about pressure or pen orientation is not available. Also person's signature differs at some cases such as using different style over time or in case of injury, mimicking the owner's signature using the other hand to sign [43].

16.4 Biometric Versus Traditional Authentication

As shown previously, each technique has its strengths and weaknesses. Choosing one of these methods depends on the user needs [29]. No technique is optimal but may satisfy what the user needs. There are a number of dissimilarities between the traditional and biometric authentications. Firstly, the traditional techniques are active that asks user to enter select carry the user credentials, while the biometric one is passive (user transparent) in which user has nothing to type select and also no devices to carry around [45]. Secondly, biometric data are linked to its owner but traditional credentials cannot do, since they can be forgotten or shared or lent or stolen [24]. Thirdly, biometric provides a reliable and natural way for identification because user has to be present at the time of authentication and can't repudiate access to system [29]. However, with traditional techniques users can deny the login by sharing the password. Last but not the least biometric data is fairly unique for each person. At the same time the biometric data is noisy which requires measurements to be accurate and this makes biometric authentication very challenging and emerging [46].

16.5 Comparison Among Authentication Techniques

This section provides a comparison among the various authentication techniques described earlier. This comparison is conducted based on the following metrics:

1. Usability (ease of use): This means that authentication should be fast and as unobtrusive as possible [47]. A determination of *High, Medium or Low* denotes how fast and easy the technique is to user.
2. Cost (need for additional hardware): With the cost here, we mean adding additional cost for a user's authentication, e.g. using camera/sensor to capture some features or from the additional support needed when mobile is blocking the access to users. Such cost should be minimized.

   A determination of *High, Medium or Low* denotes how much cost the technique requires.
3. Performance: This is related to the computational cost and time needed for a user's authentication. This includes the following characteristics:

   (a) Time complexity: This is concerned with decreasing both the calculation speed and the detection latency. The calculation speed is the time needed to build a user's model (i.e. extracting a user's features) and also to grant access to the user. Detection latency is the time consumed to detect an attacker usage of mobile and this must be minimized. It's desirable to increase user actions while decreasing waiting time for user input.

   (b) Minimum Consumption: This means to use the minimum resource requirements. Mobile can be thin client where limited computation and storage is done to minimize power (battery) consumption at mobile phones.

A determination of *High, Medium or Low* denotes how much time and computation the technique consumes.

4. Explicit or Implicit Technique(user direct interaction): This shows whether there is a need for a physical involvement of users during the authentication process or the authentication is done transparently with normal user activity without an explicit action from the users.

   A determination of Explicit or Implicit denotes if technique requires direct user interaction or not.
5. Robustness against any (aural or visual) eavesdropping: Checking whether a system is robust to various fraudulent methods and attacks that could be mounted during an authentication session, e.g. watching or listening a password during login time or selecting pictures that represent s passcode.

   A determination of *Yes or No* denotes whether a technique is robust or not.
6. Circumvention: indicates to whether a technique can detect the change of users. In other words, checking whether an illegitimate user can mimic the legitimate owner's behaviors to grant access to system.

   A determination of *Yes or No* denotes whether an owner can be mimicked or not.
7. Continuous Authentication: This is related to the length of the time during which the authentication is done either only at login time or during the runtime [9].

   A determination of Login time or Runtime denotes when authentication takes place.

Table 16.1 shows a compassion of various authentication techniques described above in relation to the previous mentioned metrics. A determination of High (H), Medium (M) or Low (L) denotes how well the technique adheres the metrics. The individual determinations are based on the authors' opinions and knowledge of techniques.

16.6 Explicit and Implicit Authentication

Traditional or biometric authentication techniques can be done explicitly or implicitly. This depends mainly on if it requires user interaction or not. This means that if it was user intrusive or not. If technique requires an explicit action from user for authentication like putting the finger on a fingerprint scanner, then this technique is considered to be explicit way for authentication. In contrast to this, technique is considered to be implicit way for authentication if it was user transparent (unobtrusive) [9], Effortless as possible and may be continuous authentication. The user transparency means that user deals normally without any explicit action because the relevant data is continuously recorded while the person is walking or writing a message for example. Continuous authentication provides protection goes beyond point-of-entry security. This means that it doesn't depend on only writing password correctly at login time but also at runtime [7].

Table 16.1 Comparison of authentication techniques

Technique	Usability	Cost	Performance	Explicit or implicit technique	Eavesdropping robustness	Circumvention	authentication authentication
PIN or PUK or alpha numeric password	High	Medium	Low	Explicit	No	Yes	Login time
Graphical password	Medium	Low	Medium	Explicit	No	Yes	Login time
Token	Medium	High	Medium	Explicit	No	Yes	Login time
Face recognition	High	Medium	High	Explicit or implicit	Yes	No	Login or runtime
Finger print	High	High	High	Explicit	Yes	Yes	Login time
Iris recognition	Medium	High	High	Explicit or implicit	Yes	No	Login or runtime
Gait recognition	Medium	Medium	High	Implicit	Yes	No	Runtime
Voice recognition	Medium	Low	Medium	Implicit	Yes	Yes	Login or runtime
Keystroke	High	Low	Low	Implicit	Yes	No	Login or runtime
Signature recognition	High	Low	Medium	Implicit	Yes	Yes	Login or runtime

16.7 Open Issues

Authentication based on behavioral biometrics have advantages over physiological ones as the former could be used to support continuous and transparent authentication system. Also, behavioral biometrics do not need any special hardware while collecting behavioral data, thus very cost-effective. Nonetheless, it is very difficult to design behavioral biometric techniques which could suite all users. So, the research should focus on how to propose an authentication system such that providing continuous and transparent authentication while not imposing additional cost for the special hardware. One way to achieve this is by developing multi-modal behavioral biometric authentication systems. In addition, these multi-model systems should be flexible and scalable. For example, a multi-model authentication system could be voice and keystroke or signature based system. Furthermore, this system should have the capability to integrate new biometric techniques while preserving the underling mechanism of the overall system design.

GP must be resistant to shoulder surfing or any eavesdropping while taking less time and effort to login. However, it isn't suitable for blind people or people with weak visions.

16.8 Conclusion

Mobile smart phones are now very important for their users. They aren't only used for communication purposes but also for storing and accessing sensitive data. In the era of cloud computing, the smart phones are a good tool to provide access to data and services on cloud and on the Internet. The first gate to protect the mobile itself and the data stored on it or the services provided by it is the authentication process. Many techniques are being used to support mobile authentications in different environments. This chapter has given an overview on the current mobile authentication mechanisms: traditional and biometric. Based on the user interaction with these mechanisms, a classification has been made. In addition, the chapter has showed the advantages and disadvantages of these mechanisms and it has conducted a comparison between the described techniques. Furthermore, the chapter has highlighted that the behavioral biometric authentication could be promising techniques for mobile authentication as they do not require any special hardware while support a continuous authentication. However, there is no a generic behavioral model to support all users, thus a multi-model (physiological and behavioral) is required to consider for further research in this direction. Before successful deployment of such potential system, great efforts of research and development are still required to investigate all aspects (e.g. power consumption and usability) of the mobile smart phone biometric system.

References

1. Tseng, D., Mudanyali, O., Oztoprak, C., Isikman, S.O., Sencan, I., Yaglidere, O., Ozcan, A.: Lensfree microscopy on a cellphone. Lab Chip **10**(14), 1787–1792 (2010)
2. Wang, H., Liu, J.: Mobile phone based health care technology. Recent Pat. Biomed. Eng. **2**(1), 15–21 (2009)
3. Fudong, L., Nathan, C., Maria, P., Paul, D.: Behaviour profiling on mobile devices. In: International Conference on Emerging Security Technologies (EST), 2010, IEEE (2010), pp. 77–82
4. Vaclav, M.J., Zdenek, R.: Toward reliable user authentication through biometrics. IEEE Secur. Priv. **1**(3), 45–49 (2003)
5. Hanul, S., Niklas, K., Sebastian, M.: Poster: user preferences for biometric authentication methods and graded security on mobile phones. In: Symposium on Usability, Privacy, and Security (SOUPS) (2010)
6. Wazir, Z.K., Mohammed, Y.A., Yang, X.: A graphical password based system for small mobile devices. arXiv preprint arXiv:1110.3844 (2011)
7. Nathan, L.C., Steven, M.F.: Authentication of users on mobile telephones-a survey of attitudes and practices. Comput. Secur. **24**(7), 519–527 (2005)
8. Mohsen, T., Ali, A.B.: Solutions to the gsm security weaknesses. In: The Second International Conference on Next Generation Mobile Applications, Services and Technologies, 2008. NGMAST'08, IEEE (2008), pp. 576–581 (2008)
9. Qian, T., Raymond, V.: Biometric authentication system on mobile personal devices. IEEE Trans. Instrum. Meas. **59**(4), 763–773 (2010)
10. Andrea, K., Valerie, S., Michael, S.: Using publicly known passwords with haptics and biometrics user verification. In: IEEE Haptics Symposium (HAPTICS) 2012, IEEE (2012), pp. 559–562 (2012)
11. Greg, E.B.: Graphical password (September 24 1996) US Patent 5,559,961
12. Haichang, G., Zhongjie, R., Xiuling, C., Xiyang, L., Uwe, A.: A new graphical password scheme resistant to shoulder-surfing. In: International Conference on Cyberworlds (CW) 2010, IEEE (2010), pp. 194–199 (2010)
13. Lawrence, O.: Comparing passwords, tokens, and biometrics for user authentication. Proc. IEEE **91**(12), 2021–2040 (2003)
14. Fadi, A., Syed, Z., Wassim, E.H.: Two factor authentication using mobile phones. In: IEEE/ACS International Conference on Computer Systems and Applications, 2009 (AICCSA 2009) IEEE (2009), pp. 641–644 (2009)
15. Parekh, T., Gawshinde, S., Sharma, M.K.: Token based authentication using mobile phone. In: International Conference on Communication Systems and Network Technologies (CSNT) 2011, IEEE (2011), pp. 85–88 (2011)
16. Clarke, N.L., Furnell, S.: Advanced user authentication for mobile devices. Comput. Secur. **26**(2), 109–119 (2007)
17. Fred, C.: A secure mobile otp token. In: International Conference on Mobile Wireless Middleware, Operating Systems, and Applications, pp. 3–16. Springer (2010)
18. Mohamed, H.E., Muhammad, K.K., Khaled, A., Tai-Hoon, K., Hassan, E.: Mobile one-time passwords: two-factor authentication using mobile phones. Secur. Commun. Netw. **5**(5), 508–516 (2012)
19. Salil, P., Sharath, P., Anil, K.J.: Biometric recognition: security and privacy concerns. IEEE Secur. Priv. **1**(2), 33–42 (2003)
20. Mohammad, O.D., Claudia, N., Patrick, B., Christoph, B.: Unobtrusive user-authentication on mobile phones using biometric gait recognition. In: Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP) 2010, IEEE (2010), pp. 306–311 (2010)
21. Anil, K.J., Arun, R., Salil, P.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 4–20 (2004)
22. Arun, R., Anil, J.: Biometric sensor interoperability: a case study in fingerprints. In: Proceedings of International ECCV Workshop on Biometric Authentication. Springer, pp. 134–145 (2004)

23. Anil, K.J., Patrick, F., Arun, A.R.: Handbook of Biometrics. Springer, New york (2007)
24. Pim, T., Anton, H.M.A., Tom, A.M.K., Geert-Jan, S., Asker, M.B., Raymond, N.J.V.: Practical biometric authentication with template protection. In: Audio-and Video-Based Biometric Person Authentication, pp. 436–446. Springer (2005)
25. Kresimir, D., Mislav, G.: A survey of biometric recognition methods. In: 46th International Symposium Electronics in Marine, 2004. Proceedings Elmar 2004, IEEE (2004), pp. 184–193
26. Patrick, G., Elham, T.: Performance of biometric quality measures. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 531–543 (2007)
27. Vibha, K.R.: Integration of biometric authentication procedure in customer oriented payment system in trusted mobile devices. Int. J. Inf. Technol. **1**(6), 15–25 (2012). doi:10.5121/ijitcs.2011.1602
28. Jakobsson, M., Shi, E., Golle, P., Chow, R.: Implicit authentication for mobile devices. In: Proceedings of the 4th USENIX conference on Hot topics in security, USENIX Association, pp. 9–9 (2009)
29. Umut, U., Sharath, P., Salil, P., Anil, K.J.: Biometric cryptosystems: issues and challenges. Proc. IEEE **92**(6), 948–960 (2004)
30. Muhammad, K.K., Jiashu, Z., Xiaomin, W.: Chaotic hash-based fingerprint biometric remote user authentication scheme on mobile devices. Chaos Solitons Fractals **35**(3), 519–524 (2008)
31. Jakobsson, M.: Mobile Authentication: Problems and Solutions. Springer Publishing Company, Incorporated, New York (2013)
32. Yooyoung, L., Phillips, P.J., Ross, J.M.: An automated video-based system for iris recognition. In: Tistarelli, M., Nixon, M.S. (eds.) Advances in Biometrics, pp. 1160–1169. Springer, Berlin (2009)
33. Park, K.R., Park, H.A., Kang, B.J., Lee, E.C., Jeong, D.S.: A study on iris localization and recognition on mobile phones. EURASIP J. Adv. Signal Process **2008**, Article ID 281943 (2008). doi:10.1155/2008/281943
34. Hamed, N., Ghada, E.T., Eman, M.: A novel feature extraction scheme for human gait recognition. Int. J. Image Graph. **10**(04), 575–587 (2010)
35. Dacheng, T., Xuelong, L., Xindong, W., Stephen, J.M.: General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. **29**(10), 1700–1715 (2007)
36. Oriana, R., Chuan, Q., Karin, S., Dimitrios, L.: Progressive authentication: deciding when to authenticate on mobile phones. In: Proceedings of the 21st USENIX Security Symposium (2012)
37. Shanmugapriya, D., Padmavathi, G.: A survey of biometric keystroke dynamics: approaches, security and challenges. arXiv preprint arXiv:0910.0817 (2009)
38. Carlo, T., Abbas, R., Ilhami, T.: Full-size projection keyboard for handheld devices. Commun. ACM **46**(7), 70–75 (2003)
39. Ting-Yi, C., Cheng-Jung, T., Jyun-Hao, L.: A graphical-based password keystroke dynamic authentication system for touch screen handheld mobile devices. J. Syst. Softw. **85**(5), 1157–1165 (2012)
40. Sevasti, K., Nathan, C.: Keystroke analysis for thumb-based keyboards on mobile devices. In: New Approaches for Security, Privacy and Trust in Complex Environments, pp. 253–263. Springer (2007)
41. Simon, L., Mark, S.: A practical guide to biometric security technology. IT Prof. **3**(1), 27–32 (2001)
42. Marcos, M.D., Julian, F., Javier, G., Javier, O.G.: Towards mobile authentication using dynamic signature verification: useful features and performance evaluation. In: 19th International Conference on Pattern Recognition, 2008. ICPR 2008, IEEE (2008), pp. 1–5
43. Ram, P.K., Julian, F., Javier, G., Marcos, M.D.: Dynamic signature verification on smart phones. In: Highlights on Practical Applications of Agents and Multi-Agent Systems, pp. 213–222. Springer (2013)
44. Anil, K.J., Friederike, D.G., Scott, D.C.: On-line signature verification. Pattern Recognit. **35**(12), 2963–2972 (2002)

45. Roman, V.Y., Venu, G.: Behavioural biometrics: a survey and classification. Int. J. Biometrics **1**(1), 81–113 (2008)
46. Kai, X., Jiankun, H.: Biometric mobile template protection: a composite feature based fingerprint fuzzy vault. In: IEEE International Conference on Communications, 2009. ICC'09, IEEE (2009), pp. 1–5
47. Rene, M., Thomas, K.: Towards usable authentication on mobile phones: an evaluation of speaker and face recognition on off-the-shelf handsets. In: Fourth International Workshop on Security and Privacy in Spontaneous Interaction and Mobile Phone Use (IWSSI/SPMU), Newcastle, UK (2012)

Part IV
Cloud Security and Data Services

Chapter 17
Cloud Services Discovery and Selection: Survey and New Semantic-Based System

Yasmine M. Afify, Ibrahim F. Moawad, Nagwa L. Badr and M. F. Tolba

Abstract With the proliferation of Software-as-a-Service (SaaS) in the cloud environment, it is difficult for users to search for the right service that satisfies all their needs. In addition, services may provide the same functionality but differ in their characteristics or Quality of Service attributes (QoS). In this chapter, we present a comprehensive survey on cloud services discovery and selection research approaches. Based on this survey, a complete system with efficient service description model, discovery, and selection mechanisms is urgently required. Therefore, we propose a semantic-based SaaS publication, discovery, and selection system, which assists the user in finding and choosing the best SaaS service that meets his functional and non-functional requirements. The basic building block of the proposed system is the unified ontology, which combines services domain knowledge, SaaS characteristics, QoS metrics, and real SaaS offers. A hybrid service matchmaking algorithm is introduced based on the proposed unified ontology. It integrates semantic-based meta data and ontology-based matching. Ontology-based matching integrates distance-based and content-based concept similarity measures. The matchmaking algorithm is used in clustering the SaaS offers into functional groups to speed up the matching process. In the selection process, the discovered services are filtered based on their characteristics, and then they are ranked based on their QoS attributes. Case studies, prototypical implementation results, and evaluation are presented to demonstrate the effectiveness of the proposed system

Y. M. Afify (✉), · I. F. Moawad, N. L. Badr · M. F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
e-mail: yasmine.afify@fcis.asu.edu.eg

I. F. Moawad ibrahim
e-mail: ibrahim_moawad@cis.asu.edu.eg

N. L. Badr
e-mail: nagwabadr@cis.asu.edu.eg

M. F. Tolba
e-mail: fahmytolba@gmail.com

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_17, © Springer-Verlag Berlin Heidelberg 2014

With the proliferation of Software-as-a-Service (SaaS) in the cloud environment, it is difficult for users to search for the right service that satisfies all their needs. In addition, services may provide the same functionality but differ in their characteristics or Quality of Service attributes (QoS). In this chapter, we present a comprehensive survey on cloud services discovery and selection research approaches. Based on survey, a complete system with efficient service description model, discovery, and selection mechanisms is urgently required to assist the user in finding and choosing the best SaaS service that meets his functional and non-functional requirements. Therefore, we propose a semantic-based SaaS publication, discovery, and selection system. We developed a unified ontology that combines services domain knowledge, SaaS characteristics, QoS metrics, and real SaaS offers. A hybrid service matchmaking algorithm is introduced based on the proposed unified ontology. It integrates semantic-based metadata and ontology-based matching. Ontological similarity integrates distance-based and content-based concept similarity measures. It is used in clustering the SaaS offers into functional groups to speed up the matching process. Case studies, prototypical implementation results, and evaluation are presented to demonstrate the effectiveness of the proposed system.

17.1 Introduction

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, storage, and applications) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Examples of leading cloud computing providers are Amazon, IBM, Salesforce, and Google [1, 2].

Service discovery is an emerging research area in the domain of cloud computing, which aims to automatically or semi-automatically detect services or service information in cloud computing environment. With the exponential growth of IT companies offering cloud services, the cloud service discovery became a real problem. In particular, cloud providers use different service descriptions and non-standardized naming conventions for their services [3, 4].

Furthermore, it is not sufficient to just discover multiple cloud services, it is also important to evaluate which is the most suitable service for user needs [5]. Cloud services differ from one another in their specification, performance, and several other attributes, which make it a challenge for service users to select the service that best suits their performance requirements within their budget constraints [6]. One reason behind the difficulty faced by the users is that the cloud providers typically publish their service descriptions, pricing policies, and Service Level Agreement (SLA) rules on their websites in various formats [4].

Analysis of the existing research work on cloud services revealed a set of crucial limitations. In order to address some of them, we propose a semantic-based system that governs the publication, discovery, and selection of cloud SaaS services by achieving the following objectives:

- Developing a unified ontology that serves as a repository for real SaaS offers.
- Using semantic approach to unify the service publications.
- Introducing a hybrid cloud service matchmaking algorithm based on the proposed unified ontology.
- Utilizing both the cloud service characteristics and QoS metrics in the selection process.
- Assessing the effectiveness of the proposed system via prototypical implementation.

This chapter is an extended version of our research work presented in [7]. Additions include an extensive survey on recent cloud services discovery and selection research work, a comprehensive introduction to the unified ontology structure, case studies to illustrate the proposed matchmaking algorithm and the characteristics-based filtering module, and new experimental results.

The remainder of this chapter is organized as follows. Section 17.2 surveys the cloud services discovery and selection research work. Section 17.3 presents the proposed system architecture with a detailed description of its components. Section 17.4 presents the proposed hybrid service matchmaking algorithm and a case study. The prototypical implementation details, experimental scenarios, and evaluation are presented in Sect. 17.5. Finally, the conclusion and future work are presented in Sect. 17.6.

17.2 Related Work

In this survey, we classified recent related work into three categories: cloud service discovery, cloud service selection, and cloud service discovery and selection.

17.2.1 Cloud Service Discovery

In this survey, approaches that tackle cloud services discovery problem can be classified into four categories: cloud-focused semantic search [8–11], cloud service annotations [9, 12, 13], semantic QoS-aware service discovery [14, 15], and other efforts [16, 17].

17.2.1.1 Cloud-Focused Semantic Search Approaches

Cloud-focused semantic search for cloud service approaches are presented in [8–11]. Unified business service and cloud ontology with service querying capabilities is proposed in [11]. The unified ontology captures the required business services in an organization and provides the mapping between business functions and the offered services in the cloud landscape. This work has the following limitations: (a) the exact matching between query and the business functions required by the user, and (b) the query representation, which is depicted in SPARQL language, greatly limits the use of the ontology to experienced users only.

The proposed system in this chapter overcomes the limitations of [11] and extends their work with the following features: (a) accept keyword-based user request, which greatly facilitates the use of the system and maximizes the users acceptance, (b) use semantic techniques to expand the user request and the services description, by retrieving key term synonyms from the WordNet [18], (c) exploit the ontology richness by integrating several concept similarity measures to find relevant services with similar functionality, and (d) provide characteristics and QoS-based selection features.

Another cloud-focused semantic search engine for cloud service discovery is proposed in [10], which is called Cloudle. Cloudle is an agent-based search engine that consults cloud ontology for determining the similarities between service specifications and service requirements. Another cloud-based ontology that is used for generic cloud services search is presented in [8]. However, they use SPARQL API for querying the ontology and they use the services attributes only as a search criterion. Finally, a cloud-focused search tool for the retrieval of services from keyword-based searches is proposed in [9]. Their platform has been tested with promising results. A more complete and thorough validation of the system is planned by applying the system to a larger set of services and by using statistical methods for analyzing the results obtained.

17.2.1.2 Cloud Service Annotation Approaches

Cloud discovery using services annotation is presented in [9, 12, 13]. In [12], cloud service annotations are used for semantic-based discovery of relevant cloud services. Their work depends on Web Services Description Language (WSDL) files. Consequently, it cannot be directly applied to SaaS offers that greatly differ in their representation. On the other hand, authors of [13] continued their efforts in semantic services discovery through extracting fragmental semantic data. They extended the Support Vector Machine (SVM)-based text clustering technique in the context of service oriented categorization in a service repository.

17.2.1.3 Semantic-Based QoS-Aware Approaches

Approaches that support semantic-based QoS-aware service discovery are presented in [14, 15]. Ontology-based discovery of cloud virtual units is presented in [14]. However, the architecture applies on IaaS services only. Another approach is presented in [15]. It is deployed as a cloud application to provide behavior-aware and QoS-aware service discovery services. Its efficiency is going to be evaluated upon deploying it into a commercial cloud.

17.2.1.4 Other Approaches

Other efforts include a service concept recommendation system in [16] and centralized service discovery architecture in [17]. A framework for a service concept recommendation system is proposed in [16], for service retrieval in the service ecosystem. The framework is integrated into a semantic service matchmaker in order to enhance the dependability of the semantic service matchmaker in the service ecosystem. In another context, a Service Discovery Architecture (SDA) for cloud computing environment is proposed in [17]. They suggested that setting of two-level service directories can achieve the most suitable massive cloud services quickly. Finally, the performance analyses of the architecture were studied and results showed its validity.

17.2.2 Cloud Service Selection

In this survey, the approaches that tackle cloud service selection problem can be classified into four categories: performance analysis [19–24], Multi-Criteria Decision Making (MCDM) [6, 25–28], recommender systems [29–31], and other efforts [32, 33].

17.2.2.1 Approaches Based On Performance Analysis

First, we present approaches based on objective assessment and quantitative evaluation of available cloud services [20–22]. A CloudCmp framework for cloud providers comparison is presented in [21, 22], which consists of a set of benchmarking tools used to systematically compare the performance and cost of common services offered by cloud providers along interesting dimensions to customers. However, they focused only on comparing the low level performance of cloud services such as CPU and network throughput. In addition, no details were given on how to derive the representation of cloud services used as inputs to the selection process. Another discussion of cloud benchmark testing is presented in [20]. They proposed a new performance measurement method, which considers the types of services executed on a virtual machine (VM) for IaaS clouds. They recommended creating standardized cloud performance measurement VMs used to measure and compare performance between different providers. However, their work can be applied on IaaS clouds only.

In the previous approaches, no subjective aspect is taken into account to reflect the overall performance of a cloud service. In contrast, other works consider the subjective aspects of a cloud service [19, 24]. A framework for reputation-aware software service selection and rating is presented in [19]. A selection algorithm is devised for service recommendation, providing SaaS consumers with the best possible choices based on quality, cost, and trust. An automated rating model is also defined to overcome feedback subjectivity issues. Another use of subjective assessment is [24], which proposed a novel framework for monitoring cloud performance to derive cloud service selection, in which the performance of a cloud service is predicted by users feedback. There is no mechanism to check the reliability of users feedback and objective assessment of cloud service was not considered.

Integrating both objective and subjective aspects is realized in [23], where a novel model of cloud service selection is presented, which aggregates the information from both the cloud users feedback and objective performance benchmark testing of a trusted third party. However, the cloud service characteristics were not considered in the selection process.

17.2.2.2 Multi-Criteria Decision Making Approaches

Some works formulate the service selection process as a Multi-Criteria Decision Making (MCDM) problem [6, 25–28]. The use of Analytic Hierarchy Process (AHP) technique for cloud service selection is proposed in [25, 28]. Another MCDM technique is presented in [26], which uses quality network models to recommend the most suitable SaaS ERP to the user according to his needs. An alternative implementation of the MCDM problem is presented in [27], which uses the ELECTRE methodology to support the selection. On the other hand, authors of [6] studied key MCDM methods for IaaS cloud service selection. Results showed that MCDM techniques are indeed effective for cloud service selection, but different MCDM techniques do not lead to selection of same service. Hence, more work is needed to identify the most effective MCDM method for IaaS cloud selection using an extended dataset and a much broader criteria set. The main drawback in these works is that the service selection is based on a limited number of features, neglecting the cloud service characteristics.

17.2.2.3 Recommender Approaches

Recommender systems are used to solve the cloud services selection problem in [29–31]. A mechanism to automatically recommend cloud storage services for cloud applications is proposed in [30]. However, the proposed schema does not comply with or take into account any standardization efforts proposed as ontologies on the semantic web. Another work proposed the recommendation of trustworthy cloud providers in [29]. A quantitative trust model for cloud computing environment is presented. It uses aggregated recommendations of the trusted user acquaintances. Experiments on real data confirm its feasibility. On the other hand, the recommendation of cloud-based infrastructure services is proposed in [31], which is a declarative decision support system. It automates the mapping of users specified application requirements to cloud service configurations. The system currently stores IaaS configurations information only.

17.2.2.4 Other Approaches

There are other related works that have different contexts [32, 33]. An efficient QoS-aware service selection approach is proposed in [33]. They proposed a novel concept, called QoS uncertainty computing, to model the inherently uncertain of QoS. However, the work is based on typical QoS criteria adopted from web service selection algorithms. In another context, authors of [32] proposed a novel brokerage-based architecture in the cloud, where the cloud broker is responsible for service selection. A cloud service selection algorithm is designed that considered services convergence. They developed efficient service selection algorithms that rank potential service providers and aggregate them if necessary.

17.2.3 Cloud Service Discovery and Selection

In this survey, we give an overview of the recent research work in cloud service discovery and selection [5, 34–36] ordered by their publication date. Through the combination of dynamic attributes, web service WSDL and brokering, authors of [35] successfully created Resources via Web Services framework (RVWS) to offer higher level abstraction of clouds in the form of a new technology. This study made possible the provision of cloud resources publication, discovery, and selection based on dynamic attributes. However, they did not consider user personalization and there is no concrete selection approach proposed in this work.

A semantic service search engine is proposed in [34]. Apart from a novel search model, they also provide a QoS-based service evaluation and ranking methodology based on provider reputation and other evaluation criteria. To address the defect of low recall rate that appeared in the experiment, authors plan to modify their matchmaking algorithm to obtain better performance.

An OWL-S based semantic cloud service discovery and selection system is proposed in [36]. A novel dynamic service matchmaking method is proposed, where service offers and goals are described with complex constraints. However, neither the service QoS metrics nor characteristics were considered in the selection process.

Another important contribution is the SMICloud framework proposed in [5]. The SMICloud is a hierarchical framework that partitions the description of a service into categories with each category being further refined to a set of attributes. They also proposed an AHP ranking mechanism for cloud services based on QoS requirements. However, no concrete matching approach is proposed, their metrics focus on quantifiable metrics in context of IaaS.

17.2.4 Summary of Key Findings

Based on this survey, a set of interesting findings can be highlighted as follows:

- The incompatibility and lack of standardization in cloud services publication [4]
- Existing work on cloud services discovery mainly focuses on IaaS services. Despite the evidential popularity of SaaS services, SaaS services discovery has not received its eligible research attention yet.
- Existing work on cloud services selection focus solely on QoS attributes (except [8]) neglecting other important cloud service characteristics, which have a great influence on the selection process.
- There is a need for a detailed ontology for each cloud computing layer [37].
- A complete system that employs efficient service description models, discovery, and selection mechanisms is urgently required to assist the user find and choose the best SaaS service that meets his functional and non-functional requirements.

In this research work, we propose a new semantic-based SaaS publication, discovery, and selection system that overcomes some of existing research work drawbacks and limitations in order to close the gap between SaaS requests and real offers.

17.3 The Proposed System

17.3.1 System Architecture

In addition to the web-based user interface and the unified ontology, the proposed system is composed of three main sub-systems. The first sub-system is the service registration sub-system, where SaaS providers directly register their services. The second sub-system is the service discovery sub-system, where the user enters his request (or refined request) in the form of a keyword-based query, and hence the matching services are retrieved as a result of the request. The third sub-system is the service non-functional selection sub-system, where the discovered services are filtered based on their characteristics, and then they are ranked based on their QoS attributes promised in SLA. The ranked services are then displayed to the user. The proposed system architecture is shown in Fig. 17.1.

17.3.2 Unified Ontology

The proposed unified ontology merges knowledge about services domain, SaaS service characteristics, and QoS metrics in addition to real offers. This ontology defines the domain model for the SaaS layer and serves as a semantic-based repository across the service publication, discovery, and selection processes.

The main contribution in this respect is the identification and collection of the most important concepts in the SaaS services domain and their definition in the unified ontology. The required services domain knowledge was collected from multiple resources: Business Function Ontology (BFO) framework [38], cloud ontologies [39–43], and the Wikipedia [44]. In addition, established industry classification standards have been used as a guiding reference: United Nations Standard Products and

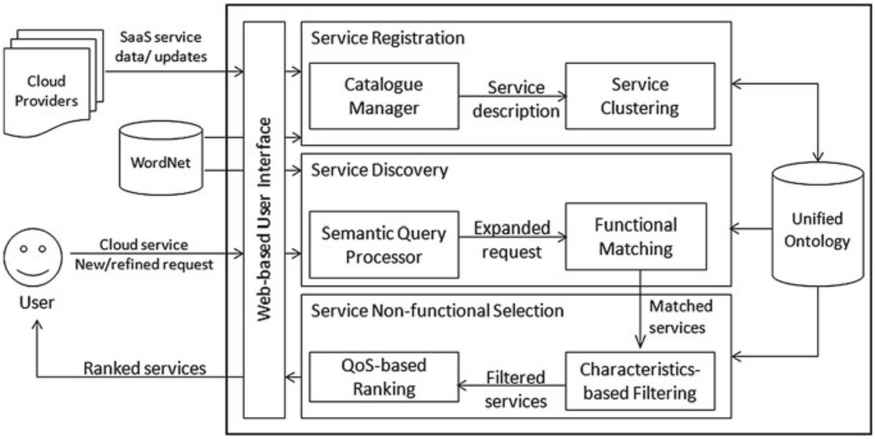


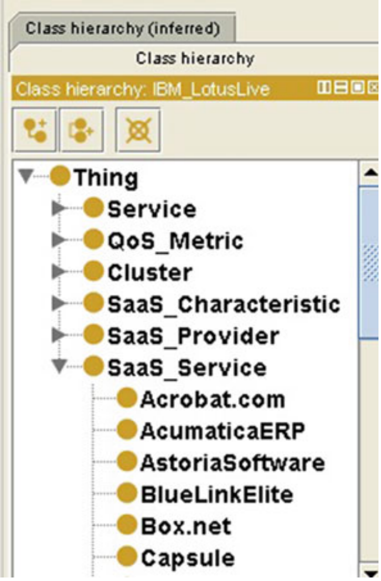
Fig. 17.1 Proposed system architecture

Services Code (UNSPSC) [45] and North American Industry Classification System (NAICS) [46]). The unified ontology currently consists of 650 concepts that represent the domain knowledge for four SaaS application domains: Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), Collaboration and Document Management (DM). These domains were chosen due to their popularity in the cloud market.

Another contribution in this respect is modeling real service offers published by the cloud providers according to the developed ontology. Cloud service offers were collected through a market research by manually visiting the cloud provider portals. The ontology was represented in the well-known knowledge representation Web Ontology Language (OWL) [47]. OWL has been widely accepted and supported by the research community. It is based on Description Logics (DL), which is a family of logic-based knowledge representation formalisms [47]. Protege 4.1 ontology editor [48] was used to implement the ontology.

Figure 17.2 shows the main concepts of the unified ontology. The *Service* concept is the root of the SaaS services domain ontology. The *QoS_Metric* concept contains SaaS QoS metric concepts. The *Cluster* concept contains the generated service cluster concepts. The *SaaS_Characteristic* concept contains cloud service characteristic concepts. The *SaaS_Provider* concept contains information about real SaaS providers (e.g. *Oracle* and *Microsoft*). Finally, the *SaaS_Service* concept contains information about the real services (e.g. *Acrobat.com* and *Box.net*). Figure 17.3 shows a snapshot of the services domain ontology. Business functions from the SaaS services domain are represented as concepts (e.g. *Payroll* and *Accounting*).

Object properties are used for describing the relationship between classes in the ontology. The semantic reasoner uses these specifications for classifying different classes of the ontology hierarchy. For the unified ontology, the following object properties have been defined. The *supportsBusinessFunction* property is used to



Class hierarchy (inferred)

Class hierarchy

Class hierarchy: IBM_LotusLive

- Thing
  - Service
  - QoS_Metric
  - Cluster
  - SaaS_Characteristic
  - SaaS_Provider
  - SaaS_Service
    - Acrobat.com
    - AcumaticaERP
    - AstoriaSoftware
    - BlueLinkElite
    - Box.net
    - Capsule

Fig. 17.2 Unified ontology: main concepts

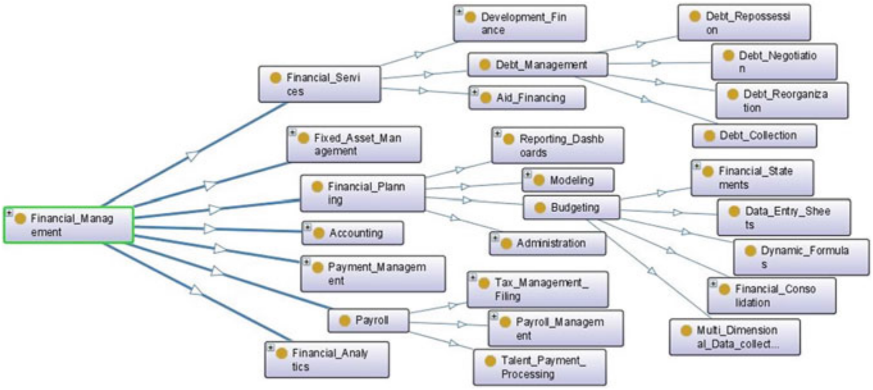


Fig. 17.3 Unified ontology: services domain ontology view

connect the service concepts and SaaS service domain concepts, they describe the business processes supported by each service. The *isProvidedBy* property is used to connect each service to its provider. The *belongsToCluster* property is used to connect each service to the generated cluster it belongs to.

The cloud service characteristics [3] are represented as concepts as shown in Fig. 17.4. The value partitions design pattern [48] is used in order to restrict the range of possible values for each characteristic to an exhaustive list, e.g. the range of possible values for the *PaymentSystem* characteristic is *Free* or *Dynamic* or *PayPerUse*.

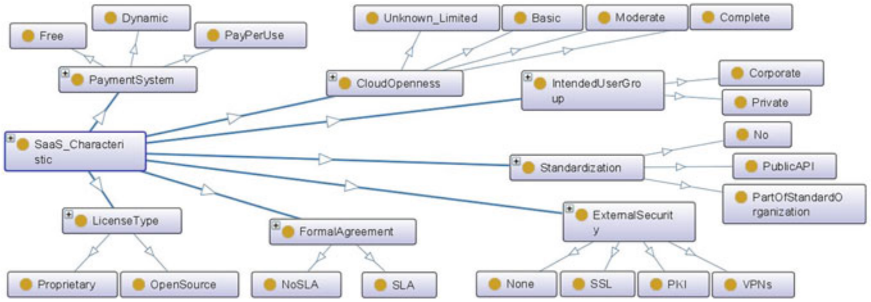


Fig. 17.4 Unified ontology: service characteristics view

The characteristics of real SaaS offers are described using object properties. For each object property, we set it to *functional* and its range to its related cloud characteristic. For example, the object property *hasPaymentSystem* is functional and its range is set to concept *PaymentSystem*. To specify that the cloud service*37Signals* follows the *PayPerUse* payment system, we use the following object restriction: *hasPaymentSystem some PayPerUse* on the *37Signals* service concept. Other object properties defined include: *hasCloudOpenness, hasExternalSecurity, hasFormalAgreement, hasIntendedUserGroup, hasLicenseType*, and *hasStandardization*.

Service QoS metrics are represented as concepts. The service QoS values guaranteed by the service providers are described using data properties. Some of the data properties defined include: *hasAdaptability, hasAvailability*, and *hasReliability*. Data properties range can be of any of the supported data types e.g. integer, string, boolean, etc. For example, in order to specify that the price of the most popular plan of the cloud service *37Signals* is *49* in the unified ontology, we use the following data property assertion: *hasServiceUseCost 49* on the *37Signals* individual.

17.3.3 Service Registration Sub-system

This sub-system is composed of two modules: the catalogue manager and the service clustering modules.

17.3.3.1 Catalogue Manager Module

In this module, cloud service providers register their services through a user-friendly web-based interface based on predefined parameters: service provider name, service name, description, URL, application domain, price/month, and characteristics. Then, the cloud providers map the service features to ontological concepts retrieved from the unified ontology.

The catalogue manager is responsible for the pre-processing stage [49] of the service description. Firstly, it applies tokenization to break the service description into tokens. Secondly, it applies stop words removal to eliminate the common words irrelevant to the service operation. Thirdly, it applies stemming to obtain the root form of service description tokens. Pre-processing aims at the unification of the service descriptions before the matching process. The catalogue manager is also responsible for accepting updates of the registered services.

In order to address the problem of non-standardized naming conventions [3, 4] in the services description, the WordNet ontology is consulted to expand the service description. The catalogue manager sends service descritpion key terms to the WordNet API and receives term synonyms. The semantically enriched service description is then stored in the unified ontology.

17.3.3.2 Service Clustering Module

This module is responsible for clustering the service offers based on their similar functionalities in order to expedite the retrieval of the most relevant SaaS services. The Agglomerative Hierarchical Clustering (AHC) [49] approach is used, which starts with each service in a separate cluster and recursively merges two or more of the most similar clusters. A hybrid service matchmaking algorithm is introduced to measure the similarity between two services. If the resulted similarity is above a threshold value, then they belong to the same cluster. The appropriateness of merging two clusters depends on the similarity of the elements of the clusters. To compare two clusters containing many services, the average inter-similarity [49] between services of the two clusters is computed. The process continues until clusters do not change for two successive iterations. The average inter-similarity [49] between clusters $c_1 = \{s_1, s_2, \ldots, s_k\}$ and $c_2 = \{ss_1, ss_2, \ldots, ss_m\}$ is calculated using Eq. (17.1):

$$sim(c_1, c_2) = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} sim(s_i, ss_j) \tag{17.1}$$

After clustering is accomplished, a cluster signature vector is created for each cluster that contains key terms including the business functions that best describe services in this cluster. The cluster signature vector will be used later in the functional matching process.

17.3.4 Service Discovery Sub-system

This sub-system is composed of two modules: the semantic query processor and functional matching modules.

17.3.4.1 Semantic Query Processor Module

In this module, the user enters his/her keyword-based query using a web-based interface. Similar to the service description, the semantic query processor pre-processes the user query. Finally, in order to improve the recall of the proposed system, the request is expanded using its token synonyms from the WordNet ontology. The expanded user request is then passed to the functional matching module.

17.3.4.2 Functional Matching Module

The functional matching module is responsible for matching the expanded user query against the cluster signature vectors in order to find the cluster that best matches the user requirements. The Vector Space Model (VSM) algebraic model [49] is exploited to present user query and cluster signature vectors. To calculate the term weights, a common term importance indicator is used, which is Term Frequency-Inverse Document Frequency (TF-IDF) model [49]. Weight vector for cluster c is: $[w_{1,c}, w_{2,c}, \dots, w_{N,c}]^T$ and calculated using Eq. (17.2):

$$w_{t,c} = tf_{t,c}. \log \frac{|C|}{1+|\{c' \in C | t \in c'\}|} \tag{17.2}$$

where $tf_{t,c}$ is a local parameter that represents the count of term t in cluster c, $|C|$ is the total number of clusters, $|\{c' \in C | t \in c'\}|$ is the number of clusters that contain the term t and $\log \frac{|C|}{1+|\{c' \in C | t \in c'\}|}$ is a global parameter that measures whether the term is common or rare across all cluster vectors. The similarity between the cluster signature vector c and the user query q can be calculated using the cosine similarity of their vector representation [49] using Eq. (17.3), where N represents the number of terms.

$$sim(c, q) = \cos\theta = \frac{\sum_{i=1}^{N} w_{i,c}.w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,c}^2}\sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \tag{17.3}$$

Services that belong to the cluster with the maximum similarity are retrieved to be processed by the selection sub-system.

17.3.5 Service Non-functional Selection Sub-system

Several cloud taxonomies [1, 3, 50, 51] describe the common cloud service characteristics. Existing research work [5, 6, 25–28, 31, 33]—except [8]—focus on QoS-based selection only and neglects the other cloud service characteristics. To extend the existing work, both characteristics and QoS metrics of SaaS cloud services are employed in the selection phase. This sub-system is composed of two modules: characteristics-based filtering and QoS-based ranking modules.

17.3.5.1 Characteristics-Based Filtering Module

In this module, the discovered services are filtered according to the characteristics that the user is interested in. We have a set of k services $K = \{s_1, s_2 \ldots, s_k\}$ resulting from the discovery process, where $k > 1$, and a set of n characteristics $C = \{c_1, c_2 \ldots, c_n\}$ selected by the user as a base for non-functional matching, where $n >= 1$. The service characteristic values form the following $k.n$ matrix V, where $v_{i,j}$ represents the value of characteristic j for service i.

$$V = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k,1} & v_{k,2} & \cdots & v_{k,n} \end{bmatrix}$$

The user specifies his required service characteristic values $R = \{r_1, r_2 \ldots, r_n\}$ from a predefined set extracted from the unified ontology. Since users may have their own preferences, the relative importance of characteristics may differ for each user. Consequently, the user enters his priority weights $W = \{w_1, w_2 \ldots, w_n\}$ for the selected characteristics such that the weights sum up to 1.

The characteristics-based filtering involves comparing the user required values R to the service characteristic values matrix V to filter the discovered services. In the proposed system, the cloud service characteristics single value types are considered. Single value types include string-based (e.g. license and intended user group) and enumeration (e.g. openness and formal agreement). The relative ranking [52] of the user required value against the service characteristic value is represented by R^c/V_i^c. In the case of string-based characteristic, two cases exist. First, the required value is the same as the service value. Second, either different values or the service value is unknown for this characteristic. The relative ranking value is calculated using Eq. (17.4):

$$R^c/V_i^c = \begin{cases} w_c & \text{if } v_{i,c} = r_{i,c} \\ 0 & \text{otherwise} \end{cases} \tag{17.4}$$

In the case of enumeration typed characteristic, the values are positively ordered i.e. values with higher position are evaluated as better than the other. The relative ranking value is calculated using Eq. (17.5):

$$R^c/V_i^c = \begin{cases} w_c & \text{if } \mathrm{pos}(v_{i,c}) >= \mathrm{pos}(r_{i,c}) \\ w_c.\frac{\mathrm{pos}(v_i,c)}{\mathrm{pos}(r_{i,c})} & \text{otherwise} \end{cases} \tag{17.5}$$

The characteristic-based matching values are calculated using Eq. (17.6). The results are displayed to the user. Results include each service and its matching value to the user preferences. Results are sorted in descending order according to the matching value. Consequently, the user specifies x, which represents the number of services to be ranked, where $1 < x <= k$. Finally, the best x services are promoted to the QoS-based ranking module.

$$max_{x<=k}(\forall_{1<i<=k} \sum_{c=1}^{n} R^c/V_i^c) \tag{17.6}$$

For example, if the user chooses two characteristics (license and openness) for the characteristics-based matching process, i.e. $n = 2$. He enters his preferences $R = \{opensource, complete\}$ and weights $W = \{0.6, 0.4\}$. Assuming 3 services resulted from the discovery process, i.e. $k = 3$. The service characteristic values are extracted from ontology and form following matrix V.

$$V = \begin{bmatrix} opensource & basic \\ proprietary & \\ proprietary & complete \end{bmatrix}$$

The calculated characteristics-based matching values for the discovered services are 0.7, 0.4, and 0 for s_1, s_3, and s_2 respectively. Whereas, if the user entered weights $W = \{0.4, 0.6\}$, the results would be 0.6, 0.55, and 0 for s_3, s_1, and s_2 respectively. If the user chooses $x = 2$, then two services s_3 and s_1 will be promoted to the QoS-based ranking module.

17.3.5.2 QoS-Based Ranking Module

The cloud services have several QoS attributes, all of which are the criteria that have to be taken into account when making a service selection decision. This is a Multi-Criteria Decision-Making (MCDM) problem [53] that involves multiple criteria with interdependent relationship. In this module, the Analytical Hierarchy Process (AHP) [53] is used to assign weights to QoS attributes considering the interdependence between them, thus providing a quantitative basis for the ranking of discovered services with matching characteristics.

In order to choose the best metrics upon which SaaS can be compared, a survey was conducted on existing quality metrics in recent research work and international standards [5, 6, 25–28, 31, 33, 54]. We selectively extracted 20 metrics as our evaluation model. The AHP SaaS ranking problem is modeled in Fig. 17.5. The hierarchy represents the selection parameter levels (QoS metrics) only, not the alternatives (service offers). Finally, the ranked services are displayed to the user.

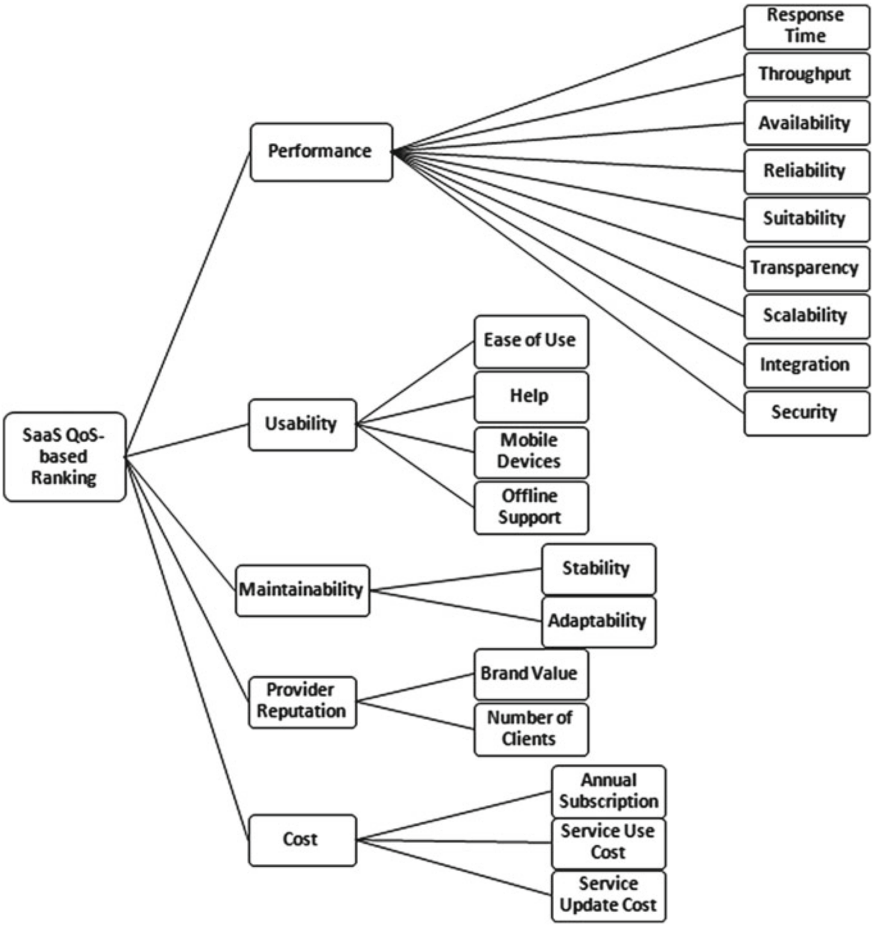


Fig. 17.5 AHP hierarchy for SaaS QoS-based ranking

17.4 Hybrid Service Matchmaking Algorithm

In order to cluster cloud SaaS services that provide similar functionality, a hybrid matchmaking algorithm is proposed that makes use of both semantic-based services metadata and ontology-based matching. The ontology-based matching exploits both concept features and hierarchical structure. The ontological hierarchical structure takes into account both distance and content similarity models [55]. Distance similarity model depends on the ontology taxonomy structure while content similarity model depends on the amount of shared information between two nodes. The algorithm is explained, using line numbers, as follows.

Line 2: Semantic-based Matching

Using Eq. (17.3), VSM is used to calculate the semantic similarity between the expanded service descriptions resulting in $sim_S(s_1, s_2)$.

Lines 3–27: Ontology-based Matching

The novelty of this work is to make use of the richness of the ontology concepts represented in the object properties in the context of cloud SaaS service discovery. Service features are modeled using object properties that relate service concept to business function concepts. Ontology-based similarity matching comprises features and hierarchical similarity matching.

Line 3: Features Similarity

Features similarity denotes common features provided by two services; i.e. the greater the number of common object properties, the more similarity between the two services. For example, for two services s_1 and s_2, the features similarity is calculated using Eq. (17.7):

$$sim_f(s_1, s_2) = \frac{|obj_{s1} \cap obj_{s2}|}{min(x,y)} \tag{17.7}$$

where $|obj_{s1} \cap obj_{s2}|$ represents the number of common object properties in two services, x is the number of object properties of s_1 and y is the number of object properties of s_2. The denominator is used for normalization to ensure that similarity value lies between 0 and 1.

Lines 4–27: Hierarchical Similarity

Hierarchical similarity measures are utilized to find any ontological relationship between the different business functions supported by the two services. In the previous example, if we have n unique business functions for s_1 and m unique business functions for s_2. We have $n.m$ comparisons to process. For each comparison, the semantic similarity between the two concepts is calculated using one of the following three cases:

Case 1: If two concepts have a child-parent relationship, then they are considered to have the maximum similarity, i.e. 1.

Case 2: If two concepts are siblings, then the distance similarity is irrelevant. Consequently, the content-based similarity is adopted to account for amount of shared information between two concepts. The content-based similarity calculates the specificity of concepts by measuring the information content (IC), which is higher for more specific concepts. The LIN measure [56] is calculated using Eq. (17.8):

$$sim_{LIN}(c_1, c_2) = \frac{2 * IC(lcs(c_1,c_2))}{IC(c_1) + IC(c_2)} \tag{17.8}$$

Lin measures the ratio of the IC of the Lowest Common Subsumer (LCS) to the IC of each of the concepts. In the proposed algorithm, the computation model proposed in [57] is used. The concepts IC is computed as the ratio between its degree of generality (expressed by the number of leaves) with respect to its degree of concreteness (expressed by the amount of taxonomical subsumers), it is calculated using Eq. (17.9):

$$IC(c) = -\log\left(\frac{\frac{|leaves(c)|}{|subsumers(c)|} + 1}{maxleaves + 1}\right) \tag{17.9}$$

where $leaves(c)$ corresponds to the number of leaves of concept c, $subsumers(c)$ is the relative depth of concept c represented by the number of its taxonomical subsumers in the ontology, and $maxleaves$ is the number of leaves corresponding to the root node of the hierarchy, which acts as a normalizing factor.

Case 3: Otherwise, the distance-based and content-based similarity models are integrated. First, the distance-based similarity calculates the minimum path of edges between the two concepts. In case of multiple inheritance, the LCS is chosen with the shortest distance between the two concepts. The Resnik semantic similarity measure [58] is calculated using Eq. (17.10):

$$sim_{edge}(c_1, c_2) = 2 * D - min(len(c_1, c_2)) \tag{17.10}$$

where D is the maximum depth of the ontology. Second, the content-based similarity is calculated using Eq. (17.8). Finally, the two ontological measures are integrated using Eq. (17.11):

$$sim_{ontH}(c_1, c_2) = (sim_{edge}(c_1, c_2) + sim_{LIN}(c_1, c_2))/2 \tag{17.11}$$

The ontological services similarity between s_1 and s_2 is calculated by computing the average maximum inter-similarity between concepts of the two services using Eq. (17.12). First, each concept from s_1 is compared with all concepts of s_2 and the maximum similarity value is taken, and then repeats for all s_1 concepts. Second, the average of the n comparisons is calculated using Eq. (17.12):

$$sim_{ontH}(s_1, s_2) = \frac{\sum_{i=1}^{n} max_{1<j<=m} sim_{ontH}(c_i, c_j)}{n} \tag{17.12}$$

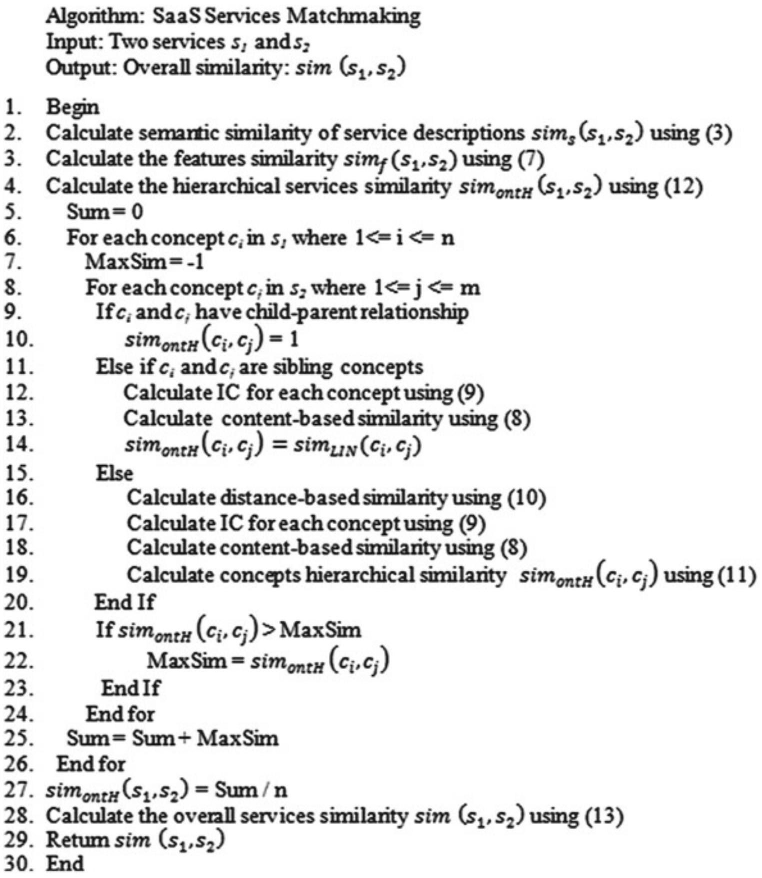
Lines 28–30: Overall Services Similarity

The overall similarity between the two services is calculated by taking weighted average of above similarity measures using Eq. (17.13):

$$sim(s_1, s_2) = a.sim_f(s_1, s_2) + b.sim_S(s_1, s_2) + c.sim_{ontH}(s_1, s_2) \tag{17.13}$$

where a, b and c are weights that reflect the importance of each similarity measure, where the weights sum up to 1. It is important to determine proper values of a, b and c to reach the optimum performance of the system, which is investigated in our experiment. Figure 17.6 presents the pseudo-code of the proposed SaaS services matchmaking algorithm.

To better illustrate the proposed SaaS matchmaking algorithm, we will introduce an example that computes the ontological similarity between two services $s_1$and s_2. We will assume the two services support some of the business functions shown in Fig. 17.3. The features supported by s_1 are {Accounting, Debt Collection, Payroll, and Reporting Dashboards}, while features supported by s_2 are {Modeling, Accounting,



```
Algorithm: SaaS Services Matchmaking
Input: Two services $s_1$ and $s_2$
Output: Overall similarity: $sim\ (s_1, s_2)$

1.  Begin
2.  Calculate semantic similarity of service descriptions $sim_s(s_1,s_2)$ using (3)
3.  Calculate the features similarity $sim_f(s_1,s_2)$ using (7)
4.  Calculate the hierarchical services similarity $sim_{ontH}(s_1,s_2)$ using (12)
5.    Sum = 0
6.    For each concept $c_i$ in $s_1$ where 1<= i <= n
7.      MaxSim = -1
8.      For each concept $c_j$ in $s_2$ where 1<= j <= m
9.        If $c_i$ and $c_j$ have child-parent relationship
10.           $sim_{ontH}(c_i, c_j) = 1$
11.       Else if $c_i$ and $c_j$ are sibling concepts
12.           Calculate IC for each concept using (9)
13.           Calculate content-based similarity using (8)
14.           $sim_{ontH}(c_i, c_j) = sim_{LIN}(c_i, c_j)$
15.       Else
16.           Calculate distance-based similarity using (10)
17.           Calculate IC for each concept using (9)
18.           Calculate content-based similarity using (8)
19.           Calculate concepts hierarchical similarity $sim_{ontH}(c_i, c_j)$ using (11)
20.       End If
21.       If $sim_{ontH}(c_i, c_j)$ > MaxSim
22.             MaxSim = $sim_{ontH}(c_i, c_j)$
23.        End If
24.     End for
25.   Sum = Sum + MaxSim
26.  End for
27. $sim_{ontH}(s_1,s_2)$ = Sum / n
28. Calculate the overall services similarity $sim\ (s_1, s_2)$ using (13)
29. Return $sim\ (s_1,s_2)$
30. End
```

Fig. 17.6 Proposed SaaS cloud services matchmaking algorithm

Development Finance, Financial Statements, and Payroll Management}. The features similarity is calculated using Eq. (17.7) as follows: $sim_f(s_1, s_2) = 1/4 = 0.25$

The hierarchical similarity reflects the ontological similarity among the unique business functions supported by the two services. We have $n = 3$ unique business functions for s_1 and $m = 4$ unique business functions for s_2. Therefore, we have 12 comparisons to process. For each comparison, the semantic similarity between the two concepts c_1 and c_2 is calculated using one of the three cases (lines 9–20). Table 17.1 shows the detailed execution of the hierarchical similarity calculation $sim_{ontH}(s_1, s_2)$.

Table 17.1 Example on hierarchical similarity calcuation

No	C_1	C_2	Case	Sim_{LIN}	Sim_{edge}	Sim_{ontH}	$Max_{1<j<=4}$
				(c_1, c_2)	(c_1, c_2)	(c_1, c_2)	$sim_{ontH}(c_1, c_j)$
1	Debt Collection	Modeling	3	0.23	0.68	0.46	0.72
2		Development finance	3	0.64	0.81	0.72	
3		Financial statements	3	0.23	0.62	0.42	
4		Payroll management	3	0.25	0.68	0.47	
5	Reporting Dashboards	Modeling	2	0.66	–	0.66	0.73
6		Development finance	3	0.26	0.75	0.5	
7		Financial statements	3	0.66	0.81	0.73	
8		Payroll management	3	0.29	0.75	0.52	
9	Payroll	Modeling	3	0.31	0.81	0.56	1
10		Development finance	3	0.31	0.81	0.56	
11		Financial statements	3	0.31	0.75	0.53	
12		Payroll management	1	–	–	1	
$sim_{ontH}(s_1, s_2) = (0.72 + 0.73 + 1)/3 = 0.81$							

17.5 Experimentation and Evaluation

Information Retrieval (IR) performance measures [49] are used to experimentally evaluate the proposed system performance. The experimental evaluation is described in terms of experimental setup, evaluation metrics, experimental scenarios, and results.

17.5.1 Experimental Setup

Experiments were conducted on an Intel Core i3 2.13 GHz processor, 5.0 GB RAM running under Windows 7 Ultimate. The system was built from scratch using Java, Jena API, and WordNet API in Eclipse IDE.

We built a data set of 500 SaaS service offers, amongst these, 26 services are live services from the Internet and the remaining are pseudo services generated by adapting the real services with some changes. Pseudo services were created on purpose, which exhibit certain characteristics in order to test the rigidity of the matchmaking algorithm. Real cloud SaaS offers were collected from the top 10 cloud providers portals listed in [2].

17.5.2 Evaluation Metrics

We used the three most widely used performance measures from the Information Retrieval (IR) field [49]: precision, recall, and F-Measure.

Precision: Precision of one cluster is calculated using Eq. (17.14). The total precision for the system is calculated using Eq. (17.15), where n represents the number of clusters.

$$precision(cluster_i) = \frac{numofcorrectclusteredservices}{numberofclusteredservices} \tag{17.14}$$

$$Precision = \frac{\sum_{i=1}^{n} precision(cluster_i)}{n} \tag{17.15}$$

Recall: Recall of one cluster is calculated using Eq. (17.16). The total recall for the system is calculated using Eq. (17.17), where n represents the number of clusters.

$$recall(cluster_i) = \frac{numofcorrectclusteredservices}{numberofexpectedclusterservices} \tag{17.16}$$

$$Recall = \frac{\sum_{i=1}^{n} recall(cluster_i)}{n} \tag{17.17}$$

The F-measure using combines precision and recall to reflect the user preference by setting the non-negative real weight β. We used $\beta = 2$.

$$F_\beta = (1 + \beta^2).\frac{Precision.Recall}{\beta^2.Precision.Recall} \tag{17.18}$$

17.5.3 Experimental Scenarios

Three experiments were conducted to achieve our objective. In experiment 1, data analysis was conducted to determine: (1) the optimum values of services matchmaking similarity weights and (2) the optimum clustering threshold value. Findings from this experiment were used as the base for the other two experiments. The objective of experiment 2 is to investigate the relevance and overhead of semantically enriching the service description during the service registration. Finally, the objective of experiment 3 is to assess the significance and consequence of clustering the service offers into functionally similar clusters.

Table 17.2 Services data analysis

Services Domain	Services	Sim_f	Sim_S	Sim_{ontH}
Same domain	Oracle CRM On demand versus intouchcrm	0.37	0.54	0.77
	Oracle CRM On demand versus intouchcrm	0.37	0.54	0.77
	Blue link elite versus NetSuite	0.4	0.16	0.82
	IBM lotus live versus CubeTree	0.33	0.31	0.75
	HyperOffice verus GetDropBox	0.66	0.21	0.67
	OrderHarmony versus Plex online	0.2	0.01	0.62
	Incipi workspace versus HyperOffice	0.16	0.19	0.79
Different domains	37 Signals highrise versus Box.net	0.0	0.22	0.3
	DocLanding versus Acumatica ERP	0.0	0.19	0.31
	Incipi workspace versus intouchcrm	0.0	0.14	0.34
	SpringCM versus salesForce.com	0.0	0.24	0.28

17.5.4 Experimental Evaluation

17.5.4.1 Experiment 1

The objective of this experiment is to apply data analysis to derive: (1) the optimum values of matchmaking similarity weights and (2) the optimum clustering threshold value. The services matchmaking algorithm integrates different similarity measures. The first step is to judge the relative effect of each similarity measure solely in order to determine their optimum weight values. The similarity between the services is computed using three scenarios: (1) features similarity only Sim_f, (2) semantic similarity of expanded services description only Sim_S, and (3) hierarchical similarity only Sim_{ontH}. Part of the results is shown in Table 17.2.

After analyzing the results, we conclude the following. First, the features similarity mirrors the exact number of features supported by the two services only. It is a direct measure that two services provide the same functionality. Consequently, we decided to assign a big weight *a* for the features similarity.

Second, the semantic similarity varies radically among services from the same application domain. The reason is that it highly depends on the terms used by the service providers in describing their services. In addition, results show that it does not precisely reflect the concrete functionalities supported by the service. Moreover, results showed that services from different application domains may have high semantic similarity. To summarize, a high semantic similarity value cannot verify that two services are similar beyond doubt. Consequently, we decided to assign a small weight *b* for the semantic similarity.

Third, the hierarchical similarity accounts for the relationships among functionalities supported by the two services even if they do not provide the exact features. It is important in the case of retrieving services with related functionalities to the user when exact matches are not available. Results show that it has a high value among

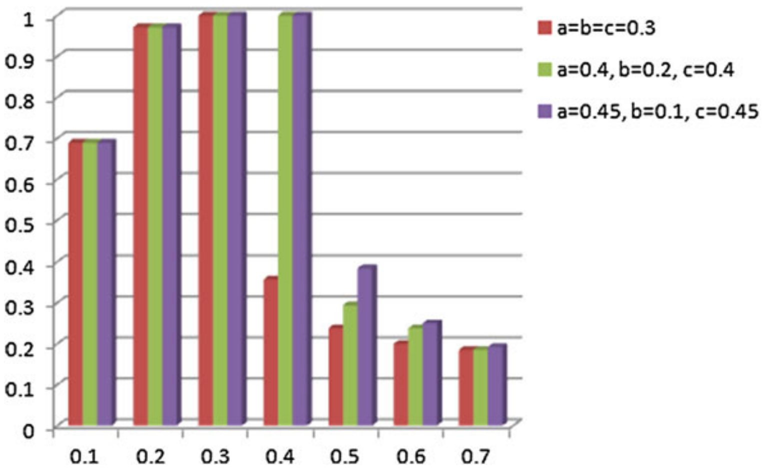


Fig. 17.7 Generated clusters F-Measure under different thresholds using different settings of **a** features similarity, **b** semantic similarity and **c** hierarchical similarity weights

services from the same domain. Consequently, we decided to assign a big weight *c* for the hierarchical similarity.

Notably, the hierarchical similarity is relatively big in the case of comparing services from different domains; the reason is the relatively small size of the ontology. It is expected that this value will decrease (for services from different domains) when the ontology size increases. Enriching the ontology with more knowledge is an ongoing work.

The second objective of this experiment is to determine the optimum value of the clustering threshold. We started the clustering threshold with 0.1, and increased it by 0.1 until it reached 0.7. For each threshold, different combinations of similarity weights were used, where a denotes features similarity weight, b denotes semantic similarity weight and c denotes hierarchical similarity weight. The F-Measure of generated clusters at each threshold value was compared with manually constructed clusters.

Summary of the results is shown in Fig. 17.7. Results show that the clustering values that result in the best F-Measure value range are 0.3 and 0.4 under all settings. To conclude this experiment, the proposed SaaS matchmaking algorithm similarity weights are as follows $a = 0.4$, $b = 0.2$, and $c = 0.4$ and the clustering threshold is 0.4.

17.5.5 Experiment 2

The objective of this experiment is twofold. First, verify the relevance of semantically enriching the services description through consulting WordNet. Second, compute its

Table 17.3 Semantic similarity of services description with and without using WordNet

Semantic similarity change (%)	Services improvement (%)	Services deterioration (%)
> 100	34.4	–
80–100	2.4	4.6
60–80	3.6	6.7
40–60	7.3	6.7
20–40	5.5	10.7
< 20	8.9	8.6
Total	62	38

overhead. To achieve the first objective, semantic similarities of services description are computed with and without using WordNet. Summary of the results is shown in Table 17.3. Results show that semantically enriching the service description usually has a positive impact on the semantic similarity between two services. However, this improvement is highly scattered because of the wide coverage of WordNet. The results heavily depend on the commonality and generalization of the terms used by the cloud providers in describing their services.

The second objective is to analyze the overhead of enriching the service description using WordNet. The overhead is computed by the average time taken with different numbers of service description terms. Results are shown in Fig. 17.10. It is evident that the overhead of the semantic enrichment of the service description is negligible. On average, it takes 70 ms for each service which clearly signifies that it can be processed online during the service registration.

17.5.6 Experiment 3

The objective of this experiment is twofold. First, identify the importance of the clustering module. Second, compute its overhead. To achieve the first objective, the average time taken to match the user request is calculated with varying number of services. Two scenarios were used: (1) using service clusters and (2) without service clusters. Results are shown in Fig. 17.9. Results clearly show that using the service clusters greatly improves the system response time by average 41 %.

To compute the clustering overhead, the time taken by the clustering algorithm is observed with varying numbers of services. This experiment was run for 20 times for each result and all results were reported on average. Results are shown in Fig. 17.10. Since the execution time of the clustering process is highly dependent on its implementation, analyzing its overhead is not so direct. However, we can get an estimation by studying the elapsed time it takes with a varying number of services. It is worth mentioning that the clustering is an offline process, which has nothing to do with processing user queries and it is executed with new service registration.

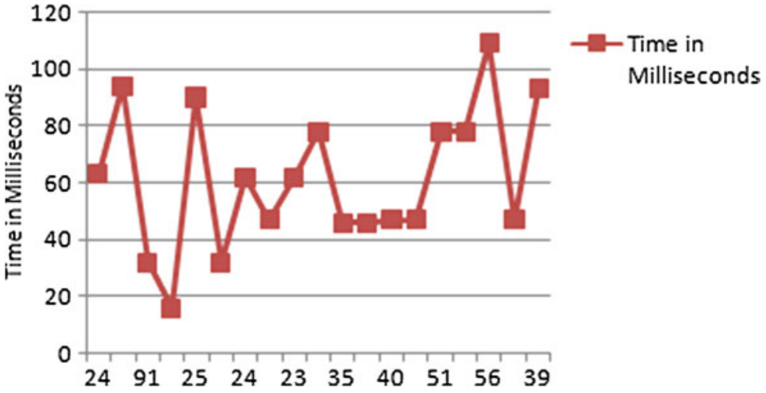


Fig. 17.8 WordNet overhead with varying number of service description terms

Fig. 17.9 Processing user query with and without service clusters

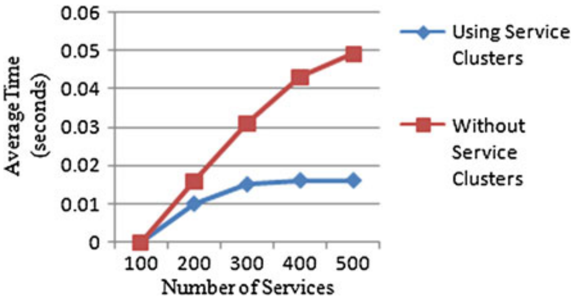
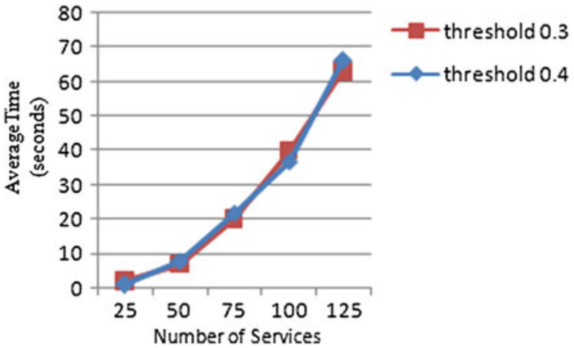


Fig. 17.10 Clustering time with increasing number of services



17.6 Conclusion and Future Work

In this chapter, a comprehensive survey to cloud services discovery and selection work was presented. The survey highlighted the need for a complete system that assists SaaS users in finding the service that meets their functional and non-functional requirements efficiently. In response to this finding, a semantic-based SaaS cloud service publication, discovery, and selection system was proposed, which facilitates the process of mapping the user specified service request to real cloud offerings. Both domain knowledge structure and services metadata are exploited, by precisely matching the service with domain ontology concepts and enriching the semantics of the service offer with descriptive metadata in the registration phase.

Unified ontology was developed, which integrates SaaS services domain knowledge, cloud service characteristics, QoS metrics, and real offers. This ontology is considered a major step towards a complete SaaS services ontology. A hybrid service matchmaking algorithm was introduced based on the proposed unified ontology. It combines semantic-based metadata and ontology-based matching. Ontological similarities include features and hierarchical similarities. We implemented a prototype for the proposed system and the results showed its effectiveness.

As part of our future work, we plan to contribute in several directions. First, develop a semantic focused crawler that collects information about real SaaS services, use this information to annotate SaaS offers, and maintain this information in the unified ontology. Second, we will investigate service fuzzy clusters, where the service offer belongs to many clusters with different degrees. Finally, further fine tune is required on the proposed matchmaking algorithm based on more experiments using different index term-based IR models.

References

1. OpenCrowd: The opencrowd cloud taxonomy. http://cloudtaxonomy.opencrowd.com. Accessed Aug 2013
2. SearchCloudComputing-TechTarget: Top cloud computing providers. http://searchcloudcomputing.Techtarget.com/feature/Top-10-cloud-computing-providers. Accessed May 2013
3. Hfer, C.N., Karagiannis, G.: Cloud computing services: taxonomy and comparison. J. Internet Serv. Appl. **2**(2), 81–94 (2011)
4. Haller, A., Strazdins, P., Zhang, M., Georgakopoulos, D., Ranjan, R.: Investigating decision support techniques for automating cloud service selection. In: Proceedings of the 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), CLOUDCOM ’12, pp. 759–764. IEEE Computer Society, Washington (2012)
5. Garg, S.K., Versteeg, S., Buyya, R.: A framework for ranking of cloud computing services. Future Gener. Comput. Syst. **29**(4), 1012–1023 (2013) (Special Section: Utility and Cloud Computing)
6. Ur Rehman, Z., Hussain, O.K., Hussain, F.K.: Laas cloud selection using mcdm methods. In: 2012 IEEE Ninth International Conference on e-Business, Engineering, vol. 0, pp. 246–251. (2012)
7. Afify, Y.M., Moawad, I.F., Badr, N., Tolba, M.F.: A semantic-based software-as-a-service (saas) discovery and selection system. In: 2013 8th International Conference on Computer Engineering Systems (ICCES), pp. 57–63. (2013)
8. Kanth, S.: Cloud service discovery system using cloud ontology. In: National conference on parallel computing technologies (PARCOMPTECH), February (2013)
9. Rodrguez-Garca, M., Valencia-Garca, R., Garca-Snchez, F., Samper-Zapater, J., Gil-Leiva, I.: Semantic annotation and retrieval of services in the cloud. In: Omatu S., Neves J., Rodriguez J.M.C., Santana J.F.P., Gonzalez S.R. (eds.) Distributed Computing and Artificial Intelligence, of Advances in Intelligent Systems and Computing, vol. 217, pp. 69–77. Springer, Heidelberg (2013)

10. Sim, K.M.: Agent-based cloud computing. IEEE Trans. Serv. Comput. **5**(4), 564–577 (2012)
11. Tahamtan, A., Beheshti, S.A., Anjomshoaa, A., Tjoa, A.M.: A cloud repository and discovery framework based on a unified business and cloud service ontology. In: IEEE Eighth World Congress on Services (SERVICES2012), pp 203–210. (2012)
12. Chen, Fei, Bai, Xiaoli, Liu, Bingbing: Efficient service discovery for cloud computing environments. In: Shen, Gang, Huang, Xiong (eds.) Advanced Research on Computer Science and Information Engineering. Communications in Computer and Information Science, vol. 153, pp. 443–448. Springer, Berlin Heidelberg (2011)
13. Wang, J., Zhang, J., Hung, P.C.K., Li, Z., Liu, J., He, K.: Leveraging fragmental semantic data to enhance services discovery. In: 2011 IEEE 13th International Conference on High Performance Computing and Communications (HPCC), pp. 687–694. (2011)
14. Dastjerdi, A.V., Tabatabaei, S.G.H., Buyya, R.: An effective architecture for automated appliance management system applying ontology-based cloud discovery. In: 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), pp. 104–112. (2010)
15. Chen, H.P., Li, S.C.: Src: A service registry on cloud providing behavior-aware and qos-aware service discovery. In: 2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA), pp 1–4. (2010)
16. Dong, H., Hussain, F.K., Chang, E.: A service concept recommendation system for enhancing the dependability of semantic service matchmakers in the service ecosystem environment. J. Netw. Comput. Appl. **34**(2), 619–631 (2011) (Efficient and Robust Security and Services of Wireless Mesh Networks)
17. Zhang, J., He, L.W., Huang, F.Y., Liu, B.: Service discovery architecture applied in cloud computing environments. Appl. Mech. Mater. **241**, 3177–3183 (2012)
18. Fellbaum, C.: Wordnet. In: Theory and Applications of Ontology: Computer Applications, pp. 231–243. Springer, Netherlands, (2010)
19. Limam, N., Boutaba, R.: Assessing software service quality and trustworthiness at selection time. IEEE Trans. Software Eng. **36**(4), 559–574 (2010)
20. Lenk, A., Menzel, M., Lipsky, J., Tai, S., Offermann, P.: What are you paying for? performance benchmarking for infrastructure-as-a-service offerings. In: 2011 IEEE International Conference on Cloud Computing (CLOUD), pp. 484–491. (2011)
21. Li, A., Yang, X., Kandula, S., Zhang, M.: Cloudcmp: comparing public cloud providers. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, IMC '10, pp. 1–14. ACM, (2010)
22. Li, A., Yang, X., Kandula, S., Zhang, M.: Comparing public-cloud providers. IEEE Internet Comput. **15**(2), 50–53 (2011)
23. Qu, L., Wang, Y., Orgun, M.A.: Cloud service selection based on the aggregation of user feedback and quantitative performance assessment. In: Proceedings of the 2013 IEEE International Conference on Services Computing, SCC '13, pp. 152–159. IEEE Computer Society, Washington (2013)
24. Ur Rehman, Z., Hussain, O.K., Parvin, S., Hussain, F.K.: A framework for user feedback based cloud service monitoring. In: 2012 6th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp. 257–262. (2012)
25. Godse, M., Mulik, S.: An approach for selecting software-as-a-service (saas) product. In: Proceedings of the 2009 IEEE International Conference on Cloud Computing, CLOUD '09, pp. 155–158. IEEE Computer Society, Washington (2009)
26. Park (Jong Huk), J., Jeong, H.-Y.: The qos-based mcdm system for saas erp applications with social network. J. Supercomputing 1–19, (2012)
27. Salaja, S., Rajsingh, E.B., Ezra, K.: Efficient service selection middleware using electre methodology for cloud environments. Inf. Technol. J. **11**(7), 868–875 (2012)
28. Sun, M., Zang, T., Xu, X., Wang, R.: Consumer-centered cloud services selection using ahp. In: Proceedings of the 2013 International Conference on Service Sciences, ICSS '13, pp. 1–6. IEEE Computer Society, Washington (2013)

29. Bedi, P., Kaur, H., Gupta, B.: Trustworthy service provider selection in cloud computing environment. In: 2012 International Conference on Communication Systems and Network Technologies (CSNT), pp. 714–719. (2012)
30. Ruiz-Alvarez, A., Humphrey, M.: An automated approach to cloud storage service selection. In: Proceedings of the 2nd international workshop on Scientific cloud computing, ScienceCloud '11, pp. 39–48. ACM, New York (2011)
31. Zhang, M., Ranjan, R., Nepal, S., Menzel, M., Haller, A.: A declarative recommender system for cloud infrastructure services selection. In: Economics of Grids. Clouds, Systems, and Services, volume 7714 of Lecture Notes in Computer Science, pp. 102–113. Springer, Berlin Heidelberg (2012)
32. Sundareswaran, S., Squicciarini, A., Lin, D.: A brokerage-based approach for cloud service selection. In: 2012 IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 558–565. (2012)
33. Shangguang, W., Zibin, Z., Qibo, S., Hua, Z., Fangchun, Y.: Cloud model for service selection. In: IEEE INFOCOM 2011–IEEE Conference on Computer Communications Workshops, 10–15 April pp. 666–671. (2011)
34. Dong, H., Hussain, F.K., Chang, E.: A service search engine for the industrial digital ecosystems. IEEE Trans. Industr. Electron. **58**(6), 2183–2196 (2011)
35. Goscinski, Andrzej, Brock, Michael: Toward dynamic and attribute based publication, discovery and selection for cloud computing. Future Gener. Comput. Syst. **26**(7), 947–970 (2010)
36. Ngan, L.D., Kanagasabai, R.: Owl-s based semantic cloud service broker. In: 2012 IEEE 19th International Conference on Web Services (ICWS), pp. 560–567. (2012)
37. Androcec, D., Vrcek, N., Seva, J.: Cloud computing ontologies: a systematic review. In: The 3rd International Conference on Models and Ontology-based Design of Protocols, Architectures and Services (MOPAS 2012), (2012)
38. Born, M., Filipowska, A., Kaczmarek, M., Markovic, I., Starzecka, M., Walczak, A.: Business functions ontology and its application in semantic business process modelling. In: Proceedings of the 19th Australasian Conference on Information Systems, pp. 136–145. Christchurch, (2008)
39. Fortis, T.-F., Munteanu, V.I., Negru, V.: Towards an ontology for cloud services. In: 2012 6th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp. 787–792. (2012)
40. Hepp, M., Radinger, A.: Eclassowl. the products and services ontology. http://www.heppnetz.de/eclassowl/. Accessed May 2013
41. Joshi, K., Yesha, Y., Finin, T.: Automating cloud services lifecycle through semantic technologies. IEEE Trans. Serv. Comput. **PP**(99), 1–1, (2012)
42. Moscato, F., Aversa, R., Di Martino, B., Fortis, T., Munteanu V.: An analysis of mosaic ontology for cloud resources annotation. In: 2011 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 973–980. (2011)
43. Youseff, L., Butrico, M., Da Silva, D.: Toward a unified ontology of cloud computing. In: Grid Computing Environments Workshop, 2008. GCE '08, pp. 1–10. (2008)
44. Wikipedia: Wikipedia, the free encyclopedia. http://en.wikipedia.org. Accessed 22 Jan 2013
45. United Nations Development Programme (UNDP): The united nations standard products and services code. http://www.unspsc.org/. Accessed 1 Jan 2013
46. North American Industry Classification System (NAICS): Naics. http://www.census.gov/eos/www/naics/. Accessed 15 Jan 2013
47. W3C Recommendation: Web ontology language (owl). http://www.w3.org/TR/owl-features/. Accessed Dec 2012
48. Horridge, M.: A practical guide to building owl ontologies using protege 4 and co-ode tools. Technical report, The University Of Manchester (2011)
49. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Inc, New York (1986)
50. The National Institute of Standards NIST and Technology: Cloud computing reference architecture. http://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/ReferenceArchitectureTaxonomy. Accessed Dec 2012

51. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud computing systems. In: Proceedings of the 2009 5th International Joint Conference on INC, IMS and IDC, NCM '09, pp 44–51. IEEE Computer Society, Washington (2009)
52. Tran, V.X., Tsuji, H., Masuda, R.: A new qos ontology and its qos-based ranking algorithm for web services. Simul. Model. Pract. Theory **17**(8), 1378–1398 (2009)
53. Saaty, T.: Theory and applications of analytic network process: Decision making with benefits, opportunities, costs and risks. RWS publications, 2005
54. International Organization for Standardization: Iso/iec 9126–1:2001 software engineering product quality - part 1: Quality model. http://www.iso.org/iso/home/store/catalogue_ics.htm. Accessed Dec 2012
55. Cross, Valerie, Xinran, Yu., Xueheng, Hu: Unifying ontological similarity measures: a theoretical and empirical investigation. Int. J. Approximate Reasoning **54**(7), 861–875 (2013)
56. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, ICML '98, pp. 296–304. Morgan Kaufmann Publishers Inc., San Francisco (1998)
57. Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Know. Based Syst. **24**(2), 297–303 (2011)
58. Resnik, P.: Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Artif. Intell. Res **11**, 95–130 (1999)

Chapter 18
Data and Application Security in Cloud

Rajesh P. Barnwal, Nirnay Ghosh and Soumya K. Ghosh

Abstract Cloud computing is an emerging technological paradigm, which provides computing resources as utility. Like other day-to-day utilities, cloud computing follows *pay-as-you-use* model, where users are charged according to the usage without regard to where the services are hosted or how they are delivered. Today, majority of companies follow an IT infrastructure-driven business model. With the growing demand, rise in customer base and market place competitions, companies prefer focusing on respective business policies and services they offer, rather than IT management overheads. Therefore, there is a high probability that the future of present day business model may shift to clouds where non-IT companies no longer have to procure, manage, and maintain IT resources. They will host applications and data to the servers, which are deployed by cloud providers, possibly in geographically dispersed locations. However, security is a major challenge before outsourcing any IT needs of business. As cloud provides a multi-tenant virtual computing environment, where competitive businesses may co-exist, hosting of sensitive information for mission-critical applications is of utmost concern. This chapter reviews the recent works reported specifically in the area of *data and application security* relevant to cloud computing. Some works which use biologically inspired phenomenon to manage security and load balancing in cloud environment, have also been studied. The aim of this chapter is to provide an insight into the present state-of-the-art cloud security problems, proposed solutions, and identify future research directions as well as scopes in various security issues.

R. P. Barnwal (✉)
Information Technology Group, CSIR-Central Mechanical Engineering Research Institute,
Durgapur 713209, India
e-mail: r_barnwal@cmeri.res.in

N. Ghosh · S. K. Ghosh
School of Information Technology, Indian Institute of Technology, Kharagpur 721302, India
e-mail: nirnay.ghosh@gmail.com

S. K. Ghosh
e-mail: skg@iitkgp.ac.in

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_18, © Springer-Verlag Berlin Heidelberg 2014

18.1 Introduction

Computing paradigm has witnessed different phases of transitions, starting from as early in 1960s. The first phase witnessed was *Mainframe* in which users with thin terminals connect to a powerful Mainframe server which is shared among multiple tenants. The second phase saw the advent of *Personal Computers (PCs)* which are stand-alone, high-performance, and powerful enough to satisfy general-purpose requirements. In the next phase, *Networks* came into existence, which connect multiple PCs locally to share the resources. In the following phase, computing paradigm shifted to *Internet*, which connects one local network to another enabling access to remote applications and resources. *Grid computing* was introduced as a parallel and distributed high-performance computing, to utilize the shared computing power and storage resources for solving computation intensive scientific applications. Grids are composed of a number of remotely placed clusters which are essentially groups of kernels that are connected by private local area network and communicate among each other over low-bandwidth links. At present, the paradigm is shifting to *cloud computing*, which in simple term exploits all the available resources on the Internet in a scalable and distributed way. This chapter presents a comprehensive survey on data and application security in cloud environment related works that have been reported between 2008–2013. We also propose potential research directions in the area of cloud computing security.

The rest of the chapter is as follows: Sect. 18.2 introduces to the concept of cloud computing and gives an insight into its security issues. Section 18.3 discusses about works on data and application security. Section 18.4 identifies the probable future research directions and Sect. 18.5 presents the overall conclusion.

18.2 Cloud Computing

There are number of reasons contributing to the advent of cloud computing. On one hand, it allows better resource utilization through scalability on-demand, while on the other, it minimizes IT resource (viz. servers, storage devices, network devices, softwares, applications, IT personnel, etc.) management overhead for non-IT companies and allows them to focus on improving core business processes. In addition to these, cloud computing reduces start-up costs which is beneficial to small-scale companies. It enables *economy of scale* by multiplexing the same physical resource among several tenants. This is also an advancement towards *green computing*, where the primary objective is to reduce carbon footprints generated due to deployment of data centers.

Cloud computing has numerous definitions coined by technical, research, business, as well as scientific communities. However, the most widely accepted definition is the one given by *National Institute of Standards and Technology* (*NIST*) [1] as follows:

"*Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable computing resources* (*e.g. networks, servers, storage, applications, and services*) *that can be rapidly provisioned and released with minimal management effort or service provider interaction.*"

NIST's definition for cloud computing encompasses all of its attributes which are given as:

- *On demand self-service* resources can be used as and when required with minimal human interaction between the user and the provider.
- *Ubiquitous network access* all cloud services are accessible over Internet through web services.
- *Resource pooling* cloud computing gives a notion of 'infinite' resource for moderately large requests whose efficient and optimal allocation enables serving multiple customers.
- *Location independence* cloud resources and customers may be located at geographically dispersed locations.
- *Rapid elasticity* resources can be scaled up or down depending on workload there by minimizing server idle time.
- *Measured services* customers are charged based on measured usage of the cloud resources

A good section of technological, scientific, and research community view cloud computing as a combination of existing technologies amalgamated with business models. Some of the technologies whose characteristics have influenced the "architecture" of the cloud computing [2] are given in Table 18.1. Cloud computing is found to be advantageous to companies that aim at improving businesses processes without managing IT infrastructures. However, there are a number of issues related to cloud computing that still prevent companies from moving their business onto public clouds. Among them, security is the primary concern that has inhibited the use of cloud to its full potential. This is evident from the outcome of a number of surveys done by various agencies. The survey (refer to Fig. 18.1) done by IDC Enterprise panel[1] during 2009 on 263 corporate executives reveals that security is considered to be a major hindrance for outsourcing business onto public clouds. Similar result is established through another survey conducted by IDC during April 2010[2] which enquired participants about the potential barriers before adopting public and private clouds. It was observed that security in public cloud is the major concern for more than 70 % participants, and that for private cloud is to around 45 % people. Other than security, the factors which are of concern to the IT decision-makers are:

- Lack of technology maturity
- Lack of personnel skill sets
- Organizational challenges
- Difficulty in integrating with existing infrastructure

[1] http://idcenterprisepanel.com/index.html.

[2] http://www.idc.com/getdoc.jsp?containerId=223077.

Table 18.1 Technologies influencing cloud architecture

Technology	Influencing characteristics
High Performance Computing	Connecting low-cost, commercially available personal computers in a network cluster to form a high-performance computing system (parallel computing)
Autonomic computing	Innovative mechanisms to manage, operate, and maintain complex systems
Utility and enterprise grid computing	Sharing, selection, and aggregation of geographically distributed autonomous resources dynamically depending upon availability, capability, performance, cost, and QoS requirements
Service consolidation	Develops services as cost-effective tools to be deployed over a shared infrastructure and restores QoS compliance, security, and governance
Horizontal scaling	implementing redundancy and reliability to loosely coupled systems by adding more of the individual resource elements, such as servers
Web services	Standard interface (described in *WSDL*) to provide communication among computing platforms running different applications
Virtualization	Rapid deployment of additional servers, effective utilization of the resources, promotion of economies of scale, and so on

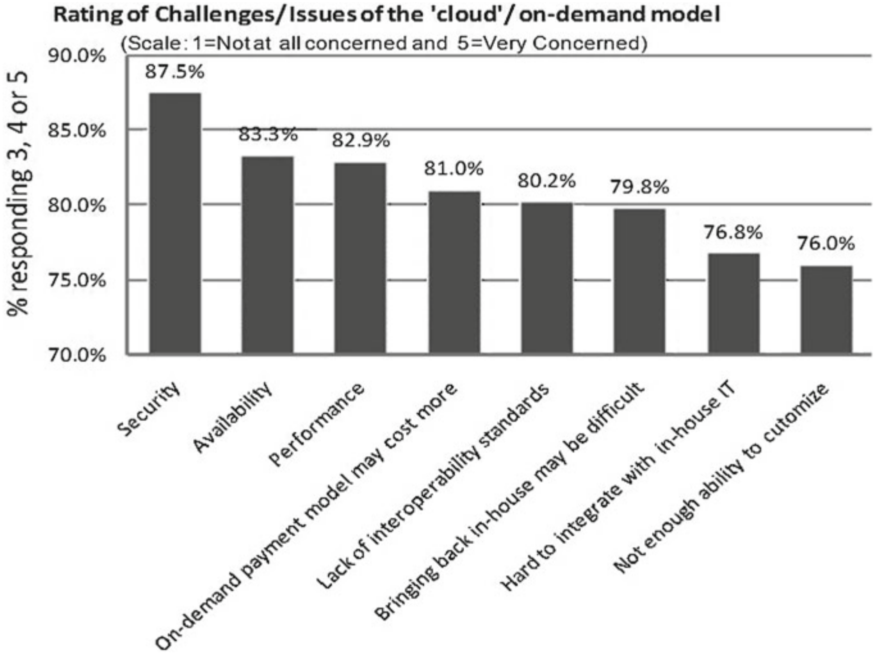


Fig. 18.1 Survey on potential cloud barriers (*Source* IDC ranking security challenges)

However, these barriers are expected to be overcome as cloud technologies evolve with time.

18.2.1 Security in Cloud Computing

Cloud computing environment is considered to be vulnerable by organizations due to its multi-tenancy and lack of customer's control over data and applications. In contrast, cloud service provider is concerned about *reputation fate sharing*, by which, malicious behavior of a single cloud customer can affect the reputation of other customers leading to seizure of their resources. Cloud security issues are similar to traditional in-house IT deployments [2]. Securing a cloud also requires addressing the well-known *CIA* (*Confidentiality, Integrity, Availability*) tenets of information and system security:

- *Confidentiality* Isolate customer instances and applications in a multi-tenant environment to support confidentiality or privacy as well as industry standard identity management.
- *Integrity* Data and applications running in the virtual instances should be prevented from tampering to restore integrity. This demands for data protection against worms, viruses, spywares, trojans, or even application-specific scripting, and injection attacks.
- *Availability* The computation service and data should be available whenever customer requires them. This requires prevention of *Denial-of-Service (DoS)* attacks which includes *SYN flooding*, *ICMP flooding*, etc.

Cloud also needs to address a myriad of security issues specific to its domain [3]. Some of these issues are as follows:

- *Privileged access* It implies sharing of responsibility between a client and a cloud service provider (CSP). Both would have conflicting requirements, the client would be more concerned about data security whereas the CSP would be more concerned about the efficiency of the cloud environment.
- *Regulatory compliance* Any CSP should be willing to undergo security audits and comply to government regulations. The laws again would differ from one country to the other.
- *Data segregation* Cloud service would need to take into account insecure communication channels between client and data center managed by CSP. Hence data would need to be encrypted.
- *Recovery* CSP would have to provide appropriate data recovery schemes in case of disaster.
- *Investigative support* Any vendor would need to have the ability to detect inappropriate activities.
- *Long term viability* Ideally, cloud computing provider should never get acquired and swallowed up by a larger company, and customers must be sure that their data will remain available even after such event.

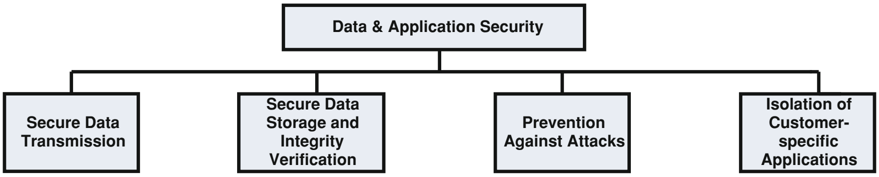

Data & Application Security
Secure Data Transmission
Secure Data Storage and Integrity Verification
Prevention Against Attacks
Isolation of Customer-specific Applications

Fig. 18.2 Classification of data and application security in cloud

Security in cloud computing is a vast area and focuses on a number of issues. The scope of this work is limited by the classification presented in Fig. 18.2.

18.3 Data and Application Security

Cloud provides a shared virtual infrastructure, where data and applications of customers are mostly under control of service providers. In many cases, data stored in a cloud is proprietary and therefore a provider needs to instil confidence and trust in users about disclosure and protection of data from unauthorized use. In cloud, a user would be interacting with many devices over ubiquitous interfaces and platforms. Therefore, ensuring end-to-end *secure data flow* among participating entities is an important issue. Cryptographic mechanisms are applied on cloud data to restore confidentiality and integrity. However, the encrypted data have to be decrypted at the time of processing, which leads to the possibility of side-channel attacks. Hence, after data reaches the end points, it is necessary to verify its integrity and consistency.

In software-as-a-service (SaaS) delivery model, almost all cloud applications are web service-based, eliminating the need to install and run application on customer's own computers. This simplifies issues of hardware compatibility, license, security patch management, and maintenance. However, like traditional IT networks, vulnerabilities related to web-based applications also prevail in cloud. Possible threats in such systems are denial-of-service attack, loss of sensitive data due to attacks originating from misconfiguration of firewall, improper input string checks, XML-specific attacks, and access rules violation attacks. Therefore, prevention against such attacks is an important requirement for data and application security in cloud.

At infrastructure-as-a-service (IaaS) layer of abstraction, cloud enables multiple customers to run respective applications on a shared physical server. Such multiplexing of resource to execute several workloads is done by a technology, known as, *virtualization*. Advantages of virtualization are: (i) increased server utilization, (ii) reduced space, and (iii) power consumption. A software layer called the *hypervisor* or *Virtual Machine Monitor* (*VMM*) manages instances called *virtual machines* (*VMs*), which run concurrently on the same physical server. However, placing different customers' workloads on same physical system leads to requirements of isolation management, workload management and access control. The following subsections review the different aspects of data and application security in a cloud environment.

18.3.1 Secure Data Transmission

Xu et al. [4] have addressed the issues of confidentiality and authentication of data in cloud environment. The major threats are data interception during transmission from client site to cloud and authorized access to documents stored in the cloud. The authors have proposed a *secure document service*, which encrypts the content of a document without altering its format. The proposed authorization model depends on encryption functions and public and private keys of the users. The model of mechanism is as follows : (i) Documents are stored in *Cloud Document Warehouse*. (ii) The document format and content separation is done by *Document Service*. (iii) Owner in client-end is responsible for encryption and decryption of document content. (iv) To save the document, client must re-encrypt partitioned content and then send it to *Document Service*. However, the encryption algorithm has not been implemented in the present work. The separation of content and format requires semantic interpretation of the document and involves additional overhead in reformatting after every update.

In [5], the authors have discussed about security of data intensive applications that requires interoperability, integration, and sharing of multiple domains. Security of such applications addresses security of infrastructure, communication network and user. The report also gives details about using *Declarative Secure Distributed System* (DS^2), an infrastructure for specifying, analyzing, and deploying secure information systems. In DS^2, network protocols and security are specified in *Secure Network Datalog* (*SeNDlog*) language. *SeNDlog* provides support for authenticated communication and confidentiality of transmitted tuples. The work covers secure query processing in a cloud computing environment which requires processing of query to authenticate the users and machines, the data transfer across machines and finally integrity of the query.

In [6], the authors have proposed a novel method for effective and flexible distributed scheme with dynamic data support to ensure correctness of users' data in the cloud. The adversary model in cloud environment consists of CSPs and adversaries. A CSP can be self-interested, untrusted or malicious. It may attempt to move infrequently accessed data to low-level storage, hide data loss incident to management error, byzantine failures etc. Depending on the level of capabilities, adversaries are of two kinds:

- *Weak adversary* Interested in corrupting users data on individual servers.
- *Strong adversary* Capable of compromising all the storage servers in order to modify data files as long as they are internally consistent.

The system model considered comprises of three actors: *users*, *cloud service provider* (*CSP*), and *third party auditor* (*TPA*). As users are not data holders, it is critical to ensure that their data is correctly stored and maintained. This can be ensured by delegating the monitoring responsibility to the optional TPA. However, assumption is that point-to-point communication channels are authenticated and reliable.

In [7], the authors have designed and implemented a new secure data flow processing system that aims at providing trust in multiple clients open distributed systems. The scheme provides confidentiality and integrity for dataflow processing applications by imposing lightweight processing for both composer and service components. Some of the assumptions in this implementation are as follows:

- Both component and component service providers use public/private key pairs to bound themselves for encrypting, decrypting and data signature.
- A party cannot forge signatures or decrypt encrypted data using other's public keys.

For data centric networks a major shortcoming is the under provisioning of resources. An application employed in such infrastructure forces the application owners to take into account the infrastructure limitations. This, in turn, implies building counter measures to ensure that applications are secure and meets the required performance. In [8], the authors proposed a *peer to cloud and peer* (*P2CP*) system in which cloud servers can communicate among each other through three secure tunnels: (i) cloud-user data transmission tunnel, (ii) clients data transmission tunnel, (iii) common data transmission tunnel.

18.3.2 Secure Data Storage and Integrity Verification

Bowers et al. [9] proposed a High-Availability and Integrity Layer (HAIL) for secure cloud storage. HAIL is aimed to manage file integrity and its availability across a collection of servers and independent storage devices. It uses POR [10] as building block to test and relocate storage resource in case of failure. The proposed solution relies on a single trusted verifier (either client or a service), who can verify the integrity of stored files and also enables a set of servers to prove to a client that a stored file is fully intact.

Wang et al. in [6] studied the problem of ensuring the integrity of data storage in cloud computing environment and then proposed a scheme for public verifiability and data dynamics in cloud storage. The work presented a method to verify the integrity of the dynamic data storage in cloud. In this scheme, BLS-based construction has been used to achieve both public verifiability and data dynamism. The authors extended the proof of retrievability (PoR) [11] model using Merkle hash tree construction to achieve fully dynamic data operation. The scheme adopts a blockless approach for remote data checking functionalities.

Verification of untrusted cloud storage (*Venus*) has been proposed in [12]. It is a service for a secure user interaction in untrusted cloud storage. The service guarantees integrity and consistency for applications accessing storage data based on keys. When a user accesses data using Venus, it completes the operations optimistically, guaranteeing at all times the integrity of the data. After the completion of the operations, it verifies the operation consistency and sends a notification to the application. *Venus* provides two guarantees: (1) *Integrity* which means that the stored objects are protected against malicious data modification. (2) *Consistency* which allows multiple clients to access stored data concurrently in a consistent fashion. The system model includes a storage service, a verifier, and multiple clients.

Chen et al. [13] proposed a new protocol for *Remote Data Checking* (*RDC*) in an untrusted environment based on *network coding* (*NC*) which has been termed as *RDC-NC*. RDC is a technique which works on the client-server concept. Traditionally, RDC enables a client to verify if data stored on an untrusted server remains consistent over time. The data at the server is checked periodically and if corruption is detected, data recovery is initiated. Other reported forms of RDC include redundancy in distributed storage through *replication* and *erasure coding*. The *RDC-NC* protocol involves three phases: (*i*) *Setup*, (*ii*) *Challenge*, (*iii*) *Repair*. The proposed protocol is expected to have additional overhead of calculation of security keys and maintenance of repair verification tag. Also the scheme can only be used for archival storage due to inherent limitations of network coding.

The authors in [14] have proposed *RunTest*, a light weight application level alteration method to verify the integrity of data processing results in the cloud. In RunTest, randomized attestation using a small subset of the input data over different subsets of cloud nodes is performed. The technique employs an attestation graph model to account for different attestation results. Further, the authors have designed a clique based attestation graph analysis algorithm to locate malicious service providers. RunTest performs integrity attestation by using the final output and does not rely on trusted hardware like *TPM* or secure kernel. Migration of data processing to a cloud architecture involves security concerns regarding privacy of the data. In that respect, this work is significant. However, some assumptions adopted for processing nodes such as, they are stateless and only have deterministic events, may not be a representative of a practical environment in a cloud. Therefore, the work is bound to have limited applicability.

Kamara et al. in [15] discussed about cryptographic storage in cloud. In the chapter, authors presented the survey of several cryptographic architectures, which are suitable for building a secure cloud storage service on top of public storage cloud. Among various contemporary cryptographic primitives, searchable encryption like symmetrical searchable encryption (SSE), asymmetric searchable encryptions (ASE), Efficient ASE (ESE) and multi-user SSE (mSSE) have been described. Moreover, the authors also enlisted other cryptographic techniques such as attribute-based encryption, which allows specifications of a decryption policy with a ciphertext. According to them, the concepts like proofs of storage (POR) [10, 16] and dynamic proofs of storage (with data updation facility) [17, 18] are the effective tools for verification of integrity of the stored data.

Popa et al. [19] presented a secure storage system named *CloudProof* to help the data-owners in detecting violations of integrity, write-serializability, and freshness of their stored data in clouds. The proposed system also allows customers to prove cloud misbehavior, if occurs, to a designated third party. It claims that the suggested method of cloud security provides highly scalable solution and maintains availability and performance of the cloud services inspite of security overhead. In [20], the authors propose a service named *SSTreasury* + which includes *encryption application* and *cloud storage service*. User's data is encrypted prior to uploading to the cloud to prevent it from being stolen during transmission or in the cloud storage.

18.3.3 Prevention Against Attacks

Many websites allow users to access their backend databases using well-defined interfaces. Such databases are referred as *Hidden Databases*. *Data harvesting* are attacks launched to extract information from *Hidden Databases*. These attacks can be carried out through web crawlers (crawling attack) or sampling of data resulting through adaptive sequence of queries (sampling attacks). In [21], the authors have proposed the *HengHa* system to prevent such *data harvesting* attacks. It consists of two subsystems:

1. *Heng* subsystem Uses frequent pattern mining to find query correlation in a session.
2. *Ha* subsystem It efficiently finds the result coverage of a session with a *coverage bit vector*.

The *Heng* subsystem uses the assumption that a data harvesting session has relatively lower query correlation to a normal user session. The *Ha* subsystem evaluates the resulting coverage of queries using a training pattern.

Faatz et al. [22] have emphasized on three perspective of information security in the clouds:

1. *Protecting data* A cloud customer relies upon cryptographic capabilities offered by the provider. Inspite of encryption, data needs to be exposed for processing and therefore, customers should be allowed to verify these techniques and, if possible, install third-party encryption software.
2. *Protecting information* Information stored in a public cloud is required to be periodically copied to make it resilient and fault-tolerant. To protect and minimize exposure of *personally identifiable information* (*PII*), the authors propose integration of enterprise *Identity and Access Management* (*IdAM*) capabilities to those of Cloud's.
3. *Monitoring and defending systems* State-of-the-art clouds lack efficient monitoring infrastructure. Moreover, audit trails from network devices, operating systems, and applications are almost inaccessible to the customers. Therefore, there is a requirement to develop new monitoring procedures for handling security incidents. Such mechanism should clearly differentiate the security responsibilities between provider and customer, and also enables each party to verify whether they are meeting the responsibilities.

In [23], *DoS* has been described as an attack which exploits an under-provisioned network in a cloud infrastructure. A cloud data center could be used by multiple clients, who have hardly any control on the underlying network. An adversary can launch the *DoS* attack by gaining access to a set of hosts, learns the topology, and then send a large amount of traffic through the upstream router to saturate the link's uplink. A possible solution to prevent such attack is by limiting the bandwidth consumption by each server.

Kupsch et al. [24] presents a novel analyst-centric technique, *First Principles Vulnerability Assessment* (*FPVA*), for cloud middlewares. *FPVA* is capable of finding flaws in a program source code that are normally not detected by some commercial tools. Some of the vulnerabilities detected by *FPVA* include erroneous or changeable configuration files, injection attacks and race condition. The security vulnerability has been rated on four-level scale which are: *Level 0 or False alarm*, *Level 1 or Zero-value vulnerability*, *Level 2 or Low-value asset access* and *Level 3 or High value asset access*. *FPVA* is primarily used to correct *Level 3* vulnerabilities, and to some extent, those at *Level 2*. The major steps of operation followed by *FPVA* are: (i) *architectural analysis*, (ii) *resource identification*, (iii) *trust and privilege analysis*, and (iv) *component evaluation*. Comparison of *FPVA* with automated source code analysis tools has been reported with several advantages, viz. detecting complex scenario type vulnerabilities, relatively lesser number of *false positives*, and so on.

In [25], the authors have discussed about the security issues relevant to various technologies that are enabling different cloud delivery models (viz. SaaS, PaaS, IaaS). As most of the services offered by cloud are web-based applications, therefore, securing web services (WS) is of great importance. Such *WS-security* requires enforcing XML security standards, that includes *XML signature* to ensure integrity and authenticity, and *XML Encryption* to ensure confidentiality and certification from trusted *Certification Authority* (*CA*). The authors have cited examples of attacks which are pertinent to cloud enabling technologies. Some of them are as follows:

1. *XML Signature Element Wrapping* (*wrapping attack*) By this attack, the original body of an XML message is moved to a newly inserted wrapping element inside the *SOAP* header, and a new body is created.
2. *Browser security* Such security issues encompasses *DNS poisoning* and *Phishing* attacks.
3. *Cloud malware injection attack* This includes creating malicious service implementation module (SaaS or PaaS) or malicious virtual machine instance (IaaS).
4. *Metadata spoofing attack* Such attack involves malicious reengineering of Web Services' metadata description (e.g. WSDL)
5. *Flooding attacks* Attacker sending huge amount of request to a certain service and causing denial of service.

In [26], the authors critically examined the known security challenges by leveraging similarity between cloud computing and traditional in-house computing. Two important observations which have been presented are as follows:

1. Attacks against injection vulnerabilities are difficult to detect in web application-based cloud services.
2. Encryption is not a feasible option to restore data confidentiality particularly when an application is data intensive and requires high I/O throughput.

The authors does not recommend *debugging* as a process to remove coding flaws in cloud-based web applications. This is because, it is difficult to debug large distributed applications whose testing and development environments are significantly different from deployment ones. Moreover, state-of-the-art debugging tools are not suitable for web-oriented frameworks. Instead, they suggest for *mutual auditability* between the cloud provider and the customer. However, collecting logs is difficult when applications are distributed and shared among multiple tenants.

Zargar et al. [27] give insight into distributed, collaborative and data-driven intrusion detection and prevention problems in cloud computing. The work proposed a intrusion detection and prevention framework, which operates at three architectural levels namely *Infrastructure*, *Platform* and *Virtual Machine*. It is proposed that to mitigate the attacks, the cloud service providers need to collaborate in a distributed manner to perform the task of intrusion detection and prevention. The proposal seems effective but the applicability of the proposed framework in real scenario needs to be established.

Akbarabadi et al. [28] discuss port-scanning attacks in cloud computing and their possible solutions. In port-scanning attacks, an attacker tries to gather some specific information about status of the ports which are potential opening for launching other kind of attacks. The paper enlisted some known methods for detection of port-scanning attacks such as *Time Independent Feature Set* (*TIFS*), *Packet Counting, Fuzzy Logic based Mechanism, IP classification, Packet Capturing, Network Forensic, Term Frequency-Inverse Document Frequency* (*TF-IDF*) *and Embedded Port Scan Detector*.

18.3.4 Isolation of Customer-Specific Data and Application

Trusted Virtual Data Center (*TVDc*) [29] is a technology which addresses the issues of isolation and integrity in a virtualized cloud environment. VMs and associated resources are grouped into *trusted virtual domains* (*TVDs*) which are security domains enforcing a uniform isolation policy across its members. This policy describes the constraints on the communication between VMs. TVDc uses existing components like *role based access control (RBAC)*, *hypervisor-based isolation* and *protected communication channels* like *VLANs* to implement the constraints. Integrity guarantees are based on the trust on the root. *TVDc* employs *Trusted Computing Group* (*TCG*) load time attestation mechanism to verify software integrity of the system. The basic goals of the TVDc are:

1. Prevent data from leaking from one customer workload to another even on the occurrence of faults in the VM.
2. Prevention of malicious code from spreading from one customer workload to the other.
3. Prevention of break-ins in one workload threatening any other resource on the same physical resource.
4. Prevention or reduction of misconfiguration of management tasks.

The work reported in [30] is an extension to the prototype proposed in [29]. It implements a controlled access to networked storage based on security labels and by implementing management prototypes it demonstrates enforcement of isolation constraints and integrity checking. The boundaries of a TVD are defined by labeling all VMs and associated resources with a unique TVD identifier called a *security label*. The TVDc isolation policies are implemented using anti-collocation policies based on security labels and RBAC. Roles are assigned to administrators and TVD security labels or *colors* are assigned as permissions. VMs can share data if they share a common color. A virtual machine monitor (VMM) can start a VM only if the colors contained in the VMM system label are a superset of colors assigned to a VM. Anti-collocation constraints are rules which describe a set of colors that conflict. The VMs with conflicting colors cannot be run on the same VMM system. Sharing, system authorization and collocation constraints are enforced by the VMM (using another VM called management VM), storage and network infrastructure. A central management application is used to define the TVDs, assign labels to physical and virtual resources, to deploy security policies and consolidate integrity measurements. The system management application defines various levels of hierarchy like *IT data center administrator*, *TVDc administrator* and *tenant administrator*. The authors have implemented the TVDc prototype using a proprietary IBM hardware and software.

Leakage of tenants' confidential data to their competitors and attackers are common problem in cloud computing systems. Kodialam et al. [31] proposed a solution for protecting cloud data using dynamic inline fingerprint checks. The proposed approach extracts the fingerprints from the white-listed documents to build a database. During the data transfer phase, the outgoing data is inspected to compute and compare the same with the fingerprint database. This approach is different from other existing solutions that performs the checking of outgoing data based on specified keywords or pre-defined patterns.

Multicloud architecture is used to reduce the potential problem of data manipulation, data disclosure, and tampering of the application processes hosted on a single cloud. On single cloud architecture, once the cloud is compromised, the complete set of data and application are exposed. Multicloud architecture prevents the external attacker from retrieving or tampering hosted data or applications of the user because of their federated nature. However, the simultaneous use of distinct multiple cloud poses some new security threats. Bohli et al. [32] introduced a model to discuss the security benefits of using different architectural patterns for distributing resources to multiple cloud providers. These four distinguished architectural patterns are *Replication of applications, Partition of application system into tiers, Partition of application logic into fragments and Partition of application data into fragments*. The use of combined architectural pattern can help in achieving combined security benefits in multiclouds. In [33], authors use a variant of Ant Colony Optimization, *Cross-Entropy Ant System* (*CEAS*), to construct balanced and dependable deployment configurations that are resilient. This approach is based on heuristic and decentralized optimization method focusing on finding suitable mappings between VM replicas and nodes. Shen et al. [34] have proposed a *shadow price guided algorithm* (*SGA*), influenced by traditional genetic algorithm. It attempts to generate a task scheduling mechanism that completes all assigned tasks with minimal energy consumption, which in turn, improves performance of cloud servers. In [35], a honey bee-based load balancing technique has been proposed to promote load balancing on the global (cloud) scale via actions and interactions at the component (individual server) level. Such load balancing technique uses a collection of virtual servers, each serving a virtual service queue of requests. Any server takes a particular bee role with certain probabilities, either to randomly choose a virtual server's queue or post a calculated "cost" on the "advert board" after successfully fulfilling a request. The executing server computes profitability of just-serviced virtual server and compares the calculated profit with the colony profit, registered on the "advert board". If the calculated profit was high, then the server returns to the current virtual service queue, else it returns to the idle/waiting behavior.

18.4 Research Directions

Data is a core entity for any business-driven application. As cloud encompasses a business model for an enterprise or organization, protecting its data is a major security challenge. Naturally, in such a multi-tenant infrastructure, mission-critical data for any organization is under threat of malicious activity. Moreover, the data owner does not have complete control on its own data once it is deployed in the cloud environment. Therefore, new security standards are to be devised to isolate data of one customer from another. As security is considered to be a shared responsibility between the provider and the customer, efficient methods are required to delegate some of the data management controls to the respective customers. Again, some management issues on provider site, such as, archiving of data depending on its frequency of usage, data life cycle management, accessibility of historical data, degree of redundancy to be maintained, disaster recovery strategies, etc. are need to be addressed.

The characteristics of most of the cloud based applications are similar to those of web services. Hence, security threats which are prevalent in web services, viz. *spoofing, phishing, man-in-the-middle, session hijacking, cross-site scripting* (*XSS*) etc. are highly applicable to cloud services. Some security management issues, such as, whether an infrastructure provider or a customer is responsible for the security of a customized application are still not resolved. Denial of service (DoS) attacks are easier in a multi-tenant environment and pose severe threat if not appropriately managed. Such an attack may disrupt normal operations of multiple organizations. This is catastrophic as far as business models for different organizations are concerned. Hence, prevention of DoS attacks launched through flooding, is of utmost importance. Thus, workload management to control cloud scalability in massively used datacenters is a pending security issue [36]. As cloud provides an illusion of "infinite compute, storage, and network" capabilities, any malicious user may use this for breaking cryptographic algorithms implemented by the provider. The preventive measures of the provider may affect other noble users as a part of "reputation fate-sharing", leading to shutting down of their instances. Hence, monitoring the usage and behavioral patterns of the customer and retaining sufficient privacy are other research challenges in the area of cloud application security.

18.5 Conclusion

Cloud computing is an emerging technological paradigm to which most of the enterprises are aiming for their business operations. It provides a number of advantages but also has few loopholes. Security is one such issue which is restricting most of the organizations from moving into the cloud. The present chapter focuses on the importance of making cloud secure and also elaborately discusses about data and application related security issues jeopardizing the cloud. The objective of the chapter is to provide a contemporary view of various published research works on data and application security in cloud computing. At the end of the chapter, some research directions in data and application security aspects of cloud have been identified. These directions may help the researchers to work on the focused areas of cloud security.

References

1. Mell, P., Grance, T.: The NIST Definition of Cloud Computing. Technical Report v15, US National Institute of Standards and Technology ITL Technical Report (2009) http://www.csrc.nist.gov
2. Krutz, R.L., Vines, R.D.: Cloud Security : A Comprehensive Guide to Secure Cloud Computing. Wiley, Indianapolis (2010) ISBN : 978-81-265-2809-7
3. Ramgovind, S., Eloff, M.M., Smith, E., Chakerian, S.: The Management of Security in Cloud Computing. In: Proceedings of the Information Security for South Asia (ISSA), IEEE Computer Society, pp. 1–7 (2010)
4. Xu, J.S., Huang, R.C., Huang, W.M., Yang, G.: Secure Document Service for Cloud Computing. In: CloudCom 2009. vol. 5931 of LNCS, pp. 541–546. Springer, Heidelberg (2009)
5. Zhou, W., Marczak, W.R., Tao, T., Zhang, Z., Sherr, M., Loo, B.T., Lee, I.: Towards secure cloud data management. Technical report, Department of Computer and Information Science, University of Pensylvania (2010) http://repository.upenn.edu/cis_reports/919
6. Wang, Q., Wang, C., Li, J., Ren, K., Lou., W.: Enabling public verifiability and data dynamics for storage security in cloud computing. In: European Symposium on Research in Computer Security. vol. 5789 of ESORICS '09., LNCS, pp. 355–370, Springer (2009)
7. Du, J., Wei, W., Gu, X., Yu, T.: Towards secure dataflow processing in open distributed systems. In: Proceedings of the 2009 ACM workshop on Scalable trusted computing (STC '09), pp. 67–72, Chicago, USA (2009)
8. Sun, Z., Shen, J.: A high performance peer to cloud and peer model augmented with hierarchical secure communications. J. Syst. Softw. **86**(7), 1790–1796 (2012)
9. Bowers, K.D., Juels, A., Oprea, A.: HAIL: A high-availability and integrity layer for cloud storage. In: Proceedings of the 16th ACM conference on Computer and communications security. CCS '09, pp. 187–198, New York, USA, ACM (2009)
10. Juels, A., Kaliski, B.: Pors: Proofs of retrievability for large files. In: ACM Conference on Computer and Communication Security. CCS '07, ACM Press (2007)
11. Shacham, H., Waters, B.: Compact proofs of retrievability. In: Proceedings of the ASIACRYPT 2008. vol. 5350 of ASIACRYPT 2008, LNCS, pp. 90–107, Springer (2008)
12. Shraer, A., Cachin, C., Cidon, A., Keidar, I., Michalevsky, Y., Shaket, D.: Venus: Verification for untrusted cloud storage. In: Proceedings of the 2010 ACM Cloud Computing Security Workshop (CCSW '10), pp. 19–29, Chicago, USA (2010)

13. Chen, B., Curtmola, R., Ateniese, G., Burns, R.: Remote data checking for network coding-based distributed storage systems. In: Proceedings of the 2010 ACM Cloud Computing Security Workshop (CCSW '10), pp. 31–42, Chicago, USA (2010)
14. Du, J., Wei, W., Gu, X., Yu, T.: RunTest: assuring integrity of dataflow processing in cloud computing infrastructures. In: Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS '10), pp. 293–304, Beijing, China (2010)
15. Kamara, S., Lauter, K.: Cryptographic cloud storage. In: Workshop on Real-Life Cryptographic Protocols and Standardization. RLCPS 2010, LNCS, Springer (2010)
16. Ateniese, G., Burns, R., Curtmola, R., Herring, J., Kissner, L., Peterson, Z., Song., D.: Provable data possession at untrusted stores. In: ACM Conference on Computer and Communication Security. CCS '07, ACM Press (2007)
17. Ateniese, G., Pietro, R.D., Mancini, L.V., Tsudik, G.: Scalable and efficient provable data possession. In: Proceedings of the 4th international conference on Security and Privacy in Communication Netowrks. SecureComm '08, pp. 1–10, ACM Press (2008)
18. Erway, C., Kupcu, A., Papamanthou, C., Tamassia, R.: Dynamic provable data possession. In: ACM conference on Computer and communications security. CCS '09, pp. 213–222, ACM Press (2009)
19. Popa, R.A., Lorch, J.R., Molnar, D., Wang, H.J., Zhuang, L.: Enabling security in cloud storage SLAs with cloudProof. In: Proceedings of the 2011 USENIX conference on USENIX annual technical conference. USENIXATC'11, pp. 31–31, Berkeley, USA, USENIX Association (2011)
20. Huang, K.Y., Luo, G.H., Yuan, S.M.: SSTreasury+: A secure and elastic cloud data encryption system. In: Proceedings of the Sixth International Conference on Genetic and Evolutionary Computing (ICGEC), pp. 518–521 (2012)
21. Wang, S., Agrawal, D., Abbadi, A.E.: HengHa: data harvesting detection on hidden databases. In: Proceedings of the 2010 ACM Cloud Computing Security Workshop (CCSW '10), pp. 59–64, Chicago, USA (2010)
22. Faatz, D., Pizette, L.: Information security in the clouds. Technical Report Case: 10–3208, System Engineering at Mitre (2010) http://www.mitre.org/work/tech_papers/2010/10_3208/
23. Liu, H.: A new form of DoS attack in a cloud and its avoidance mechanism. In: Proceedings of the 2010 ACM Cloud Computing Security Workshop (CCSW '10), pp. 65–75, Chicago, USA (2010)
24. Kupsch, J., Miller, B.P., Heymann, E., Cesar, E.: First principles vulnerability assessment. In: Proceedings of the ACM Cloud Computing Security Workshop (CCSW '10), Chicago, USA (2010) http://www.cs.wisc.edu/mist/papers/ccsw12sp-kupsch.pdf
25. Jensen, M., Schwenk, J., Gruschka, N., Iacono, L.L.: On technical security issues in cloud computing. In: Proceedings of the 2009 IEEE International Conference on Cloud, Computing. pp. 109–116 (2009)
26. Maggi, F., Zanero, S.: Rethinking security in a cloudy world. Technical Report 2010–11, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Italy (2010) http://home.dei.polimi.it/fmaggi/downloads/publications/2010/
27. Zargar, S.T., Takabi, H., Joshi, J.B.: Dcdidp: A distributed, collaborative, and data-driven intrusion detection and prevention framework for cloud computing environments. In: CollaborateCom 2011, pp. 332–341, IEEE (2012)
28. Akbarabadi, A., Zamani, M., Farahmandian, S., Zadeh, J.M., Mirhosseini, S.M.: An overview on methods to detect port scanning attacks in cloud computing. Environment **1**, 22–25 (2013)
29. Berger, S., Caceres, R., Pendarakis, D., Sailer, R., Valdez, E., Perez, R., Schildhauer, W., Srinivasan, D.: TVDc: Managing security in the trusted virtual datacenter. ACM SIGOPS Oper. Syst. Rev. **42**(1), 40–47 (2008). doi:10.1145/1341312.1341321
30. Berger, S., Caceres, R., Goldman, K., Pendarakis, D., Perez, R., Rao, J.R., Rom, E., Sailer, R., Schildhauer, W., Srinivasan, D., Tal, S., Valdez, E.: Security for the Cloud Infrastructure: Trusted Virtual Data Center Implementation. IBM Journal of Research and Development 53(4) (2009) 6:1–6:12.

31. Hao, F., Kodialam, M., Lakshman, T., Puttaswamy, K.: Protecting cloud data using dynamic inline fingerprint checks. In: INFOCOM, pp. 2877–2885, 2013 Proceedings IEEE. (2013)
32. Bohli, J.M., Gruschka, N., Jensen, M., Iacono, L., Marnau, N.: Security and privacy-enhancing multicloud architectures. IEEE Trans. Dependable Secure Comput. **10**(4), 212–224 (2013)
33. Csorba, M.J., Meling, H., Heegaard, P.E.: Ant system for service deployment in private and public clouds. In: Proceedings of the 2nd workshop on Bio-inspired algorithms for distributed systems, pp. 19–28, ACM (2010)
34. Shen, G., Zhang, Y.Q.: A shadow price guided genetic algorithm for energy aware task scheduling on cloud computers. In: Advances in Swarm Intelligence, pp. 522–529, Springer (2011)
35. Randles, M., Lamb, D., Taleb-Bendiab, A.: A comparative study into distributed load balancing algorithms for cloud computing. In: Advanced Information Networking and Applications Workshops (WAINA), 2010 IEEE 24th International Conference on, IEEE, pp.551–556 (2010)
36. Vaquero, L.M., Rodero-Merino, L., Moran, D.: Locking the sky: A survey on IaaS cloud security. Computing **91**(1), 93–118 (2011). doi:10.1007/s00607-010-0140-x

Chapter 19
Security Issues on Cloud Data Services

Nour Zawawi, Mohamed Hamdy El-Eliemy, Rania El-Gohary and Mohamed F. Tolba

Abstract In Cloud environments, resources are provided as services to endusers over the internet upon request. Resources' coordination in the Cloud enables users to reach their resources anywhere and anytime. Ensuring the security in Cloud environment plays an important role, as customers often store important information on Cloud storage services. These services are not always trusted by the data owners. Customers are wondering about the integrity and the availability of their data in the Cloud. Users need to save their data from outsider and insider attackers (i.e. attacker within service providers' coordination's). Moreover, any collateral damage or errors of Cloud services provider arises as a concern as well. Most of the vital security needs and issues regarding data Cloud services are mentioned in this chapter. The purpose of this chapter is to examine recent research related to data security and to address possible solutions. Research of employing uncommon security schemes into Cloud environments has received an increasing interest in the literature, although these schemes are neither mature nor rigid yet. This work aspires to promote the use of security protocols due to their ability to reduce security risks that affect users of data Cloud services.

N. Zawawi (✉) · M. H. El-Eliemy · R. El-Gohary · M. F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University,
El-Kalifa Al Ma'mounst., Abbasyia, Cairo 11565, Egypt
e-mail: nourzawawi@gmail.com

M. H. El-Eliemy
e-mail: m.hamdy@cis.asu.edu.eg

R. El-Gohary
e-mail: dr.raniaelgohary@fcis.asu.edu.eg

M. F. Tolba
e-mail: fahmytolba@gmail.com

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_19, © Springer-Verlag Berlin Heidelberg 2014

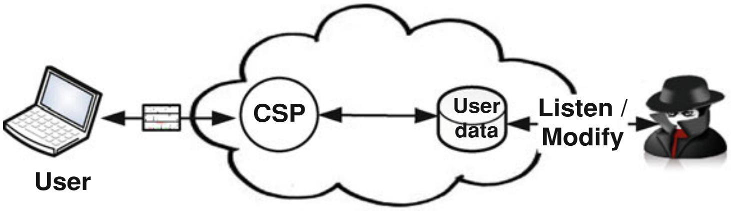


Fig. 19.1 Users' scenario for cloud environment

19.1 Introduction

In a public Cloud environment, resources and IT related capabilities are provided as services to the outer customers using the internet [1]. It depends on sharing information and computing resources instead of using local servers or personal devices to manage applications. It receives an increasing importance in commercial organizations. Moreover, it offers pay per use charge for the different required service. Essential characteristics of Cloud [1, 2] are on demand self service, broad network access, resource pooling, rapid elasticity, and measured service.

The prime revolutionary aspect of Cloud is its ability to deploy location independent services. At the same time, Service consumers (SCs) are no longer locked in with their providers. Cloud services take full advantage of the service oriented paradigm with a focus on the key attributes of statelessness, low coupling, modularity, and semantic interoperability [3, 4].

There are three, known types of Clouds: Infrastructure as a Service (IaaS), Platform as a service (PaaS) and Software as a service (SaaS). IaaS refers to the provision of virtualized hardware on which the client can run their operating system and software stack. In PaaS, the operating system and environment are provided and maintained for the client, who then runs their applications. In SaaS a Cloud Service Provider (CSP) runs and organizes the entire software system and provides software services [2].

Although users run their programs and applications depending on applications which have been physically deployed on their servers, reason for moving into Cloud computing is arising. It allows users to gain access applications from anywhere at any time through the internet. The CSP benefits are flexibility, disaster recovery, software updates automatically, pay per use model and cost reduction [5, 6].

Cloud computing still involves many risks concerning security, integrity, network dependency and centralization. Many security issues are considered based on the sensitivity and confidentiality of customers' data [6, 7]. Figure 19.1 represents the problems that prevent data owners of moving to depend on data services on the Cloud. The key issue of handling these challenges is empowering the trust between users and the service providers. Trust is the degree which clients will rely on for the assertions or security services provided by the cloud provider. Therefore, providing

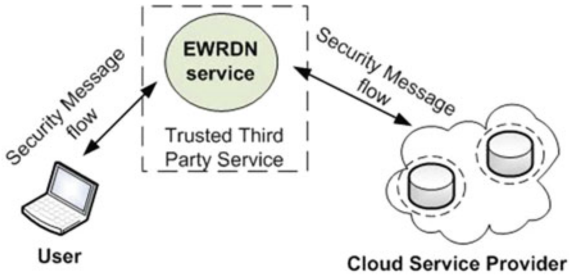


Fig. 19.2 EWRDN service proposed scenario

a trusted and secure data storage service in a public Cloud environment remains a challenge for service providers.

Availability, performance and security are the three main challenges when it comes to Cloud adoption. Nevertheless, performance needs to be measured according to time and space. In typical public cloud environments millions or even more simultaneous users are managed. This means that at full capacity, the system can handle these user and their data sets with minimal failures. The better an application's scalability, the more users it can handle simultaneously [8, 9]. Security in terms of integrity is the most important aspect of a Cloud environment. In this chapter, trusted security services which work against such security threats are illustrated. It achieves the missing trust enabling the parties in a Cloud environment to operate on their applications and services.

This chapter focuses more on the issues related to the data security and privacy aspects that may shape the trust in a public Cloud environment, such as data integrity, data confidentiality and service availability. As data and information will be shared with a third party, customers want to avoid unsafe service providers. Protecting private and important information, such as credit card details or a patient's medical records from attackers or malicious insiders has a critical importance. In previous work, several methods were proposed for securing data into the Cloud. This chapter discussed those methodologies and various techniques used to effectively and safely store data on the Cloud. An analysis for the advantages and drawbacks of those techniques is presented. Moreover, one discusses deeply the given aspects and criteria of one of the mature approaches that can handle efficiently these security issues. EWRDN service is [10] introduced as a trusted security service which tries to solve the previously mentioned scenario.

Figure 19.2 illustrates the proposed scenario of EWRDN service. EWRDN was built to be part of trusted third party service between user and CSP. EWRDN relies on changing the database schema by adding new columns. The function is used in constructing the new record as well as the secret key (K). In general, it combines some important features of database security and privacy like non repudiation, integrity, copyright protection and recovery. Moreover, it gives data owner more monitoring capabilities over their data. These features come from the missing trust between CSP and users.

EWRDN service uses watermarking techniques to prove if data has been tampered. By, saving data watermarks with users' original data. If the values match together, then data is tampered free. If not, then data owner has a legal evidence to prove that his data has been tampered with when it was at a CSP. Moreover, it provides a way to recover data, if unauthorized changes or errors happened. To the best of our knowledge, it is the first practical trusted Cloud privacy and copyright solution that solves previously mentioned lack of trust problems.

The rest of this chapter is organized as follows: Sect. 19.2 discusses the main security issues that affect in data over the Cloud. It has been divided into traditional and Cloud security challenges. It is followed by the main issues considered as the main security problems. Each section illustrates some of the recent research designed to overcome some security issues. Section 19.3 discusses data integrity over Cloud. Section 19.4 discusses data availability and Sect. 19.5 discusses data confidentiality. Section 19.6 illustrates the problems facing data security service over the Cloud with an introduction to a new service used to overcome the previous issues. Finally, Sect. 19.7 concludes the work discussed in this chapter.

19.2 Security Issues

Although CSP provides benefits to their clients, security risks play an important role resisting the development of any public cloud environment [11]. Users of online data sharing or network facilities are aware of the potential loss of privacy [12]. Protecting private and important information against attackers or malicious insiders is vital. So, an important question arises about the way to protect data from being exposed. In cloud scenarios there are three suspicious individuals for exposing data. They are:

- Cloud Service Provider (CSP)
- Internal Users
- Outsider Attacker

Therefore, how to save data from the three attackers is an important question that needs to be answered. Also, how to prove who has exposed data in order to take the related action needs to be answered. Moreover, there must be a well known technique to deal with data if errors or crashes appear.

Moving databases to large data centers involves many security challenges [9] such as accessibility vulnerability, and privacy. Also, there are control issues related to data accessed from a third party including data integrity, confidentiality, and data loss or theft.

In this section, one has categorized the main cloud security issues. This has been divided into two types; traditional and cloud (new) security challenges. Since, the common security challenges of traditional communication systems are inherited as well.

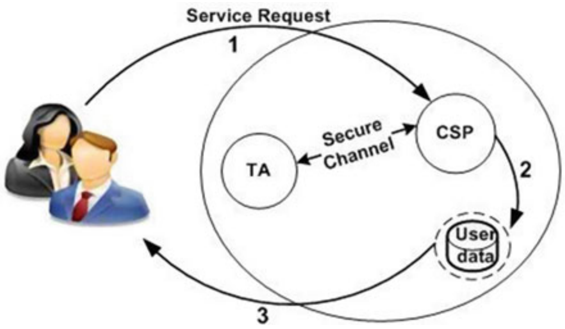
19.2.1 Traditional Security Challenges

Most businesses need some form of database security to protect confidential records and logical property from both external and internal threats. With violations of sensitive data, enhancement for security features appeared to upgrade the database security services. Consider a complete database security system to balance the security and privacy of your data with employee access. Developers recommend a database security model that sets controls to provide limited use. But because data needs to be stored, copied and made available instantly, database security remains a collaborative effort for companies.

Depending on your database security services database security measures can be set using various properties [13, 14]. They can be summarized as followed:

- Authorization: is finding out if the person, once identified, is permitted to have the resource. This is usually determined by finding out if that person is a part of a particular group, if that person has paid admission, or has a particular level of security clearance.
- Authentication: is any process by which you verify that someone is who they claim they are. This usually involves a username and a password, but can include any other method of signifying identity, such as a smart card, retina scan, voice recognition, or fingerprints.
- Confidentiality: it is the system policies that limits access or make restrictions on certain types of information. It could be explained as the way of protecting information from being exposed by unauthorized users. Organizations private data needed to be protected.
- Verification: The evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition. Where, it uses a digital signature to combine a public key with an identity. The certificate can be used to verify that a public key belongs to an individual. Also, it demonstrates the authenticity of a digital message or document.
- Encryption: is the process of transforming information (plaintext) using an algorithm (cipher) to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key.
- Integrity: refers to the consistency and accuracy of the data stored in a database. Data Integrity is based on how much accuracy of data, dependability of information and protection from unauthorized modification is required. Data only counts when it is correct, while tampered data prove to be costly.
- Backups: refers to the copying and archiving of computer data so it may be used to restore the original data after a data loss event. Backups have two distinct purposes: (1) Recover data after its loss where it is data deletion or corruption. Data loss can be a common experience of computer users. (2) Recover data from an earlier time, according to a user predefined data maintenance policy. It represents a simple form of disaster recovery, and should be part of a disaster recovery plan; in the same time backups should not be considered alone as a disaster recovery.

Fig. 19.3 Trust circle



19.2.1.1 Cloud Security Challenges

In order to, move data to the Cloud, data owner needs to trust CSP. So, trust circle is introduced in Fig. 19.3. It defines CSP identity adheres to by signing a business agreement, in order to support secure transactions among members. Also, it contains the way to access any service inside a public cloud environment. The circle of trust includes three parties. They are:

- Users who send requests into CSP coordination.
- Services which have effects on users' data.
- A third party service coordination which represents the Trusted Authority TA of all parties or participants.

The issues of establishing trust in the Cloud have been discussed severally in this chapter. Security and privacy are the two major concerns about using any of data exchange Cloud services. In the Cloud, virtualization lets user access computing power that exceeds that contained within their business infrastructure. To enter this virtual environment a user is required to transfer data throughout the Cloud [9]. It is concerned with protecting the confidentiality, integrity and availability of data regardless of the form the data may take [11].

There are currently many open problems in Cloud security that should be addressed by CSP in order to convince end users to use the technology. The most important concerns, in our view, are to guarantee that user data integrity and confidentiality are attained while they are stored in the Cloud systems. In a long, non transparent provider chain, it is difficult for an end user to even determine what security mechanisms are applied to data in the Cloud.

Other important security challenges are [9]:

- Loss of control, where users have no control over their private and personal data,
- Lack of trust (mechanisms), due to lack of Service Level Agreement (SLA) standards availability between users and providers,
- Multi-tenancy, which refers to a single instance of a software application serving multiple customers.

- Resources Location, end-users use the services provided by the CSP without knowing exactly where the resources for such services are located.
- System monitoring and logs, customers may request that CSP provide more monitoring and records for the customers' personnel data.
- Cloud standards, where standards are needed to achieve interoperability among clouds and to increase their strength and security.

There is number of key security elements that should be considered as an integral part of the Cloud application development and deployment process. In the meantime, there are a few technical issues like browser security, secure browser based authentication and Attacks on browser based Cloud authentication that needs to be built [15].

In the following, one presents a selection of security issues related to Cloud. Each issue has a major real-world measured impact [16].

- XML Signature: It is the way to save data against attacks for authentication or integrity. These types of attacks focus on the way to attack Simple Object Access Protocol (SOAP).
- Browser Security: The way to protect the users' computers. Since it used only for Input and Output. Also, it used for authentication and authorization of commands to the Cloud. So, CSPs need to develop some standards or a universal platform (standard Web browser).
- Cloud Integrity and Required Issues
- Flooding Attacks: The impact of such attacks is expected to be amplified drastically. This is due to the following types of attacks: direct and indirect denial of service

The next sections address three security factors that particularly affect Clouds, namely data integrity, confidentiality, and service availability.

19.3 Data Integrity

One of the most important issues related to trust and security risks is data integrity. Data is stored in the Cloud, as tuples, may suffer from damage during transition operations from or to the CSP. Moreover, data may be stored for several years. Data owners my do not have any mean to check if their data are similar to the form that they stored it initially. Customers are wondering about attacks on the integrity and the availability of their data in the Cloud from malicious insiders and outsiders, and from any collateral damage of Cloud services.

Cloud storage can be an attractive means of outsourcing the day to day management of data, but ultimately the responsibility and liability for that data falls on the company that owns the data, not the hosting provider. With this in mind, there are four points to understand

- The causes of data corruption,
- How much responsibility CSP holds,
- Best practices for utilizing Cloud storage safely,
- Methods and standards for monitoring the integrity of data.

The computing power to the Cloud environment is provided through a collection of data centers or cloud data storages (CDSs) present at different location and connected by high speed networks. With the emergence of cloud computing the CDSs is also emerging. The integrity within cloud storage consists of two techniques, integrity of data being transmitted from CDS and integrity of CDS. It faces an important issue that is security. Also, CDSs data integrity is an extremely important issue.

The work proposed by Rawat et al. [17] discusses the model based on Multi Agent Systems (MAS) architecture of Cloud and data encoding mechanism to enhance the integrity of CDSs. Where, MAS is used basically in artificial intelligence area for finding solution to the problems. It uses two agents in client side layer for data integrity. In Cloud they are used for developing architecture for data integrity at CDSs. Data encoding is one of the basic mechanisms of providing security. It combines MAS architecture for CDS and data encoding using hash values together to give a new mechanism. This can be done inserting a hash value concept in CDIBA agent of MAS architecture. CDIBA is responsible for the maintenance of cloud storage when data is entered into it, and if the data goes out of the cloud storage the hash values being used can be used to verify the data being transmitted is in correct format. Hence the complete process of reliable retrieval and reliable data transmission is guaranteed at the same time.

Motghare et al. [18] proposes a framework of data integrity. It uses Cooperative Provable data possession (CPDP). CPDP is a technique for ensuring the integrity of data in storage outsourcing. So, it addresses the construction of an efficient CPDP scheme and dynamic audit service for distributed Cloud storage as well verifying the integrity guarantee of an entrusted and outsourced storage which support the scalability of service and data migration. It offers two main contributions:

1. Efficiency and security: It is safer to rely on a public and private key encryption. In this every time parameters are generated and key exchange takes place this becomes more secure than symmetric and asymmetric algorithms. However, it is more efficient than the other techniques. Because it does not require lots of data encryption in outsourced and no additional posts on the symbol block, and the ratio is more secure.
2. Public verifiability: It provides public validation which allows users for information on the CSP has proved challenge(rewrite previous sentence). However, it is more efficient because it does not need the information for each block encryption.

Data corruption can happen at any level of storage and with any type of media. Bit rot (the weakening or loss of bits of data on storage media), controller failures, reduplication metadata corruption, and tape failures are all examples of different media types causing corruption. Metadata corruption can be the result of any of the vulnerabilities listed above, such as bit rot, but are also susceptible to software glitches outside of hardware error rates. Unfortunately, a side effect of reduplication is that a corrupted file, block, or byte affects every associated piece of data tied to that metadata. The truth is that data corruption can happen anywhere within a storage environment. Data can become corrupted simply by migrating it to a different platform, i.e. sending your data to the Cloud. Cloud storage systems are still data centers, with hardware and software, and are still vulnerable to data corruption.

Based on the case studies proposed for home healthcare applications, one has chosen the following studies. They ensure data integrity on users' private data. Home Healthcare system is presented in [19], monitors, diagnoses and assists people outside of hospital setting. Specifically, it focuses on using TClouds on depressed patients. Establishing trust in the Cloud is a big challenge that requires collaborative efforts from academia and industry. It builds on the previous work [20]. Establishing trust is a fundamental requirement especially for Cloud's potential future as an Internet scale critical infrastructure. This requires the following: a. understanding and defining such services and their interdependencies, b. defining functions of the services which help in establishing their trustworthiness and c. building protocols and prototyping based on the defined functions. Specifically, it starts by developing the functions (e.g. LaaS, ACaaS, and PRaaS). In parallel, it establishes trust protocols based on the identified middleware functions. Once these are done home healthcare applications are deployed on the Clouds' platform architectures.

Current health cloud services offerings require full trust to CSP, where threats of malicious insiders have become one of the most dangerous attacks to protected data and applications in the Cloud. The work presented by Deng et al. [21] proposed a design for a trustworthy healthcare platform as a service. It is built on top of a trusted Cloud infrastructure that addresses technical issues and provides users with control over their personal data. Next to that, the platform addresses the needs to further decrease development and porting costs, while supporting rapid development of healthcare applications. The healthcare platform is proposed to host numerous healthcare applications, and also provide storage for medical data such as personal health records. The underlying trustworthy Cloud infrastructure is designed to increase the level of trust both for storage and computing. Various techniques are employed to provision security, data protection and resilience against data center outage and Cloud network failures. Two benchmark applications are implemented as a proof of concept to demonstrate features of the proposed the health platform. Where, it provides major benefits for both end users and service providers.

The other case one needs to discuss is Software as a Service (SaaS) applications. It exploits the potential of elastic Cloud infrastructures naturally are enabling new ubiquitous access scenarios for nomadic users, such as market salesmen and home healthcare medical assistants. SaaS applications typically require transferring data and resources to the Cloud infrastructure site. It raises several challenging issues spanning from access control to privacy protection of resources, ownership, and security of the data of the final SaaS users. However, although encryption of personal and enterprise data is strongly recommended by existing Cloud infrastructures typically they do not provide yet adequate encryption and key management support. Corradi [22] presented a real use case of home healthcare SaaS application deployed on Amazon Web Service (AWS), and discusses the challenges and changes needed to add cryptography and key management capabilities to enable SaaS data protection. It shows experimental results that benchmark the new security functions over Amazon, demonstrating their applicability to SaaS production deployments.

The last case one discusses is the way to establish trust. Trust establishment in Clouds requires collaborative efforts from industry and academia. Abbadi and Alawneh [23], presented a framework for establishing trust in the Cloud. The framework uses the dynamic domain concept. It is composed of the following: Cloud Management Domain (MD), Cloud Collaborating Management Domain (CMD), Organization Outsourced Domain (OD), Organization Collaborating Outsourced Domain (COD), and Organization Home Domain (HD). But, the framework is not enough by itself, and requires further extensions as establishing trust in Clouds is a complex subject. Also, it discusses how the framework could be extended with their previous work on secure virtual infrastructure management. They have considered Clouds resources management and infrastructure properties and differentiated between the secure management of infrastructures data and user's applications data.

Li and Ping [24] analyzed several trust models used in large and distributed environment and then introduced a novel Cloud trust model to solve security issues in cross Clouds environment. Where, customer can choose different providers' services and resources in heterogeneous domains can cooperate. The model is domain based. It divides one CSP's resource nodes into the same domain and sets trust agent. It distinguishes two different roles customer and server. Then, it designs different strategies for them. In this model, trust recommendation is treated as one type of Cloud services just like computation or storage. The model achieves both identity authentication and behavior authentication. The results of emulation experiments show that the proposed model can efficiently and safely construct trust relationship in cross Cloud environment.

19.4 Data Availability

Whenever data is lost, especially valuable data, there is a propensity to scramble to assign blame. Often in the IT world, this can result in lost jobs, lost company revenue, and, in severe cases, business demise. As such, it is critical to understand how much legal responsibility the CSP, per the service level agreement (SLA), has and to ensure that every possible step has been taken to prevent data loss. As with many legal documents, SLAs are often written to the benefit of the provider, not to the customer. Many CSPs offer varying tiers of protection, but as with any storage provider they do not assume liability for the integrity of your data. Creating a trusted SLA for the Cloud that contains explicit statements if data is lost or corrupted is common practice is still under research.

The first work that identified the Cloud middleware services and their interdependencies focusing on application layer is presented in [20] and at Cloud virtual layer is presented in [25]. The work presented by Abbadi [25], is concerned about defining, exploring, and analyzing middleware self managed services at Cloud virtual layer. Most importantly it discusses the interdependency across such services in context of Cloud environment. It presents a conceptual model of self managed services and identifies the factors, which affect services' decisions. This model helps in understanding the required functions and their interdependency when providing self managed services in Cloud. Also, it helps in realizing the challenges involved in providing automated management functions. Finally, it discusses the challenges and requirements for managing and providing automated services security and privacy by design. It is considered as the first work to explore this area and especially discussing self managed services' interdependency.

The work presented by Abbadi [20], considered as the first work to identify Cloud middleware types focusing on application layer management services and their interdependencies. Where, establishing trustworthy Cloud infrastructure is the key factor to move critical resources into the Cloud. In order to move in this direction for such a complex infrastructure, we virtually split Cloud infrastructure into layers, each layer relies on the services and resources provided by the layer directly underneath it, and each layer services' rely on messages communicated with both the layer directly underneath it and above it. Each two adjacent layers have a specific middleware that provides self-managed services. These services' implementations are based on the layer they serve. Also, different types of middleware services coordinate amongst themselves and exchange critical messages. Then, it demonstrates services interdependencies and interactions using multitier application architecture in Cloud context. Then, it discusses the advantages of middleware services for establishing trust in the Cloud.

The work of Bajpai et al. [26] proposed an authentication and authorization interface to access a Cloud service. The proposed model explains the messages involved in the process of authenticating employees of an enterprise and providing them access of the distributed Cloud services. The trust is established between the end user and the service provider through the authentication and authorization interface. Access control rights of a user for a particular service are considered before granting the service access to the user of that service. Also, to make the system more securely intact, the access rights are not shared with the authentication and authorization interface. Service selection is acquired via monitoring security measures provided by a service provider through Security Service Level Agreements (Sec-SLAs). Security measures are considered while referring the service to an end user in order to provide an end user with a more efficient Cloud service. Denial of service attack, man in the middle attack and robustness of the system are efficiently handled by this methodology that overcomes the drawbacks of previously defined models. In the initial authentication step the enterprise handles the security measures provided by CSP. So, it relieves the end user from up to 80 % of the basics of CSPs in subsequent phases as compared to the models proposed in the past that consider the handling of security measures through end users.

While, Ko et al. [27] establish the urgent need for research in accountability in the Cloud, and outline the risks of not achieving it. By, proposing new approaches in order to increase data accountability. Two trusting approaches have been introduced;

detective and preventive. Detective approach used to identify the occurrence of a privacy or security risk that goes against the privacy or security policies and procedures. While, preventive used to mitigate the occurrence of an action from continuing or taking place at all. Detective approaches complement preventive approaches as they enable the investigation not only of external risks, but also risks from within the CSP. Detective approaches can also, be applied in a less invasive manner than preventive approaches. Also, it presents the trust Cloud framework, which addresses accountability in Cloud via technical and policy based approaches. Where, it can be used to give Cloud users a single point of view for accountability of the CSP.

Campbell et al. [28] proposed the properties and building blocks of a middleware that assured Cloud can support critical missions, where the middleware must include sophisticated monitoring, assessment of policies, and response to manage the configuration and management of trusted resources. Specifically, it considers applications in which assigned tasks or duties are performed in accordance with an intended purpose or plan in order to accomplish an assured mission. Mission critical Cloud may possibly involve hybrid (public, private, heterogeneous) Clouds and require the realization of end to end and cross layered security, dependability, and timeliness. It proposed the properties and building blocks of a middleware to support critical missions. In this approach, it assumed that mission critical Cloud must be designed with assurance in mind. In particular, the middleware in such systems must include sophisticated monitoring, assessment of policies, and response to manage the configuration and management of dynamic systems of systems with both trusted and partially trusted resources (data, sensors, networks, computers, etc.) and services sourced from multiple organizations.

19.5 Data Confidentiality

Another security risk that may occur with a CSP is data confidentiality. It is one of the main concerns for users of public Cloud services. The work proposed by Arasu et al. [29] discusses the main problem of protecting sensitive key data from being accessed by Cloud administrators who have root privileges and can remotely inspect the memory and disk contents of the Cloud servers. While encryption is the basic mechanism that can leverage to provide data confidentiality, providing an efficient database as a service that can run on encrypted data raises several interesting challenges. It outlines the functionality of Cipher base. It has a novel architecture that tightly integrates custom designed trusted hardware for performing operations on encrypted data securely such that an administrator cannot get access to any plaintext corresponding to sensitive data.

The work proposed by Yonghong [30], provides secure database service, but it needs to integrate many security techniques, such as data access control, network transformation control, database queries and privacy protection. It focuses on privacy protection in distributed secure database service architecture, and proposes a new security model which can use the set of attributes consisting of a quasi identifier to partition data to different database system to achieve privacy protection. The theoretical analysis and experimental results show that the new method is feasible and provide privacy protection and query execution efficiently, and supports horizontal fragmentation and semantic attribute decomposition. Moreover, it proposes an automatic attribute detection partition method and a new security model which partitions data in unencrypted form to distributed secure database servers.

The work proposed by Itani et al. [31] proposed a PasS (Privacy as a Service); a set of security protocols for ensuring the privacy and legal compliance of customer data in Cloud architectures. The security solution relies on secure cryptographic coprocessors for providing a trusted and isolated execution environment in the computing Cloud. It discussed the PasS protocols and described the privacy enforcement mechanisms supported by them. Also, it presented a description of a proof of concept implementation of the privacy protocols. It allows for the secure storage and processing of users; confidential data by leveraging the tamper proof capabilities of cryptographic coprocessors.

Ranchal et al. [32], illustrates an approach for Identity Management with the ability to use identity data on untrusted hosts. The approach is based on the use of predicates over encrypted data and multiparty computing for negotiating a use of a Cloud service. It uses active bundle which is a middleware agent that includes PII data, privacy policies, and a virtual machine that enforces the policies, and has a set of protection mechanisms to protect it. An active bundle interacts on behalf of a user to authenticate to Cloud services using user's privacy policies.

The usage of a Trusted Platform Management (TPM) is to establish trust in the Cloud and provide remote attestation is proposed in [33, 34]. Wang et al. [34] approach combines the public key based homomorphic authenticator with random masking to achieve the privacy preserving for a public Cloud data auditing system. It happened by proposing a privacy preserving public auditing system for data storage security in Cloud environment. It utilize the homomorphic authenticator and random masking to guarantee that Trusted Privacy Auditing (TPA) would not learn any knowledge about the data content stored on the Cloud server during the efficient auditing process, which not only eliminates the burden of Cloud user from the tedious and possibly expensive auditing task, but also alleviates the users' fear of their outsourced data leakage. Considering TPA may concurrently handle multiple audit sessions from different users for their outsourced data files. To support efficient handling of multiple auditing tasks, it explores the technique of bilinear aggregate signature to extend main result into a multi user setting, where, can perform multiple auditing tasks simultaneously. Trusted Cloud proposals generally assert that the Trusted Computing Base (TCB) of the Cloud should be clearly defined and attested to. Extensive analysis shows that the proposed schemes are provably secure and highly efficient.

However, specific characteristics of trust in the Cloud make such solutions difficult to implement in an effective and practical way. The work by Ruan and Martin [33] establishes trust between Cloud entities based on their dynamic behavior which is not accurate and might affect the availability of CSP and their resilience. It presents RepCloud, a reputation system for managing decentralized attestation metrics in the

Cloud. Moreover, it finds that as trust evidence generated by the Trusted Computing Group (TCG) framework can be efficiently transmitted within the Cloud. In a web of nodes with high connectivity and mutual attestation frequency, corrupted nodes can be identified effectively. By modeling this web with RepCloud, it achieved a fine grained Cloud TCB attestation scheme with high confidence for trust. Cloud users can determine the security properties of the exact nodes that may affect the genuine functionalities of their applications, without obtaining much internal information of the Cloud. Also, it showed that as achieving fine grained attestation RepCloud still incurred lower trust management overhead than existing trusted Cloud proposals. Aradhana and Chana [35] determine process for managing trust with specifying trust policies for different Cloud scenarios. Where, trust policies arerepresented in the form of a decision table that helps in the implementation of these policies.

Sato et al. [36] work proposed a trust model to secure Cloud. It proposed a new Cloud trust model. In addition to conventional trust models, it considers both internal trust, and contracted trust that controls CSP. It calls the Cloud platform that meets the Cloud trust model as Security Aware Cloud. In a security aware Cloud, internal trust must be established as the firm base of trust. By implementing TPM of security such as Id management and key management on internal trust, we obtain a firm trust model. Moreover, by controlling levels of quality of service and security by contract, one can optimize Return on Investment (ROI) on service and security delegated to a Cloud.

19.6 Security and Trust Cloud Data Service

In this section, a security and trust service alternative for public Cloud environment is presented in more details. This alternative, EWRDN, can easily and mostly meet all of the previously mentioned security issues in public Cloud environment. Most of Cloud security techniques aim mainly at protecting the data from being altered or viewed. Due to, the nature of the Cloud, where users have no authority over their private data; there are no guarantees to prove integrity of data. That is one reason why organizations do not trust the data services over their private and sensitive data. So, an important question arises about the way to protect data from being showing.

Therefore, how to saves data from attackers with the ability to prove which one has exposed data is a very important point. Also, an agreement about the techniques used to deal with data in case of errors or crashes happened.

19.6.1 EWRDN Service Model

A Novel Watermarking Approach for Data Integrity and Non Repudiation in Rational Databases (WRDN) is introduced [37]. It prevents the impacts of tampering dataset and localizing any changes made. By giving the database owner more control over his

Table 19.1 Difference between WRDN and EWRDN model

	WRDN	EWRDN
Granularity level	Tuple level	Tuple and attribute level
Is a part of	A relational database management system or a database engine	A service and be involved in a trusted authority service coordination
Verifiability	Blind, private	Blind, private, deterministic
Intent	Ownership prove, Data integrity	Ownership prove, data integrity, Tamper detection
Backup	Does not allow	Allow
Trace users activity	Trace some activity (add or modify)	Data owner has the ability to trace all users activity
Eliminated attacks	Insertion and deletion	Insertion, deletion, and alternation
Overhead space	Depends on function used	Depends on the required data quality level and the compression used

data. Besides, it concentrates on proving the data integrity and copyright protection of database against any type of attack. The main idea is to apply WRDN as a trusted security service on cloud computing. But some problems arise. These problems can be summarized in hiding and locking technique. Moreover, one needs to have the ability to recover data if unauthorized changes or errors appeared.

WRDN proves the data ownership and integrity of database. It survives against two types of attacks that face the database (Insertion, Deletion). It is based on adding a watermark over a hidden column then locks this column. It is designed to be a part of Database Management System (DBMS). So, there are no fears over watermark data detection. The main idea of this chapter is creating a data security service over the Cloud. Unfortunately, applying WRDN directly to be a Cloud service is not feasible due to the following reasons: Over the Cloud, there are no guarantees that a CSP will apply the hidden mechanism over the watermark tuple. At the same time, data and the users could be not in the same country.

Therefore, CSP will also have the authority and the ability to unlock and view the watermark column. Moreover, the system will fail enormously to prove changed attributes or to recover data to its origin. To overcome these problems, some enhancement of WRDN model was made. An Enhancement model for WRDN (EWRDN) is presented [10]. It does not prevent copying, but it deters illegal copying by providing a means of establishing the ownership of a redistributed copy. Table 19.1 summarizes the difference between WRDN and EWRDN.

Therefore, the framework is composed to provide security to the data throughout the entire process of Cloud service coordination at a Trusted Third Party (TTP). Figure 19.4 illustrates EWRDN service architecture. First users need to send tuples to EWRDN service. It works as follows:

1. EWRDN service calculates watermark value for each tuple then sends result to next step.

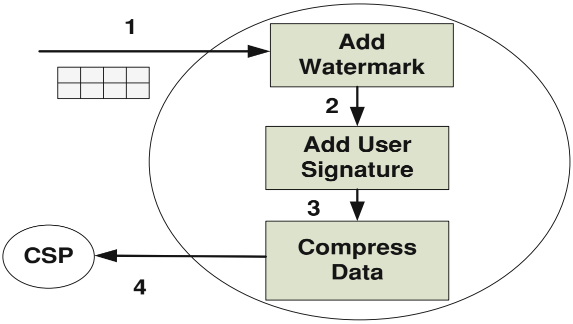


Fig. 19.4 EWRDN service architecture

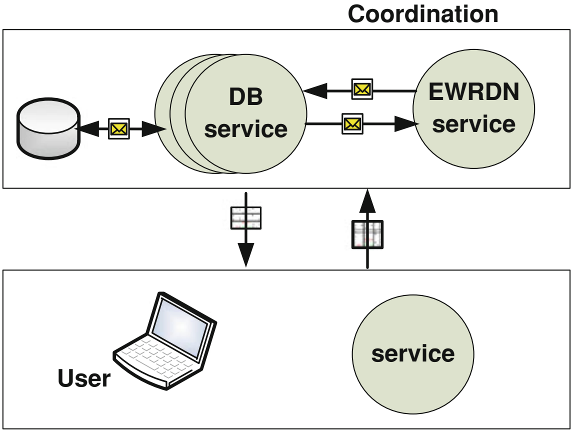


Fig. 19.5 EWRDN service coordination

2. EWRDN service as part of TA service checks over user authority. It adds users Signature using private key (PrK) over each attribute.
3. It compresses signed data, and then saves a copy into operational registry data into service.
4. Send original tuples with watermark value to CSP.

EWRDN service requires the transmission of multiple messages. The challenge lies in coordinating these messages in sequence where the actions performed by the message are executed properly with configuration of overall task. These messages are called Message exchange path (MEP). MEP represents a set of models that provide a group of sequences for the exchange of messages. The more complex an activity, the more context information it will bring. Every activity introduces a level of context into an application runtime environment. Something executing has meaning during its lifetime, and the description can be classified as context. So, a framework is required to provide a means for context information in complex activities to be managed, updated, and distributed to activity participants. Coordination which establishes such a framework is shown in Fig. 19.5.

19.6.2 EWRDN Utilization of Resources

EWRDN does not prevent copying, but it deters illegal copying by providing a means of establishing the ownership of a redistributed copy that can form a trust mean. EWRDN service prevents the worse impacts of tampering data set by localizing any changes made. Besides, it has the ability to recover data to its origin if any changes appear which gives the database owner more control over his data. It uses the compression technique to save tuples and recover them if needed. The data is prevented from any type of attacks by tracing users work to recognize authorization from unauthorized users.

To determine how the compressed data fit, and the estimated disk unit capacity, one needs to apply the following equation

$$\text{Capacity} = \text{Logical\_Data} + (2 \times \text{Free\_Space}) \tag{19.1}$$

The previous equation needs to be calculated for the users when using EWRDN service over the Cloud. That is because more space in the Cloud means extra money. Meanwhile, the previous equation covers space calculation needed to move data over to the Cloud. This proves that in order to; calculate original disk capacity one needs to learn about space of both actual data size and free space on disk. The test experiments made on EWRDN service prove that it consumed less space than WRDN. Where, it has nearly a fixed value of overhead space in EWRDN.

Space complexity depends on compressed values and compression ratio. It has a complexity of $O(1+\beta)$x n where β is the compressed value between [0,1] and n is the number of attribute. But, β differ according to the importance of data.

$$\text{Space Saving} = 1 - \frac{\text{Compressed Size}}{\text{Uncompressed Size}} \tag{19.2}$$

It has been proven that the type of compression used affect the quality of data. There are two types of compression techniques: Lossless Compression and Loose Compression. Lossless compression technique saves more space than loose compression techniques. Equation 19.2 proves the previous assumptions. It will be found that lossless compression techniques save space of 50 % of the original data size. Loose compression techniques increased the saved space to 60 % of the original data size.

EWRDN service proves data integrity by calculating watermarking data. It adds a watermark with original tuples. Then, it sends original tuples with watermark data to CSP. Also, it saves a copy of watermark value inside operational registry. Where, it gives data owner the ability to prove integrity and ownership of data anytime. Furthermore, it is built to be part of the TTP service coordination. Service Level Agreement (SLA) is made between users and CSP in order to, save data from being missing. Moreover, it adds a user's signature over each attribute being sent to Cloud. It proves data confidentiality and gives data owner the ability to differentiate authorized from unauthorized users. The main goal behind building EWRDN service is establishing trust in tracing the authorized and unauthorized changes in data services. Data privacy is considered as one of the biggest concerns affecting data storage over the Cloud. Cloud service providers need to find ways to prove and have tools that enable clients to trust their data exchange and storage. Moreover, the service providers need to differentiate between the quality levels provided to the different data owners and users. A data owner needs to track the data at all times with the ability to define errors and data failures, if any, and recover data to its origin. Also, owners need to check over any malicious modification and minimize the effects of server attacks or failures. Moreover, they need to have the same level of assertion every time they operate on the data. At the same time, concurrent users need to log on over the data each time with minimum overhead.

19.7 Conclusion

Clearly, as the use of Cloud environment has rapidly increased, security is still considered the major issue in the Cloud environment. Where, Cloud has become the future technology. Keeping this view in mind, this chapter has attempted to discuss the issues connected with data security. One of the most important issues related to trust and security challenges in Cloud environments is data integrity. Data may suffer from damage during transition operations from or to the CSP. Data owners are wondering about attacks on the integrity and the availability of their data in the Cloud from malicious insiders and outsiders, and from any collateral damage of Cloud services.

It is critical to understand how much legal responsibility the CSPs has and to ensure that all securing procedures have been applied to prevent data deformation, quality and other security threats. Moreover, this can be contacted between clients and CPS in means of a Service Level Agreement (SLA). Creating a trusted SLA for the Cloud that contains clear statements in cases of data loss is a common practice that has an increasing interest of research. A user's privacy and confidentiality risks differ with regards to the terms of service and privacy policy established by the CSP. The location of information in the Cloud effects the data confidentiality protection. The loss of service availability has caused many problems for many customers. Furthermore, data intrusion leads to many problems for the users of the Cloud.

The purpose of this work is to survey the recent research to address the security risks and solutions.

Most of Cloud security techniques aim mainly at protecting the data from being altered or viewed. Users have no authority over their private data; there are no guarantees to prove integrity of data. That is one reason why organizations do not trust the data services over their private and sensitive data. So, an important question arises about the way to protect data from showing. Therefore, how to save data from attackers with the ability to prove which one has exposed data is a very important point.

Also, an agreement about the techniques used to deal with data in case of errors or crashes.

EWRDN service is presented as one of the effective approaches that can handle easily the mentioned trust and security threats in public Cloud environment. It is based on some enhancement made on WRDN technique. It guarantees data integrity, privacy, and non repudiation with the ability to recover data to its origin. The main goal behind building EWRDN service is establishing trust in tracing the authorized and unauthorized changes in data services. It is built to prove data integrity by calculating watermark values. It provides monitoring capabilities between clients and Cloud Service Provider (CSP). Also, it gives users more control over their data by, tracing authorized users activity over database. It proves data confidentiality by adding a user's signature over each attribute. It gives data owner the ability to differentiate authorized from unauthorized users.

References

1. Grandison, T., Maximilien, E.M., Thorpe, S.S.E., Alba, A.: Towards a formal definition of a computing Cloud. In: 6th WorldCongress on Services, SERVICES, pp. 191–192. IEEE Computer Society (2010)
2. Mell, P., Grance, T.: The NIST definition of Cloud Computing. National Institute of Standards, USA (2011)
3. Fensel, D., Facca, F.M., Simperl, E., Toma, I.: Service science. In: Semantic Web Services, chapter 3, pp. 25–35. Springer (2011)
4. Papazoglou, M.P.: Service-oriented computing: concepts, characteristicsand directions. In: Fourth International Conference on Web InformationSystems Engineering, WISE 2003, pp. 3–12. IEEE Computer Society (2003)
5. Carroll, M., Kotzfie, P., van der Merwe, A.: Secure Cloud computing: benefits, risks and controls. In: Information Security SouthAfrica, ISSA, pp. 1–9. IEEE Computer Society (2011)
6. Modi, C., Patel, D., Borisaniya, B., Patel, A., Rajarajan, M.: A survey on security issues and solutions at differentlayers of Cloud computing. J. Supercomput. **63**(2), 561–592 (2013)
7. Zissis, D., Lekkas, D.: Addressing Cloud computing security issues. Future Gener. Comput. Syst. **28**, 583–592 (2012)
8. Modi, C., Patel, D., Borisaniya, B., Patel, A., Rajarajan, M.: A survey on security issues and solutions at differentlayers of Cloud computing. J. Supercomput. **63**(2), 561–592 (2013)
9. Rong, C., Nguyen, S.T., Jaatun, M.G.: Beyond lightning: a survey on security challenges in Cloud computing. Comput. Electr. Eng. **39**(1), 47–54 (2013)
10. El-Zawawi, N., Hamdy, M., El-Gohary, R., Tolba, M.F.: A database watermarking service with a trusted authority architecture for Cloud environment. Int. J. Comput. Appl. **69**(13), 1–9 (2013)
11. Ryan, M.D.: Cloud computing security: the scientific challenge, and a survey of solutions. J. Syst. Softw. **86**(9), 2263–2268 (2013)
12. Pearson, S., Yee, G.: Privacy and Security for Cloud Computing. Springer, London (2013)
13. Elmasri, R.: Fundamental of database systems, 6th edn. (chapter 24). Pearson Education, London (2011)
14. Imran, S., Hyder, I.: Security issues in databases. In: International Conference on Future Information Technology and ManagementEngineering. IEEE Computer Society (2009)
15. DhineshBabu, L.D., Venkata Krishna, P., Mohammed Zayan, A., Panda, V.: An analysis of security related issues in Cloud computing. In: 4th International Conference Contemporary Computing, IC3. Springer (2011)

16. Jensen, M., Schwenk, J., Gruschka, N., Iacono, L.L.: On technical security issues in cloud computing. In: IEEE International Conference on Cloud Computing. IEEE Computer Society (2009)
17. Rawat, S., Chowdhary, R., Dr. Bansal, A.: Data integrity of cloud data storages (cdss) in cloud. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(3), 588–592 (2013)
18. Motghare, S., Mohod, P.S., Khandait, S.P., Jaiswal, A.: Framework of data integrity for cross cloud environment using cpdp scheme. Int. J. Adv. Res. Comput. Sci. **4**(4), 55–59 (2013)
19. Abbadi, I.M., Deng, M., Nalin, M., Martin, A., Petkovic, M., Baroni, I., Sanna, A.: Trustworthy middleware services in the Cloud. In: Third International Workshop on Cloud Data Management, CloudDB '11, pp. 33–40. ACM Digital Library (2011)
20. Abbadi, I.M.: Middleware services at Cloud application layer. In: First International Conference Advances in Computing and Communications (ACC 2011), vol. 193 of CCIS, pp. 557–571 (2011)
21. Petkovic, M., Baroni, I., Deng, M., Nalin, M., Marco, A.: Towards trustworthy health platform Cloud. In: SecureData Management, vol. 7482 of SDM, pp. 162–175. Springer (2012)
22. Corradi, A.: Database security management for healthcare SAAS in theamazon awsCloud. In: IEEE Symposium on Computers and Communications, ISCC '12, pp. 812–819. IEEE Computer Society (2012)
23. Abbadi, I.M., Alawneh, M.: A framework for establishingtrust in the Cloud. Comput. Electr. Eng. **38**(5), 1073–1087 (2012)
24. Li, W., Ping, L.: Trust model to enhance security and interoperability of Cloud environment. In: First International Conference, CloudCom, vol. 5931, pp. 69–79. Springer (2009)
25. Abbadi, I.M.: Middleware services at Cloud virtual layer. In: 11th IEEE International Conference on Computer and Information Technology, CIT, pp. 115–120. IEEE Computer Society (2011)
26. Bajpai, D., Vardhan, M., Kushwaha, D.S.: Authentication and authorization interface using security service level agreements for accessing Cloud services. In: 5th International Conference onContemporary Computing-IC3, volume 306 of Communications in Computer and Information Science, pp. 370–382. Springer (2012)
27. Ko, R.K.L., Jagadpramana, P., Mowbray, M., Pearson, S., Kirchberg, M., Liang, Q., Lee, B.S.: TrustCloud: a framework for accountability and trust in Cloud computing. In: IEEE World Congress on Services (SERVICES). IEEE Computer Society (2011)
28. Campbell, R.H. Montanari, M., Farivar, R.: A middleware for assured Clouds. J. Internet Serv. Appl. **3**, 87–94 (2012)
29. Arasu, A., Blanas, S., Eguro, K., Joglekar, M., Kaushik, R., Kossmann, D., Ramamurthy, R., Upadhyaya, P., Venkatesan, R.: Secure database as a service with cipherbase. In: ACM SIGMOD International Conference on Management of Data, SIGMOD '13, pp. 1033–1036. ACM Digital Library (2013)
30. Yonghong, Y.: Privacy protection in secure database service. In: International Conference on Networks Security, Wireless Communications and Trusted Computing. IEEE Computer Society (2010)
31. Itani, W., Kayssi, A., Chehab, A.: Privacy as a service: privacy-aware data storage and processing in Cloud computing architectures. In: IEEE International Conference on Dependable, Autonomic andSecure Computing (2009)
32. Ranchal, R., Bhargava, B., Othmane, L.B., Lilien, L., Kim, A., Kang, M., Linderman, M.: Protection of identity information in Cloud computing without trusted third party. In: IEEE Symposiumon Reliable Distributed Systems, SRDS, pp. 368–372 (2010)
33. Ruan, A., Martin, A.: RepCloud: achieving fine-grained Cloud TCB attestation with reputation systems. In: The sixth ACM Workshop onScalable Trusted Computing, STC '11, pp. 3–14. ACM Digital Library (2011)
34. Wang, C., Wang, Q., Ren, K., Lou, W.: Privacy-preservingpublic auditing for data storage security in Cloud computing. In: IEEEINFOCOM, pp. 1–9 (2010)
35. Aradhana, M., Chana, I.: Developing trust policies for Cloud scenarios. In: 2nd International Conference on Computer and Communication Technology, ICCCT, pp. 389–393. IEEE Computer Society (2011)

36. Sato, H., Kanai, A., Tanimoto, S.: A Cloud trust model in a security aware Cloud. In: 10th IEEE/IPSJ International Symposium on Applications andthe Internet (SAINT), pp. 121–124 (2010)
37. Zawawi, N., El-Gohary, R., Hamdy, M., Tolba, M.F.: A novel watermarking approach for data integrity and non-repudiation in rational databases. In: Advanced Machine Learning Technologies and Applications, vol. 322 of Communications in Computerand Information Science, pp 532–542. Springer, Heidelberg (2012)

Chapter 20
A Reputation Trust Management System for Ad-Hoc Mobile Clouds

Ahmed Hammam and Samah Senbel

Abstract Most current cloud systems involves a data center model, in which clusters of machines dedicated to run cloud infrastructure software. Ad-hoc cloud model, in which infrastructure software distributed over resources harvested from machines already existed and used for other purposes, are gaining popularity. And as a try to utilize mobile devices power an ad-hoc mobile clouds model introduced. In this chapter, a trust management system (TMC) for mobile ad-hoc clouds is proposed. This system considers availability, neighbors? evaluation and response quality and task completeness in calculating the trust value for a node.

In the last few years mobile devices in addition of its spread and its main advantage which is the mobility it became more powerful in terms of processing power and memory, which encourage researchers to try to utilize mobile devices in many fields.

One of the research areas was Mobile Ad-hoc Network (MANET) which is used mainly in military operation. It made up of mobile devices shares specific tasks. Mobile devices connected through a network which continuously changing its topography with no fixed data routes from one node to another because of the node is moving from one position to another.

As a parallel research area Ad-Hoc clouds was emerged, Ad-Hoc Clouds provides a new model of cloud computing. In normal cloud computing provisioning involves data center model, in which clusters of machines are dedicated to running cloud infrastructure software. In Ad-hoc cloud model infrastructure software is distributed over resources harvested from machines already existed and used for other purposes.

By the time mobile Ad-Hoc clouds evolved by merging MANET and Ad-Hoc clouds, Then new model of Mobile Ad-Hoc cloud computing is based on harvesting

A. Hammam (✉) · S. Senbel
College of Computing and Information Technology, AASTMT, Cairo, Egypt
e-mail: ahmed@hammam.me

S. Senbel
e-mail: senbel@aast.edu

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_20, © Springer-Verlag Berlin Heidelberg 2014

resources from mobile devices to take advantage of these devices like computing power and mobility. Because of the similarity between mobile Ad-Hoc clouds and P2P systems the need of having trust management systems becomes obviously perceptible. In this chapter, a reputation trust management system (TMC) for mobile ad-hoc clouds is proposed.

TMC system considers availability, neighbors evaluation and response quality and task completeness in calculating the trust value for a node. The trust management system is built over PlanetCloud which introduced the term of ubiquitous computing. EigenTrust algorithm is used to calculate the reputation trust value for nodes. Finally, performance tests were executed to prove the efficiency of the proposed TMC in term of execution time, and detecting node behavior.

20.1 Introduction

Cloud computing is a computing paradigm, where a large pool of systems are connected in private or public networks, to provide dynamically scalable infrastructure architecture for application, data and file storage. With the advent of this technology, the cost of computation, application hosting, content storage and delivery is reduced significantly.

This architecture opened the door to new applications and solutions and a new model that provides everything as a service "XAAS". Generally, cloud computing is based on the existence of huge data-centers, which need INTERNET connection to reach its services. But, there are some applications which require scalable computing in absence of INTERNET connection, such as in natural crises, to reach computing resources, which led to the ad-hoc clouds.

In ad-hoc cloud resources are harvested from machines already in existence within a network. And the increase in computing power of mobile devices led to the use of mobile ad-hoc clouds. In this type of architecture, resources are owned by users who cannot be trusted and may be malicious. To solve this problem, trust management systems were developed.

A new architecture was introduced that forms ad-hoc mobile clouds [1, 2]. This architecture is based on a spatio-temporal calendering mechanism that automatically adjusts resources to each cloud. Client agents' calendars is stored in resource servers RS which allows them to pick suitable members to form clouds. However, cloud agents need to verify that participants are reliable, available, and not malicious.

The main objective in this chapter is to avoid malicious participations in the clouds by building a reputation trust management system that will be used to select client agents with high reputation values. Which will affect the performance of the cloud positively.

To simulate the proposed reputation Trust management system, the P2P Trust Simulator [3] was used to simulate the transactions between cloud agent and client agents and to calculate trust values. The Sect. 20.2 provides an overview of the related work in this field of research. In Sect. 20.3, a brief description of the PlanetCloud [4] system is given. In Sect. 20.4, the proposed trust system for ad-hoc mobile clouds TMC is described. Section 20.5 describes the system performance test and displays the results that were obtained, and Sect. 20.6 concludes this work.

20.2 Related Work

Ad-Hoc cloud is a model that was introduced for cloud [5], in which infrastructure software is distributed over resources that are harvested from machines that are already in existence within an enterprise. In contrast to the data center cloud model that in which resources dedicated exclusively to the cloud. Ad-hoc cloud allows partial virtualization of non-dedicated hardware based in distributed voluntary resources. Ad-Hoc cloud utilize already existing and unused resources within the network but it opens the door to malicious resource providers to affect the performance of the cloud.

Mobile Cloud Computing (MCC) was defined in surveys [6, 7] as a composition of mobile technology and cloud computing infrastructure where data and the related processing takes place in the cloud and then can be accessed through a mobile device. Or in which mobile device is being part of the cloud by voluntarily provides its resource to the cloud provisioning system. MCC inherits the issues that was emerged in the ad-hoc clouds model but it benefits from the increasing power of the mobile devices and wireless networks.

SETI@home [8] is an Internet based public volunteer computing project built originally for an ambitious space program at the University of California. As a development of SETI@HOME concept Cloud@Home [4] was introduced. It is a new Cloud infrastructure Architecture in which Commercial/business and the volunteer/scientific viewpoints coexist. This infrastructure is able to provide adequate resources to satisfy user requests also taking into account QoS requirements. This Architecture was built on resources voluntarily by their owners or administrators, following a volunteer computing approach and provided to users through a cloud-service interface.

Mobile ad-hoc networks MANET is a group of mobile devices that are connected through wireless link. These nodes are Self-organization which means they are autonomous, with no fixed infrastructure or centralized administrative node. And each node can move in any direction independently from other devices. This network has main three constraints Bandwidth, Computer power and Battery power. Ad-hoc cloud is built on resources that are collected from a network. Network in this case is MANET like in which the physical layer is mapped to mobile devices client agents. The neighborhood relationships are among those client agents which are in the radio range because that MANET is the suitable network to allocate resources for MCC a comparison between two trust management systems for MANET which were proposed in [9, 10] and TMC is provided in Sect. 20.4.

To build an effective Trust management there is a need to outline various issues involved in the design of reputation-based P2P system [11]. To answer the question of how trust can be applied in distributed computing [12, 13], an investigation has been conducted in trust management systems proposed for cloud computing. The proposed models/systems have been compared with each other based on a selected set of cloud computing parameters.

Another Trust Management System architecture was introduced for cloud computing marketplace [14]. This architecture reflects the multi-faceted nature of trust assessment by considering multiple attributes, sources and roots of trust. It aims at supporting customers to identify trustworthy services providers as well as trustworthy service providers to stand out.

To build a reputation trust management system a reputation algorithm is required to calculate reputation values. EigenTrust algorithm had been introduced [15] to be used in P2P networks to decrease the number of downloads of inauthentic files. It computes agent's trust scores in P2P networks through repeated and iterative multiplication and aggregation of trust scores along transitive chains until the trust scores for all agent members of the P2P community converge to stable values. By these global reputation values to choose the peers from whom they download, the network effectively identifies malicious peers and isolates them from the network.

A decentralized trust management middleware for ad-hoc, peer-to-peer networks had been introduced [13]. In this middle ware, reputation information of each peer is stored in its neighbors and piggy-backed on its replies to requests for data or services. This middleware relies on the lack of network structure to manage reputation information in a secure way.

To conduct the performance test a general-purpose evaluation framework is used [3]. This framework introduced to evaluate reputation management for distributed systems, peer-to-peer (P2P) networks, and ad-hoc mobile computing. This framework was chosen because it was built specially to evaluate reputation algorithms in P2P and ad-hoc networks. And it provides a mechanism to plug new algorithms.

PlanetCloud, a new architecture was introduced for forming ad-hoc mobile clouds. It is designed especially for emergency and critical situations [1, 2]. This architecture was chosen to build our trust management system on because it has resource servers (RS) which are distributed servers that are used to store spatio-temporal calendar and other information. RS can be used to store trust values for client agents to be used by cloud agents in the cloud formation request. PlanetCloud is described in more details in the Sect. 20.3.

20.3 PlanetCloud in Brief

PlanetCloud [1, 2] is an architecture that was developed to enable resource provisioning and dynamic spatio-temporal resource calendaring to form ad-hoc cloud. It aims to utilize the increased number of mobile devices available to provide "on-demand" scalable distributed computing capability, especially in critical situations. This architecture has three main components Client Agent, Cloud Agent and Resource Server (RS). Each of these components will be discussed in brief.

20.3.1 Client Agent

Client agent is the application used by the user who has a resource which he is willing to share. This application manages the client spatio-temporal resource calendar: which resource can be shared and when. It handles incoming requests for cloud formation, notifies a user of the next incoming clouds, connects with all other agents involved in the cloud formations, and synchronizes the calendar's content with the Resource Server's data.

20.3.2 Cloud Agent

Cloud agent is the application used by the user who wants to form a cloud. This application is deployed on a high capability client to manage and store the data related to spatio-temporal calendars for all clients within a cloud. The cloud agent uses local data repository to store the user profiles, and spatio-temporal calendars of clients within a cloud. This application can query data repositories in RS for extra information.

20.3.3 Resource Server

Distributed RSs operate on the updated data from clients' calendars and clouds data, which are stored in a data repository. The RS provides resource forecasting using an implemented prediction unit. It has a data store that is used to store calendars and other information and it has other three sub components, "Information Bases" where it stores information about cloud formation requests and users information, "Account manager" which contains billing information, and "Synchronizer" that synchronize data between RSs and users.

Figure 20.1 explains the high level architecture of PlanetCloud. RSs are distributed not centralized, data are replicated through RSs. Client agents and cloud agents can directly communicate to RSs. Cloud agent can communicate directly to client agents.

20.4 Proposed Trust management System for Ad-Hoc Mobile Cloud TMC

20.4.1 TMC System Design

The goal of the proposed system is to avoid the participation of malicious client agents which would affect the performance of the formed cloud. This is due to the

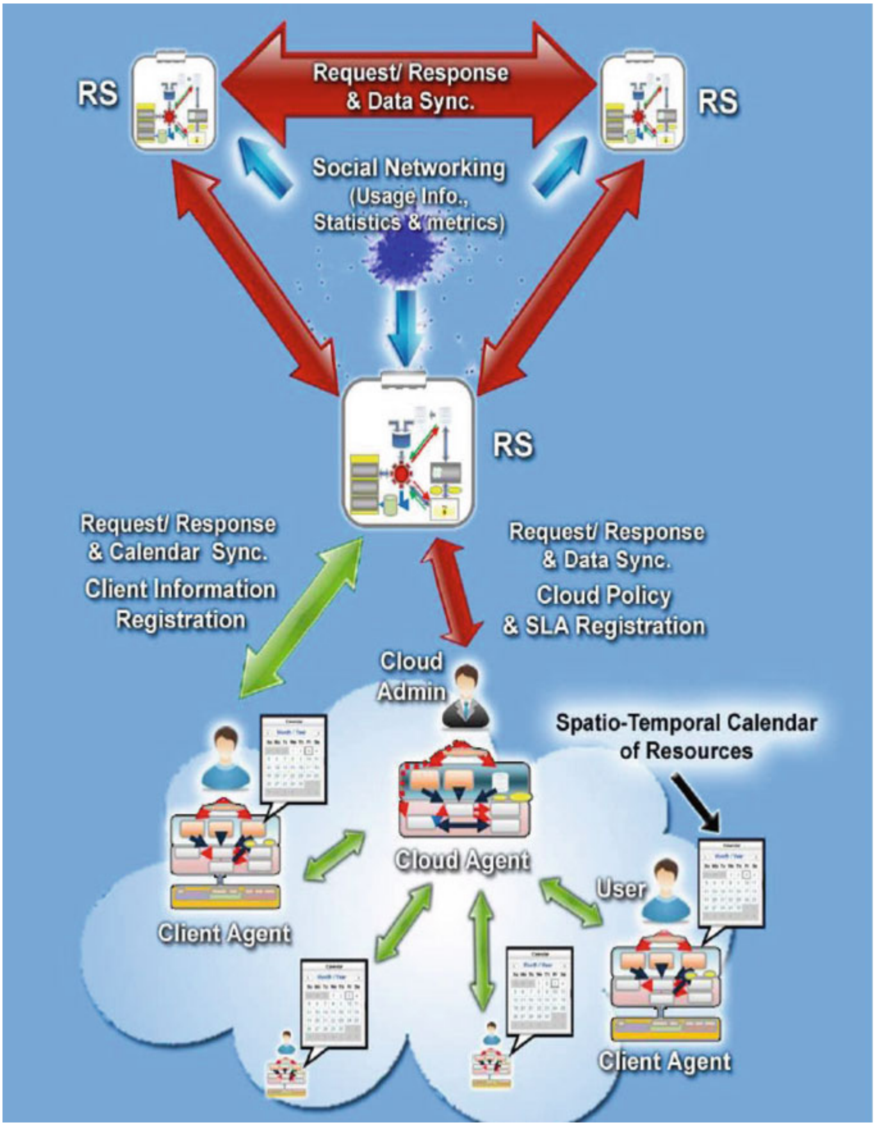


Fig. 20.1 PlanetCloud architecture [2]

fact that this system is used in critical situations which require zero tolerance with malicious behavior.

A trust system for ad-hoc mobile cloud is proposed that evaluates members of a cloud and computes trust value during its interaction in a cloud. This trust value is stored in the cloud agent's local repository then it synchronizes these values with the data repository of RSs Table 20.1 contains sample data. Cloud agent can use these

Table 20.1 Client agents information in RS/Cloud agent store

C1	C2	C3	C4	C5
4	0.70759	6	0.98742	0.07101
13	0.78732	9	0.89629	0.92652
17	0.30486	8	0.89628	0.19472
14	0.94364	9	0.86261	0.49694
10	0.38741	7	0.45707	0.47136
5	0.39277	6	0.57597	0.62726
12	0.87789	4	0.56254	0.69499

trust values in the next cloud formation to select the most high trusted client agents. Other cloud agents can query updated trust values of client agents from RSs data repository.

In Table 20.1 columns names are C1 which is Client_agent_ID, C2 which is Trust_value, C3 which is Cloud_participation, C4 which is Availability and C5 is Response_quality.

Because these ad-hoc clouds are made up mainly from mobile devices, the availability of the client agents is considered. Also, the neighbors evaluation is considered, because data are transferred from client agents to cloud agents and vice versa, though a path formed of other client agents. Malicious client agents cannot evaluate other client agents to avoid affecting trust values. Moreover, response speed and completeness of tasks are considered. Finally, the number of clouds that the client agent participated in is taken in consideration. After calculating the new trust value of a client agent, it is stored in the RSs repository to be used in the next selection process.

The EigenTrust algorithm was used to calculate the trust value of each client agent during its participation in a cloud. Figures 20.2 and 20.3 explain the proposed trust management system. In the next paragraphs, the system workflow will be described.

If a cloud agent is forming his first cloud Fig. 20.2 it sends requests to RSs to form new cloud and set its preferences and settings. RSs query its data store to find the most suitable client agents. Then RSs send approval request to client agents. If a client agent sends his approval to RSs, it checks if client agent is complying with cloud formation settings and preference.

Next, the cloud agent can form his cloud and after a cloud is formed, cloud agents start to evaluate cloud members and receive each client neighbors evaluation to include it in the final evaluation stored in the local data store, which is then synchronized with RSs. This evaluation is used to calculate the average trust value of each client agent to be used in next cloud formation.

In the subsequent as shown in Fig. 20.3 cloud formations, a cloud agent checks if it can establish a connection to RSs then it will check trust values for its trusted client agents only in RSs data store. Then RSs will query the data store to query if the number of clients to form the cloud is not covered by trusted client from cloud agents. Then it request approval from client agents before complete cloud formation request.

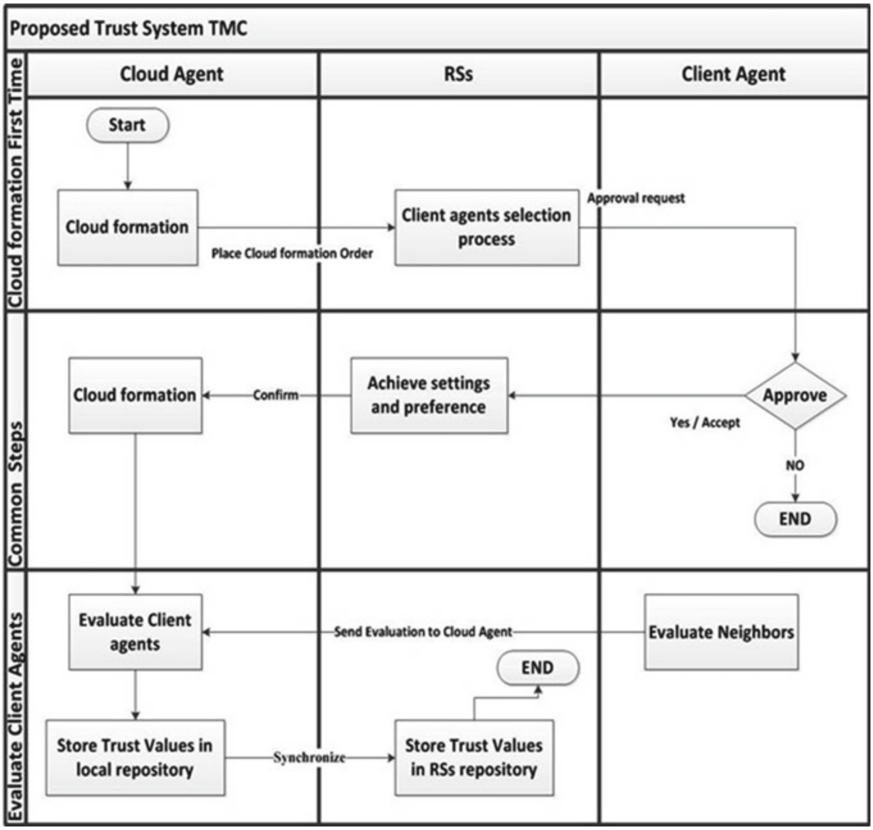

Proposed Trust System TMC
Cloud Agent
RSs
Client Agent
Cloud formation First Time
Common Steps
Evaluate Client Agents
Start
Cloud formation
Place Cloud formation Order
Client agents selection process
Approval request
Cloud formation
Confirm
Achieve settings and preference
Approve
Yes / Accept
NO
END
Evaluate Client agents
Send Evaluation to Cloud Agent
Evaluate Neighbors
END
Store Trust Values in local repository
Synchronize
Store Trust Values in RSs repository

Fig. 20.2 First cloud formation workflow

If the connection between cloud agent and RSs is not established, Cloud agent queries its local data store for client agents. Then it selects the client agents using the same criteria used by RSs. Then it send approval request to client agents, if a cloud agent accepts the cloud formation request it sends confirmation to cloud agent. Then cloud agent forms its cloud and then the client agent starts to evaluate its neighbor client agents during its participation in the cloud as previously described. And then cloud agent synchronizes trust values with RSs once it can establish a connection with it. In Table 20.2 system messages are listed with brief descriptions.

20.4.2 Simulator

P2P Trust Simulator [3] is used to execute performance tests. It was originally implemented to be fed with fixed numbers of pre-Trusted users, good Behaving users and other values for other malicious behaviors. Then it generates random values for each

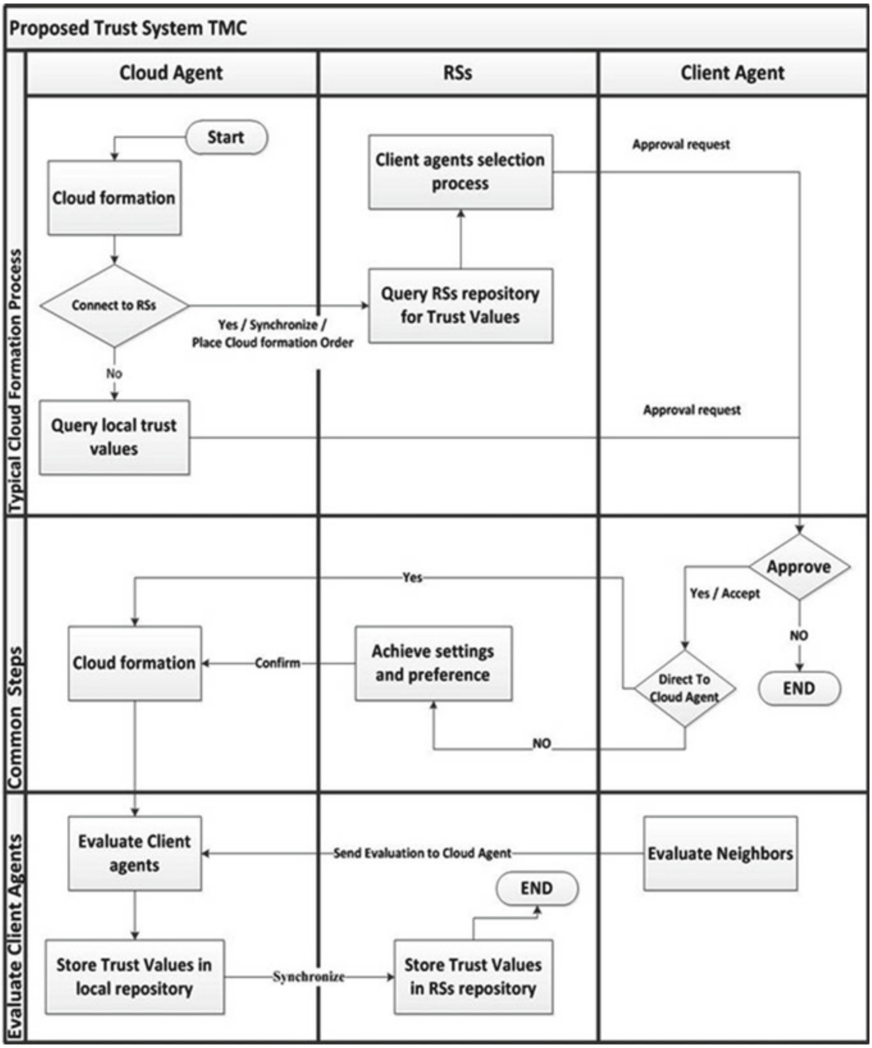

Proposed Trust System TMC
Cloud Agent
RSs
Client Agent
Typical Cloud Formation Process
Common Steps
Evaluate Client Agents
Start
Cloud formation
Connect to RSs
Yes / Synchronize / Place Cloud formation Order
No
Query local trust values
Client agents selection process
Query RSs repository for Trust Values
Approval request
Approval request
Approve
Yes / Accept
NO
END
Direct To Cloud Agent
Yes
NO
Cloud formation
Confirm
Achieve settings and preference
Evaluate Client agents
Send Evaluation to Cloud Agent
Evaluate Neighbors
END
Store Trust Values in local repository
Synchronize
Store Trust Values in RSs repository

Fig. 20.3 Typical cloud formation workflow

node to match its assigned behavior. Then it calculates number of good and bad transactions.

This simulator works in two steps. In first step, it generate a trace file that contains nodes and its trust values also this file contains the transactions between nodes and files this file is generated based up on the parameters that are passed to trace generator. In second step the trace file is used as an input for trace simulator which will apply the reputation algorithm then it write its output and statistics in a file this steps are shown in Fig. 20.4.

Table 20.2 System Messages

Message	From	To	Description
Place cloud formation order	Cloud agent	RS	set cloud settings by cloud agents
Approval request	Cloud agent/RS	Client agent	Asks client agent to join a cloud
Client agent accept	Cloud agent	RS	Client agent accepts to join the cloud
RS confirm	RS	Cloud agent	Confirms that client agent is compliant with SLA and other cloud roles that was set by cloud agent
Send evaluation	Client agent	Cloud agent	Cloud agent sends its evaluation of its neighbors
Synchronize	Cloud agent	RS	Cloud agent sends its new evaluation of client agents and request to update its local store with new computed client agents' trust values

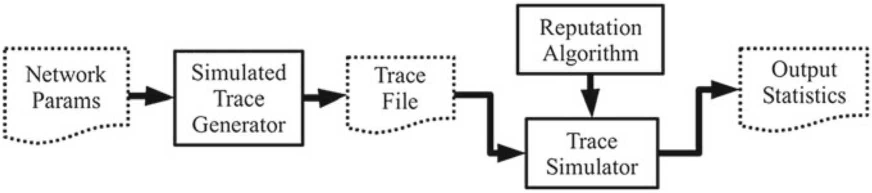


Fig. 20.4 Overview of evaluation architecture [3]

In this simulator, new calculated trust values for nodes are not used in the next rounds. To use this simulator in the performance test it is needed use the calculated trust values in next rounds. Also there is a need to dynamically detect nodes behavior according its Honesty and Response Quality as shown in Table 20.3.

Numbers in Table 20.3 are guided by numbers in Table 20.1 [3]. In this table Pre-trusted Client Agents referees to client agents which are trusted by TMC or by cloud agent. Good Client Agents are client agents which behave honestly during the test round. Malicious Client Agents are client agents which were not performing honestly during the test round. Client Agent behavior is not detected are the agents which TMC could not categorize them into one of the previous behaviors.

A simulation of RSs behavior is done by storing and retrieving these values for client agents to be used by cloud agents. Client agent will be represented by a node and there is a only one Cloud agent.

Pre-trusted client agents are detected in the start of a round. The new client agent who did not participate in any cloud will be assigned an initial trust value between 0 and 1, that is less than a static value called new_trust_value_threshold. The behavior of the new client agent can be controlled by changing the new_trust_value_threshold

Table 20.3 Values used to detect user's type

User type	Response-quality (%)	Honesty (%)
Pre-trusted client agents	90–100	90–100
Good client agents	90–100	70–100
Client agents behavior is not detected	50–70	50–70
Malicious client agents	0–50	0–50

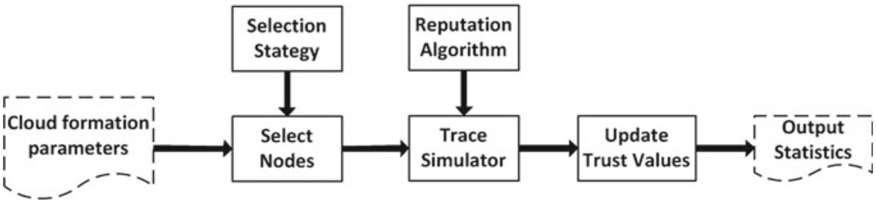


Fig. 20.5 Overview of evaluation architecture after enhancements

value. Unlike the default P2P Trust Simulator behavior, TMC detects client agent behavior from retrieved trust value Honesty and response-quality. Refer to Table 20.3.

In Fig. 20.5 explains the changes that had been done in the simulator. Now it only has two arguments passed to it the required nodes number and the selection strategy. Node selection is done using passed selection strategy then selected nodes are used in the trace simulator step. Then new calculated trust values are updated. Finally statistics are written to a file.

To calculate the trust value (c) of a client, last trust (t) and response-quality (r) values are loaded from database also neighbor-evaluation (n) and availability-evaluation (a) are taken into consideration based on the users' behavior. 100,000 transactions are generated to be used in the simulation to calculate trust value (t). However, these values cannot exceed the new_trust_value_threshold specified for a new member.

These values n, r and a are used in an Eq. 20.1 to calculate the final trust (t) value which is stored in RSs.

$$c = (0.3 \times t) + (0.1 \times n) + (0.2 \times r) + (0.4 \times a) \tag{20.1}$$

These weights were chosen for each element to reflect its importance in the mobile cloud environment. The availability is the highest importance because it is what makes the deference in mobile environment so it has the weight of 0.4. The second in its importance is the computed trust value which has been assigned the weight of 0.3. Computed trust value is based on behavior of the client agent during cloud session which reflects the trustworthy of the agent. The third in its importance is response-quality based on response speed and assigned task completeness, it has assigned the weight of 0.2. Then finally neighbor-evaluation which has the weight of 0.1 because it is affected by evaluator behavior.

Table 20.4 Comparison between TMC and MANET trust management systems

	TMC	Buchegger, Le Boudec [9]	Velloso et al. [10]
Calculation	Cloud agent for client agents which participated in a cloud and every client agent for its neighbors	node for neighbors only	each node for every in radio range
Store	Stored centralized in RSs and locally in cloud agents	Locally	Locally
Reusability	Reused by the TMC system to recommend client agents to cloud agents	Used during current network session	Used during current network session
Exchange	between TMC and cloud agents	between neighbors nodes only	between all nodes in the radio range

RSs client agent selection process is simulated, by ordering client agents using their Honesty value, response-quality and number of participation in the clouds. Then select from lists top the number of client agents requested by cloud agent.

20.4.3 Comparison Between TMC and Existing Trust Management Systems for MANET

There exists number of trust management systems were built for MANET. In the next table a comparison between TMC and two trust management systems were built for MANET. Table 20.4 compares between TMC and trust management systems [9, 10] in term of trust value calculation, trust value store, trust value reusability, and trust value exchange.

20.5 System Performance Test

In this section, the behavior of TMC is simulated and then its results are compared with plain PlanetCloud. To achieve this goal three performance tests were conducted.

In each performance test, the simulator was run 10 test rounds using the proposed system TMC. Then the simulator was run another 10 test rounds with plain PlanetCloud, under the same conditions. Then, the two results are compared. The comparison is done between the numbers of pre-trusted client agents, good behaving client agents, malicious client agents and unknown behavior client agents and between the execution times.

Pre-trusted client agents are trusted by the system before it entered a round. Good behaving client agents are the agents that have good behavior during a round. Malicious client agents are the agents that have malicious behavior during the round. Unknown behaving client agents are the agents that system could not categorize them into one of the previous behaviors.

In each experiment there is a set of 1,000 client agents to select 100 clients out of them to form a cloud. Number of performed transactions in each experiment is 100,000. In the Performance Tests a machine that has Intel core I3 processor with 2,261 MHZ CPU speed and 4 GB RAM was used.

20.5.1 Performance Test: One

In this performance test, the new_trust_value_threshold was set to be 0.7 which is near to the good type range. There is a set of 1,000 client agents to choose 100 client agents out of them. Number of transaction used was 100,000.

Figure 20.6 shows that the number of transactions done by good client agents within the ten rounds using TMC outnumbers the same number when not using TMC.

Figure 20.7 shows that the number of pre-trusted client agents that entered the ten rounds using TMC are increased from round to round more that those entered the rounds without TMC. Also the total number of them using TMC is more than it without using TMC. Figure 20.8 shows that, the number of the good behaving client agents in each round with using TMC is more than that number without using TMC.

Figures 20.9 and 20.10 show that TMC detects and eliminates malicious and unknown-behavior client agents and that this improves round after round. It has been clear from these experiments that adding the trust system to the existing PlanetCloud system will not only improve the performance of the system, but will also provide a reliable repository of client trust values which is useful in assigning future cloud groups.

TMC Results show that the number of good client agents and pre-trusted client agents are increasing from round to the next which means that TMC successfully eliminates bad behaving client agents in every round and recommends good behaving client agents to cloud agents to form their clouds. While the default behavior of the PlanetCloud is randomly recommending client agents based only their calendar, so that the participants in a cloud might be bad behaving client agents or new client agents.

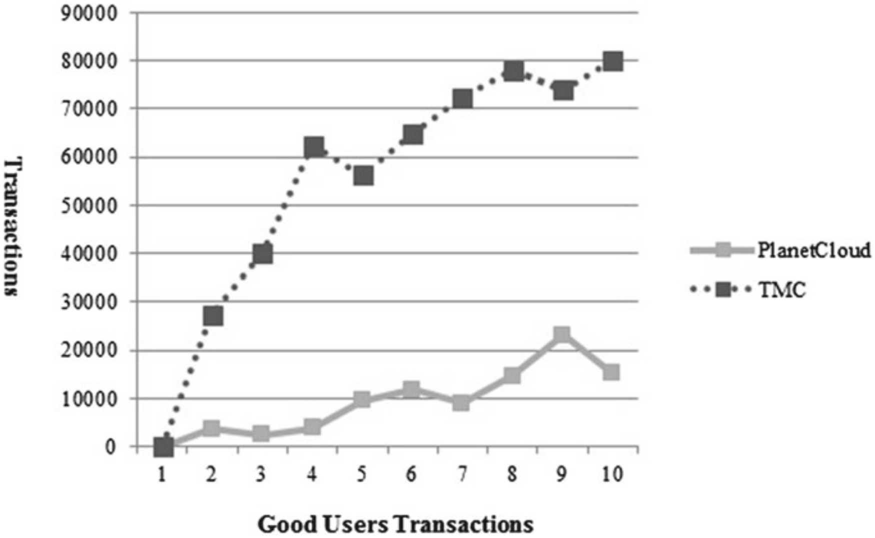


Fig. 20.6 Performance test one: number of transactions done by good users

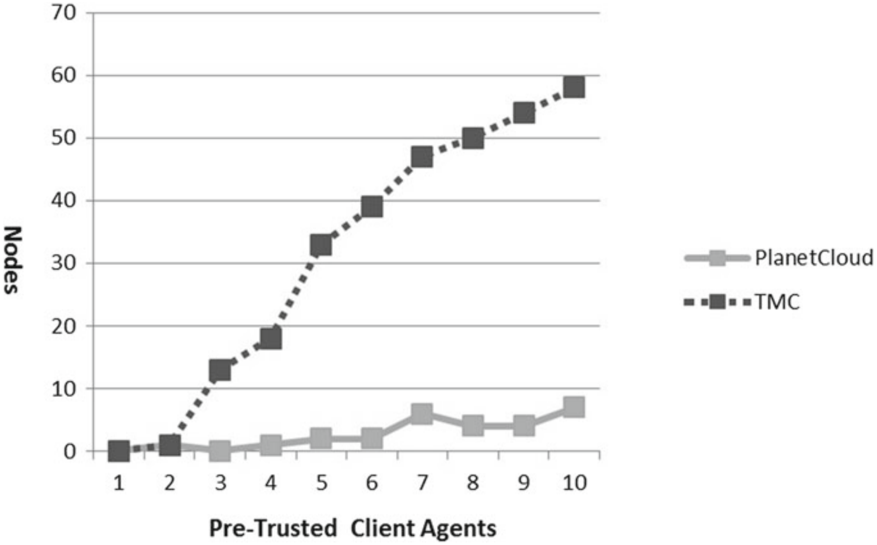


Fig. 20.7 Performance test one: pre-trusted client agents

20.5.2 Performance Test: Two

This performance test was done with the same scenario as performance test one but with new_trust_value_threshold set to be 0.5 which is near to malicious type range.

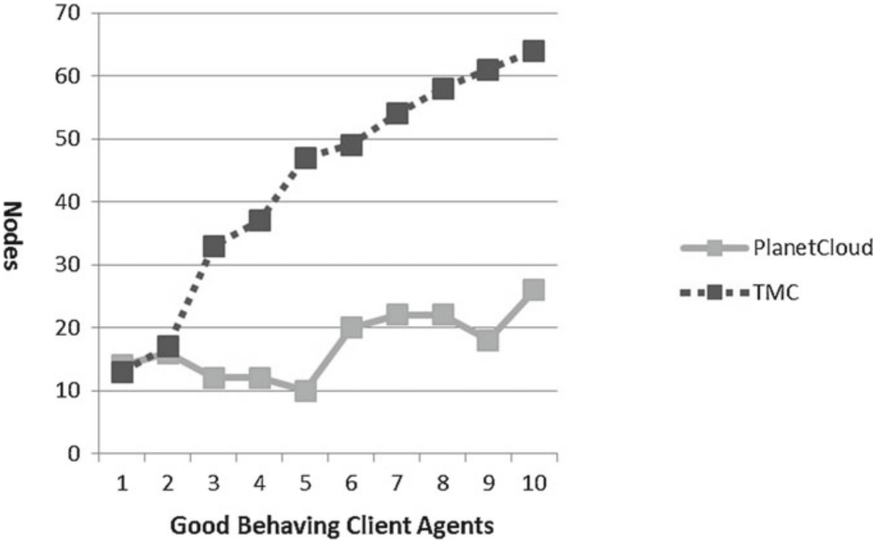


Fig. 20.8 Performance test one: good behaving client agents

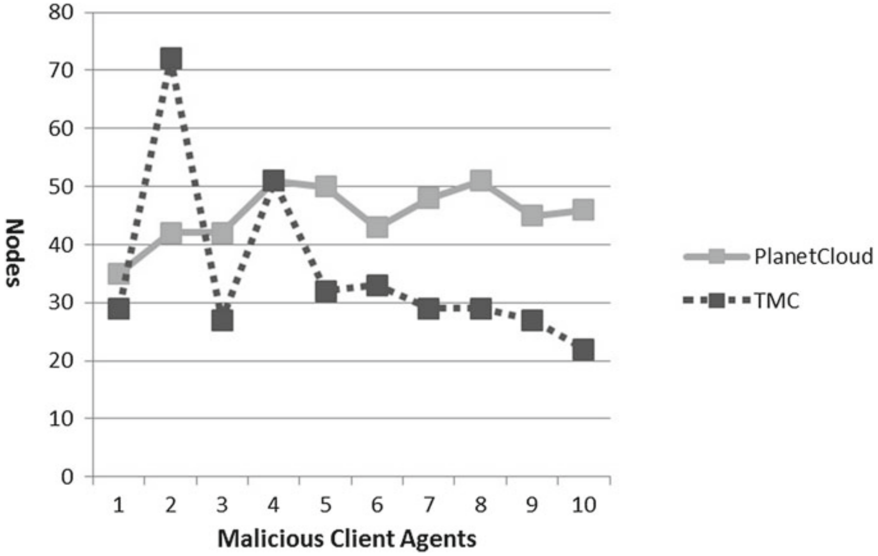


Fig. 20.9 Performance test one: malicious client agents

This means that new members are originally assumed to be malicious or close to it. There is a set of 1,000 client agents to choose 100 client agents out of them. Number of transactions used was 100,000.

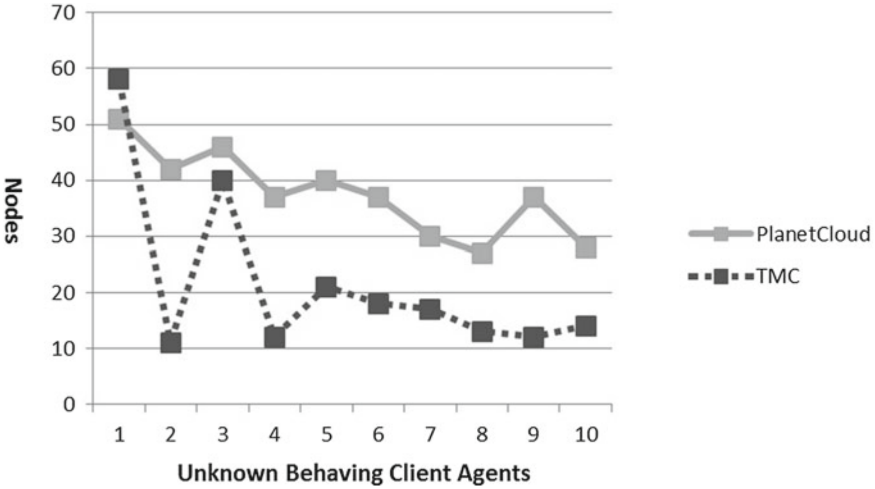


Fig. 20.10 Performance test one: unknown behaving client agents

Results show no improvement over the original PlanetCloud, and are therefore not shown in this chapter. This was because the new_trust_value_threshold was near to malicious behavior type. So that reputation and trust values improvement cannot be noticed in 10 rounds.

To solve this problem a change has been made in the simulator then run performance test three.

20.5.3 Performance Test: Three

In this performance test a change was made to Eq. 20.1 that is used to calculate final trust value. RHS of Eq. 20.2 is multiplied the by step (s) percent, after the new trust value (t) is calculated.

Step (s) value is a number that has two possible values, high value if the client agent's good transactions are greater than its bad transactions and the low value is used in the other cases.

The low percent is set to fixed value 0.99 and the high value is set to be 1.1. This step will move some client agents from bad behaving ranges to good behaving range, based on number of good transactions and bad transactions.

$$c = s \times ((0.3 \times t) + (0.1 \times n) + (0.2 \times r) + (0.4 \times a)) \qquad (20.2)$$

In this performance test the new_trust_value_threshold is set to be 0.5 which is near to malicious type range. There is a set of 1,000 client agents to choose 100 client agents out of them. Number of transaction is 100,000.

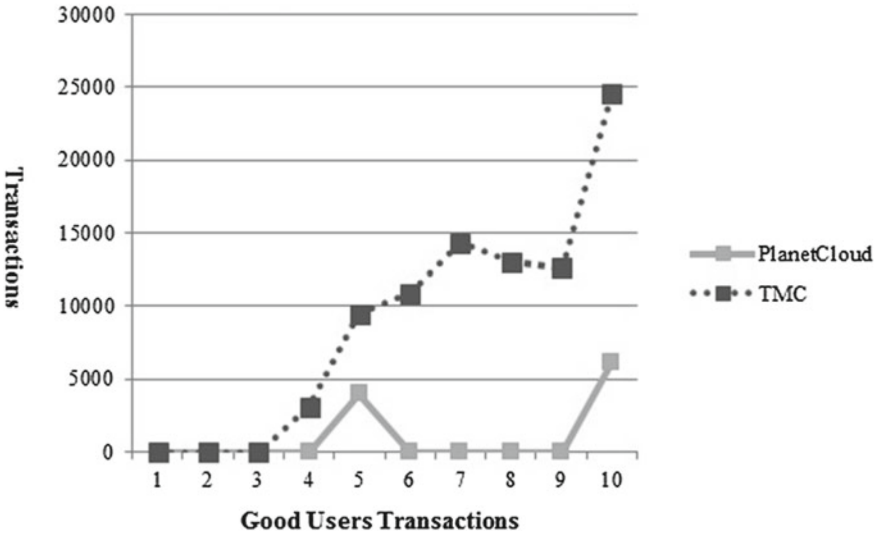


Fig. 20.11 Performance test three: number of transactions done by good users

Figure 20.11 shows that the number of transactions done by good behaving client agents with TMC outnumbers the transactions done by good behaving without it, with the TMC system giving an enhancement of about 20

Figure 20.12 shows that the pre-trusted client agents who joined cloud request using TMC are more than who joined without it. Figure 20.13 shows that the number of good behaving client agents using TMC is greater than that number without using it. Figure 20.14 shows that the number of malicious client agents is less than that number without using it.

TMC in round 1 could not detect malicious client agents because all client agents had unknown behavior then in the next rounds the number was increased. Figure 20.15 shows that TMC could categorize agents efficiently.

In this performance test we still generate client agents in the same range that is near to malicious range but to distinguish and categorized the behavior of client agents faster than performance test two in ten rounds we used the step value. Again TMC Results show that the number of good client agents and pre-trusted client agents was increasing from round to the next which means that TMC could eliminate bad behaving client agents in every round and it recommended good behaving client agents to cloud agents. PlanetCloud is randomly recommended client agents based only their calendar, so that the participants in the clouds might be a bad behaving client agents or new client agents.

Finally based on these three performance tests state that TMC enhance the performance of the PlanetCloud in term of numbers of Pre-trusted Client Agents, good behaving client agents, Malicious Client Agents and Client Agents with unknown behavior. This will reflect positively on the cloud performance, and was experimentally proven to be at least 20 %.

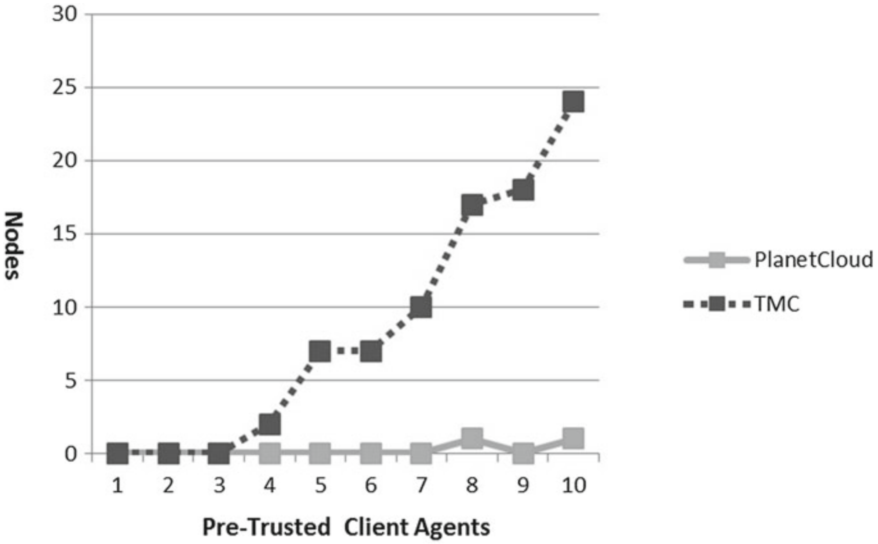


Fig. 20.12 Performance test three: pre-trusted client agents

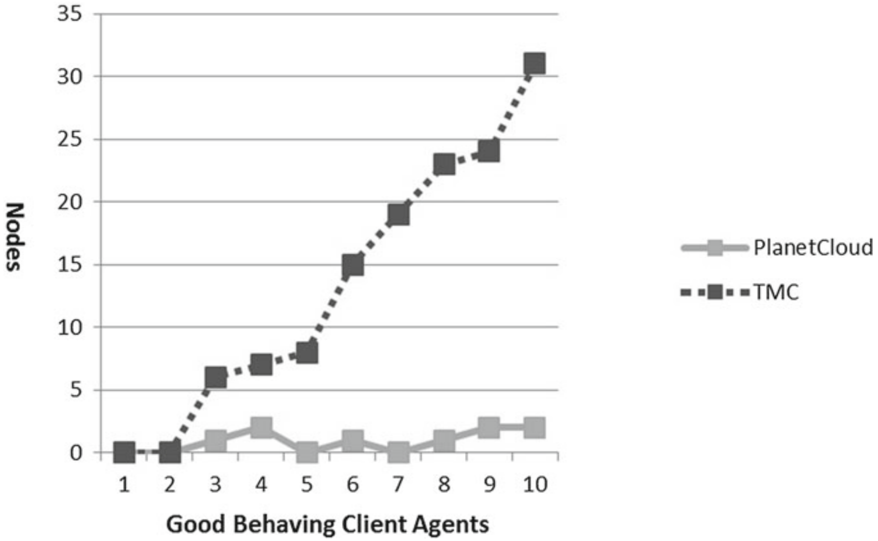


Fig. 20.13 Performance test three: good behaving client agents

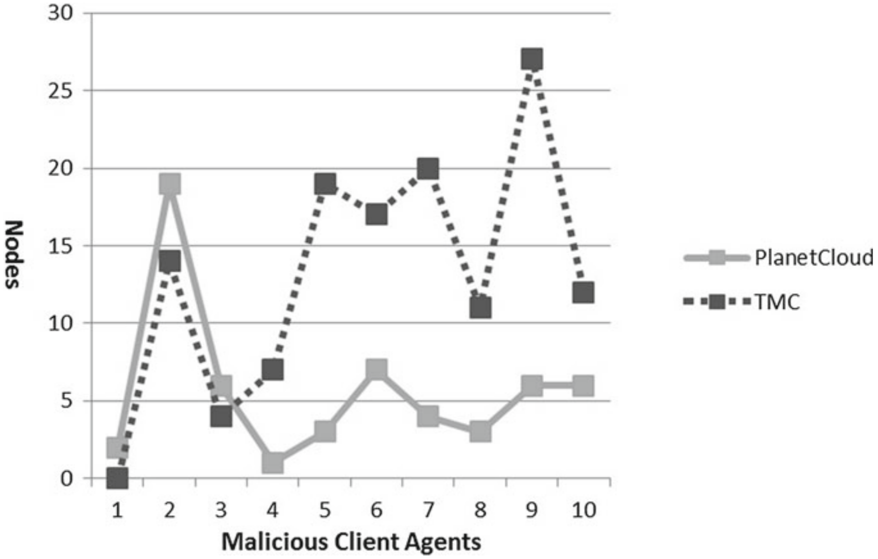


Fig. 20.14 Performance test three: malicious client agents

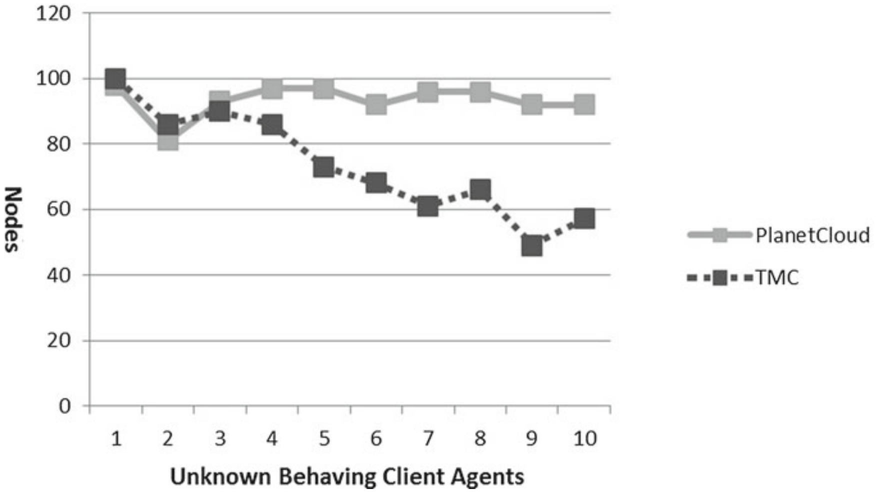


Fig. 20.15 Performance test three: unknown behaving client agents

20.6 Conclusion

The proposed system monitors the client agents behaviors that join the ad-hoc clouds in the PlanetCloud to provide better results for cloud agents by providing them by trusted Client agents. Centralized store for this information may set some limitation on forming new ad-hoc cloud, but already cloud agents and client agents need to access RSs frequently to read or update spatio-temporal calendar. And also cloud agents have a local repository which can be used in cases that RSs are not reachable. Scalability of PlanetCloud will not be affected because Cloud agents just need to connect to RS in the selection phase of client agents. RS are not interfered during cloud sessions and cloud agents need only to synchronize its new calculated values when it is possible. And it is possible for cloud agents to depend on their local data store if they cannot connect to RS.

It has been shown that the number of good behaving client agents and pre-trusted client agents is enhanced from one round to the next. And on the long run large numbers of good behaving, pre-trusted and bad behaving agents will be identified and recorded in the RSs what will help cloud agents to safely form clouds in low time cost especially in crises and emergency situations.

References

1. Khalifa, A., Eltoweissy, M.: A global resource positioning system for ubiquitous clouds. IEEE International Conference on Innovations in Information Technology (IIT) (2012)
2. Khalifa, A., Hassan, R., Eltoweissy, M.: Towards ubiquitous computing clouds. FUTURE COMPUTING 2011, The Third International Conference on Future Computational Technologies and Applications (2011)
3. West, A.G., Kannan, S., Lee, I., Sokolsky, O.: An evaluation framework for reputation management systems. In: Yan, Z. (ed.) Trust Modeling and Management in Digital Environments: From Social Concept to System Development, pp. 282308. IGI Global, Hershey (2010)
4. Distefano, S., Puliafito, A.: Cloud@ Home: toward a volunteer cloud. IT Prof. **14**(1), 27–31 (2012)
5. Kirby, G., et al.: An approach to ad hoc cloud computing. Technical Report, University of St Andrews (2010)
6. Dinh, H.T., et al.: A survey of mobile cloud computing: architecture, applications, and approaches. Wirel. Commun. Mob. Comput. **13**, 1587–1611 (2011)
7. Fernando, N., Loke, S.W., Rahayu, W.: Mobile cloud computing: a survey. Future Gener. Comput. Syst. **29**, 84–106 (2012)
8. Anderson, D.P., et al.: SETI@ home: an experiment in public-resource computing. Commun. ACM **45**(11), 56–61 (2002)
9. Buchegger, S., Le Boudec, J.-Y.: A robust reputation system for mobile ad hoc networks. Technical Report. IC/2003/50, EPFL-DI-ICA, 2003
10. Velloso, P.B., et al.: Trust management in mobile ad hoc networks using a scalable maturity-based model. IEEE Transactions on Network and Service Management, vol. 7, issue 3, pp. 172–185 (2010)
11. Maini, S.: A survey study on reputation-based trust management in p2p networks, p. 117. Technical Report, Department of Computer Science, Kent State University (2006)
12. Firdhous, M., Ghazali, O., Hassan, S.: Trust management in cloud computing: a critical review. Int. J. Adv. ICT Emerg. Reg. **2**:24–36 (2011)

13. Repantis, T., Kalogeraki, V.: Decentralized trust management for ad-hoc peer-to-peer networks. Proceedings of the 4th international workshop on Middleware for Pervasive and Ad-Hoc Computing (MPAC 2006). ACM (2006)
14. Habib, S.M., Ries, S., Muhlhauser, M.: Towards a trust management system for cloud computing. IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom 2011) (2011)
15. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The eigentrust algorithm for reputation management in p2p networks. Proceedings of the 12th International Conference on World Wide Web. ACM (2003)

Chapter 21
Secured and Networked Emergency Notification Without GPS Enabled Devices

Qurban A. Memon

Abstract Lately, many people have become aware of privacy issues that emanate from wide spread use of GPS enabled or similar systems. There are two concerns with GPS systems. On one side, many people are becoming more aware of privacy issues, and on the other hand emergency help is of great importance due to fading of signals in presence of jammers that inhibit GPS or similar systems.Thus, there is a need for location detection technique or a solution that not only helps in detecting location but also relieves the person of the privacy concern. Such systems find widespread application in military and personal communications. In this chapter, firstly introduction to various well known systems is presented, followed by existing location detection methods and technologies. The privacy issues are also highlighted. The comparative study is done to highlight strengths and weaknesses of various location detection approaches. Furthermore, various position equation methods are derived to estimate position location accuracy. In order to address privacy, a location detection system is developed for a limited geographical area as a case study. The simplicity and security of data transfer are also addressed by encryption using special purpose microcontrollers. The standardization efforts for location identification are also summarized.

21.1 Introduction

Emergency alert and response systems are of great interest in this age of personal safety and security. Generally, there are situations in operational area like drilling, mining, defence, sub-surface or any other indoor location where global positioning devices are either in-efficient or expensive to use, and typical local/indoor or private

Q. A. Memon (✉)
Department of Electrical Engineering, College of Engineering, UAE University,
15551 Abu Dhabi, UAE
e-mail: qurban.memon@uaeu.ac.ae

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_21, © Springer-Verlag Berlin Heidelberg 2014

area networking is only available option. Response to different types of emergencies in the indoor environment is critical in order to protect resources including human life. The tracking and navigation in such situations is expected to be the key to the development of user specific services and applications. In these situations, the requirement is to track the object during emergency in real time and provide rescue services dependant upon nature and priority.

A study carried out in [1] discusses the future avenues for tracking and navigation in Western Europe. In this study, it is investigated that personal navigation in Western Europe during 2007–2012 grew by about 10 % annually. With the launch of iPhone and similar ones by its competitors, this market is expanding dramatically in the next few years. Similar pattern of growth is expected to be witnessed in the middle-east, especially Gulf Cooperation Council (GCC) region. There have been four fundamental market requirements to such location based services: Convenience or simplicity, Efficiency, Security, and Reliability. The fulfilling of these requirements leads to wide acceptance of such devices into the market.

For tracking personnel, vehicles or objects, there are a number of approaches suggested in literature that may be investigated to provide a solution during an operational environment. The environments include areas where GPS tracking can be deployed to provide safety and security. GPS has found applications in military and civilian domain to the extent that it can be considered a utility, similar to water, electricity, and gas. But GPS receivers are power-hungry [2] and an application may require a low sampling rate in situations when mobile device battery is low. In those situations, lower sampling is likely to result in sparser set of track entries. For purpose of high precision, even the Assisted GPS (A-GPS) could also be too inaccurate and costly in a densely populated area with large points of interest. Current solutions thus lack in providing fool-proof privacy and are not efficient and reliable in some areas like indoors, sub-surface, etc. Additionally, the systems involving GPS have been reported to be ineffective in situations where GPS signal is affected by jammers or interferers.

Most of the 802.11 device drivers broadcast their existence to the infrastructure on a timely basis, and it is harder to tell that no GSM cell phone reports itself to the infrastructure. The cell phone providers, restaurants and other related businesses do not care if the existence and location of their network points are known, but this may not be acceptable to many individuals and corporations with information about their access points being listed in public domain. A large number of people have, thus, become very concerned about their safety and security due to use of a variety of hi-tech devices. Though, there are significant advantages associated with their use, but the devices pose direct challenge to privacy matters of the individual or the corporate. The fear is what clever people can do with the data that is accessed or is floating around the networks. A good number of research works and projects are being executed to address these issues and the priorities, but the concerns still remain. As regards to privacy,the countless benefits of location based services (LBS) are lost when compared to social hazards encountered on these platforms. Inherent in this concept is the potential within a central location to routinely control time, location, speed, and direction for each and every movement of the client device or, indeed, of many clients simultaneously. Furthermore, users usually visit places via set routes as part of their daily routine. The tracks once stored in database expose the daily routine of the device user. Another side effect of this technology use surfaces when the navigator regularly navigates by pushing buttons blindly and accesses "black boxes". Thus, he/she will not be prepared to improvise solutions in emergency using basic principles.

In a wireless environment, security may be understood by knowledge of security standards like 802.11 WEP, 802.11 WPA and WPA2 (802.11i). One of the major flaws in WEP is its limited 40-bit key length, which can be broken in few hours with the help of computer machines [3]. With new release as WEP2, its length is increased to 104-bit key. Besides, WEP2 has key management problem, and that the WEP does not support mutual authentication. Switching to WPA and WPA2 technologies has solved some problems on both user and company levels, but it is difficult to say that wireless networks are secure. WRAP (Wireless Robust Authenticated Protocol) is the recent AES encryption standard implemented on LAN platform to improve security.

For encryption purposes, four encryption algorithms may be quoted: AES, XTEA, SKIPJACK® and an algorithm based on a pseudo-random binary sequence generator. Briefly, it can safely said that in today's era of easy data access, electronic data needs to be encrypted to stand a better chance of remaining secure. As said before, many encryption algorithms can be quoted to provide protection against someone reading the hidden data or against tampering. In most of these algorithms, the decryption process causes the entire block of information to be trashed if there is a single bit error in the block prior to decryption.

The Data Encryption Standard (DES) algorithm adopted in 1997 [4], became a worldwide standard for data encryption by ISO (International Standards Organization) [5, 6]. The Advanced Encryption Standard (AES) is adopted by the National Institute of Standards and Technology (NIST) on October 2, 2000 to encrypt and decrypt data. This is a symmetric block cipher, which utilizes a secret key. Its typical implementation is based on a 16-byte block of data with a 16-byte key size, with 10 rounds of encryption. To fit into the data matrix structure, the plain text, which needs to be encrypted, is broken down into appropriate size blocks, with padding of any leftover space. Once a 128-bit key is selected and data partitioned into blocks, the encryption cycle begins.

Nowadays, the level of integration allows processor vendors to integrate encryption engine into processors/microcontrollers. This makes the system secure from the beginning. It also accelerates computations to a level that transactions may be completed in real-time with practically no noticeable delay. Indirectly, it benefits user in saving his waiting time and thus relatively more transactions per minute may be handled. Vendor products in the category of these types of MCUs range from 8-bit to 32-with dedicated encryption engines, random number generators, etc. to protect user data on communication channels.

21.2 Location Based Service Applications

Real-time locating systems (RTLS) belong to local positioning systems that help inreal time tracking and identifying objectlocations, and providing active or passive collection of location information. The term RTLS was coined in approximately 1998. The RTLS systems exclude Cellnet base station segment locators and passive radio frequency transponder indexing (RFID indexers).

Practically speaking, the location systems should be designed based on where people use their time. The scope in current systems is either limited to interior of a building with installed sensing infrastructure or outdoor working environment. Such applications will fail if they only work for a fraction of users or only during a fraction of a user's day. This dictates that LBS application should be deployed in frequently-visited public locations like transportation points, shopping centers, sports facilities etc. Typically, LBS services also include parcel or vehicle tracking services, mobile commercein the form of coupons or advertising directed at customers' current location, etc. The advertising includes location-based games or even personalized weather services.

The first LBS service 'friendzone' was launched by Swisscom on May 2001 in Switzerland, using the technology of Valis Ltd. The first commercial LBS service in Japan was launched by DoCoMo, based on triangulation for pre-GPS handsets in July 2001, and by KDDI for the first mobile phones equipped with GPS in December 2001. In May, 2002, go2 and AT&T launched the first mobile LBS local search application, which used Automatic Location Identification (ALI) technologies mandated by the FCC. Recently, in 2010, location-based services activate Mobile Local Search to enable discovery of persons, places, and things within an identifiable space defined by distinct parameters that evolve with time into a structure that is vertically deep and horizontally broad data category. The service parameters include social networks, individuals, cities, landmarks, and actions that are relevant to the searching person's past, current, and future location. Some well-known examples of location-based services are:

- Suggesting social events in a city [7]
- Mobile advertising based on location
- Nearest business or service request, such as a restaurant
- Alert services, such as sale on petrol or early warning of a traffic jam
- Active RFID combined with asset recovery to find, for example, stolen assets in containers where GPS signals undergo fading
- Locating people with the help of a mobile phone display

21.3 System Concepts and Related Work

Many system concepts sail under the tag of real-time locating systems. However these approaches are very different in respect of cost-to-benefit ratio, and thus need a little attention to describe them here:

a. Locating at choke points

There is a simple locating system that applies no physical measurement, but simply communicates at coincidence of transceiver and transponder as long as communication may happen. This locating mechanism can be related to simple RFID technologies according to an equivalent standard. In fact, this communication mechanism is the only option to apply passive RFID tags for locating objects. The reach of the RFID reader determines the choke point, and determines the accuracy of locating.

b. Locating in relative coordinates

This solution is also termed as fuzzy locating as many references describe locating at relative coordinates. The coordinates may be radial distances compared with reference to known locations with no angular directions. No exact metrics is required, as long as reference points are intelligible.

c. Locating in absolute coordinates

The appearance of satellite navigation systems has enabled setting the requirements for locating of objects. However, determination of absolute coordinates is a challenging task. An ultra-short pulse communication is used instead of electromagnetic waves in these concepts. The solutions based on this concept, however, do not serve results, when targets move. This fact is proven by a large number of publications in literature, but with small number of references on installed systems.

d. Locating in contiguity

This is a newer approach for defining a location just as the contiguous ambience of the person looking for something to be located. This approach may be related to locating at choke points; however, the accuracy is the main achievement with locating in contiguity. This is enabled by the tuned transmission power level of an active RFID tag as an intermittent beacon as opposed to steady illumination of the tag with the reader in locating at choke points. The easy option is to apply graded active RFID tag for economized locating, and thus the reach of RFID receiver determines the base point. The operational suitability may be defined by an algorithm that varies the minimum transmission reach of the beacon. Example of such solution may be quoted as very simple electronic leashes or more complex designs. Another common application can be found in electronic wireless lock solutions. Some advanced applications use autonomous software agents in combination with tag operation, for example smart phonesuse this for monitoring manually controlled services in systems.

As far as products are concerned, one of the approaches is to automatically determine the location of an individual using a tag in the form of a badge [8]. This badge emits a unique code for approximately one tenth of a second. A network of sensors placed around the building catch these periodic signals. Another master node connected to this network polls these sensors for that badge, and processes this badge data. This data can be shown in a suitable visual display, as required by the user. Typically, this kind of badge may be designed in a package size of roughly 55 × 55 × 7 mm with a weight of a mere 40 g. As people usually move slowly in an office environment, this kind of polling interval of about 10 s seems acceptable to detect the location of the badge.

RADAR [9] is the first indoor positioning system based on Wi-Fi signal-strength. The idea is based on the distance between access point (AP) and the mobile receiver. The smaller the distance, the stronger the signal, and vice versa. This signal-strength is exploited by RADAR to estimate the device's location inside a building. Based on this, a radio map is built as a lookup table to hold information regarding packet signal strengths measured and corresponding building locations. Once this radio map is built, finding a user location becomes easy. The searching of this lookup table for best match signal strength entry is the only task to be done once user device receives the signal strength from the AP within its measuring range.

The Place Lab [10] system consists of three key parts: Fixed radio beacons (like APs) spread in the environment; a database that holds information about locations of these beacons; and Place Lab clients that exploit this database to estimate their current position. The Place Lab clients like laptops, PDAs and cell phones estimate their position by listening to cell IDs of these wireless access points, and then referencing the beacons' positions in a cached database. Typically, Place Lab system generates position estimates in two dimensions (latitude and longitude) only.This type of solution creates a problem in multistory buildings, where floor number in the form of altitude becomes a key factor to determine location.

The authors in [11] describe enhanced position location system (EPLS) serves all three services as a position location, identification, communications, and (sometimes) navigation system. The system consists of two primary components, a network control element and a network of Radio Sets (RSs). Although EPLRS can use GPS inputs when available, one of the key features of EPLRS is that it does not depend on GPS to provide position and location data, thus avoiding GPS jamming vulnerabilities.

The cricket system [12] is intended to generate fine-grained location information (in the form of space identifiers, position coordinates, and orientation) to applications running sensors, laptops, and PDAs. Typically, it is developed for indoors use or in areas where outdoor systems like GPS don't perform well. The wall and ceiling-mounted beacons are placed through a building to send information on an RF channel. With each RF signal transmission, the beacon also transmit a concurrent ultrasonic pulse. The receivers (in devices) listen for RF signals, and then follow up on ultrasonic pulse after few bits of RF signals. After this pulse arrives, the listener calculates a distance estimate using the difference in propagation speeds between RF (speed of light) and ultrasound (speed of sound). This distance estimate pertains to the corresponding beacon.

The author in [13] describes in his thesis how modern, simulation-based methods can be used to monitor and, if necessary, to take over the GPS function on a vessel. In fact, the radar installed on a vessel is used to measure the distance to surrounding shores, and this data is then compared with a digital sea chart. Like-wise, in a submarine, the information from sonar equipment is compared with a digital depth chart. Thus, in combination with data about movement of the vessel, the correct position may be calculated. The approach is based on a mathematical algorithm, a so-called particle filter, installed as a program in the vessel's computer system.

The StarTrack [14] system enables operations on tracks, which are discrete and sampled representations of a continuous route. Each participating mobile device in StarTrack is assumed to have a means of determining its current location through available localization technology, such as GPS, GSM localization, Wi-Fi hotspots, etc. The mobile devices collect tracks and opportunistically upload them to a StarTrack central server. The StarTrack operations include facilities for storing, comparing, clustering, indexing and retrieving tracks. Thus a large-scale track-based service is built.

The GPS-free indoor navigation and path prediction architecture of the system proposed in [15] comprises three core elements: effective localization; map representation and route planning; and plan recognition. The data flow in the system originates with the localization module, which receives information from various sensors. A particle filter combines this raw data, from WiFi and dead reckoning, to generate an estimate of the user location that indicates the dispersion of probability of user's presence within the area or the building.

A LocataNet [16] positioning signal system mainly involves a terrestrial segment (TS) and a user segment (US), without a separate control segment. The TS includes a number of LocataLite transceivers located within a pre-defined service area. The US can be any number of fixed or moving Locata user receivers (Rovers) operating within that service area. These derive locations and time within the service area using signals emitted by the LocataLites in the TS. LocataNets may span areas as large as several tens of kilometers in extent, and is supported mostly by the availability of adequate line-of-sight geometries between the various elements of the LocataNet. Using adequate signal power, prototype networks have demonstrated LocataLite-Rover operating ranges of up to 50 km. LocataNets can deploy any easy-to-use coordinate reference system, including WGS-84, or other global, regional, local, or custom profile. As LocataNet's overall concept derives from the Navstar Global Positioning System (GPS), many of its elements therefore are similar to GPS. The LocataLites use the same role as used by GPS satellites, and the Locata user receiver operates similar to a GPS receiver. The techniques used to calculate position and time are similar to those used in GPS.

There is a wide collection of systems concepts and designs to provide real-time locating. For well-known tracking systems, as discussed earlier, a summarized version is listed in Table 21.1. Another list is shown in [17].

Nowadays, Internet has become another source for enabling geolocation detection. Apart from Global Positioning System (GPS), other sources of location information include location inferred from network signals in the form of IP address, Bluetooth MAC address, GSM/CDMA cell IDs, as well as user input. Location detection with HTML5 browser is pretty simple. Generally, browsers either support navigator.geolocation or they don't. If they do,one can see a whole world of local

Table 21.1 Well known tracking systems

Sr. No.	Product/System developed	Development timeline
1	Active badge	1992
2	Radar	2000
3	Place lab	2003
4	EPLRS	2003
5	Cricket	2005
6	Particle filtering based	2005
7	RSN Program	2007
8	Star-Track	2009
9	CMU-RI-TR	2011
10	LocataNet	2011

information available to network users. Another dimension is a deeper and more geographic level of statistics available on server side.

21.4 Location Methods and Technologies

Locating objects or people is generally done in one of the following ways:

a. A single reader in a sensory network identifies nodes using ID signals emitted by them, which is done due to coincidence nodes and the reader.
b. A multiple of readers in a sensory network pick up ID signals from nodes, and a position is estimated/calculated using one or more locating algorithms.
c. Fixed signposts (like APs) with unique identifiers transmit their locations to moving nodes, and which finally relay this information using, for example, a wireless channel to central server for location processing.
d. Mobile nodes mutually interact each other and estimate metering distances.

One of the examples, discussed earlier, for location estimation uses time difference of arrival between the start of the RF message from a beacon and the corresponding ultrasonic pulse to infer its distance from the beacon. Each time a listening device receives information from a beacon, it provides that information together with the associated distance to the attached host processor. The listening host processor infers its position coordinates based on distances from multiple beacons whose positions are known.

The measurement of travel time of radio waves is typically done using two different principles:

- Trilateration method calculates the travel time of the radio signal from a metering unit. It measures and computes the distance with respect to speed of light in vacuum. There is another worth noting method, known as multilateration that uses distances or absolute measurements of time-of-flight from three or more sites.
- Triangulation method calculates the travel time of a pair of synchronous radio signals from a metering unit with two transmitters. It measures and calculates the

difference of distance with respect to speed of light in vacuum, as an angle versus the baseline of two transmitters.

The indoor positioning estimate could also be accomplished by several other methods: Cisco Radio Frequency (RF) fingerprinting, AP triangulation or Received Signal Strength (RSS) lateration. The list also includes cell of origin—the simplest way to determine the originating position (similar to the associated access point in 802.11,). But, this approach may give inaccurate results if mobile device is not associated to the nearest AP. For greater accuracy, this approach could be combined with Received Signal Strength Indicator (RSSI).

Distance calculation may involve any of the following measurements:

- Time of Arrival (ToA): This requires that transmitting node or device be synchronized. ToA shows the measured signal's travelled time, and helps in determining the distance (velocity multiplied by time) to the target. The three neighboring APs create a triangulation to help determine the location.
- Time Difference of Arrival (TDoA). This may be used if transmitting node is not synchronized, but receiving devices are synchronized with each other. The relative time is now measured between these several receiving devices as they detect the same signal in different locations. Three neighboring APs typically create a hyperbolic trilateration.
- Angle or angulation—Angle of Arrival (AoA): This method determines the angle of incidence of the received signal from the mobile device. When signals from two APs are compared, it is possible to determine the originating location.

Ranging, as a term for measuring distance, has also been termed as a prerequisite for locating. Determining the distance may involve either a non-cooperative scanning process, as with RADAR or LIDAR, or a cooperative direct distance measuring process, typically used with RTLS.

APs typically transmit the information about the received signal from any user device (for example WLAN phone, RFID tag etc.) toward the WLAN main node, and further to wireless location application. This information using this or any of the previously described location-tracking information may be combined with tagging technology to open new technological avenues. The location calculation application may have a database that is checked against the user's real-time available location. Thus the location may be shown on the map of the floor plan, displayed on browser-based console.

In nutshell, technologies for locating methods include:

- Active radio frequency identification (Active RFID)
- Active radio frequency identification and Infrared as a hybrid (Active RFID-IR)
- Bivalent systems
- Bluetooth
- Infrared (IR)
- Low-frequency signpost identification
- Optical locating
- Radio beacon

- Semi-active radio frequency identification (semi-active RFID)
- Ultrasound Identification
- Ultrasonic ranging (US-RTLS)
- Ultra-wideband (UWB)
- Wide-over-narrow band
- Wireless Local Area Network (WLAN, Wi-Fi)

As discussed before, many location estimation approaches have been proposed in the literature, but it is difficult to say, which technique provides the best-fit solution for a specific problem. The positioning methods are diversified and the number is large. The choice of a specific method is often based on user specifics. A general model for investigating various methods and their selection as a best solution for a locating problem is constructed at Radboud University of Nijmegen [18].

21.5 Positioning Equations

Technically, the problem of locating a user device (UE) using only the transmitted signals can be divided in two parts: (a) measuring the channel parameters from the transmitted signals, and (b) computing the position from the estimated parameters. Typically, these functions are assigned to different components in the network: the first one to a signal measurement function; the second one to a computing function. As discussed in previous section, the conventional methods for wireless positioning include estimation of direction/angle-of-arrival (AOA), time-of-arrival (TOA) or time-difference-of-arrival (TDOA). The estimation takes place once multiple timing or angles of arrival are measured at a required number of receiving nodes or base stations (BS). In order to estimate location accuracy, each of the homogeneous measurements (for example AOA, TDOA or TOA) or heterogeneous measurements as ahybrid (using combination of these) is discussed in this section.

TOA measurements: In TOA-based radio location technique, the distance between transmitting node/mobile station (MS) and a receiving node/base station (BS) is obtained by finding the time between that MS and that BS and multiplying it by the velocity. With multiple BSs, the MS location can be resolved by triangulation.

Assuming t_i is the one way TOA between the MS and BS$_i$, then

$$t_i = \frac{d_i}{c} + e_i$$

with d_i being the propagation path length between the MS and the BS$_i$ and e_i being the arrival-time measurement error that includes receiver noise, propagation anomalies, and errors in the assumed station position. In fact, d_i depends nonlinearly on the (x, y) co-ordinates of the MS as follows:

$$d_i = ||r - s_i|| = \sqrt{(x - x_i)^2 + (y - y_i)^2}$$

with vector $r = [xy]^T$ and vector $s_i = [x_i\ y_i]^T$ containing the BS$_i$ co-ordinates. Assuming that there are N radio links, the following non-linear system of equations arises:

$$t = \frac{d}{c} + c \equiv f(r) + e \tag{21.1}$$

with vector t = [t$_1$ t$_2$ ... t$_N$]^T, vector d = [d$_1$ d$_2$ ... d$_N$]^{T}and vector e = [e$_1$e$_2$... e$_N$]^T.

This triangulation problem can be solved by Torrieri's approach [19] by linearizing the function f(r) by expanding it in a Taylor series around a reference point, denoted by r_0. Once linearized, the maximum likelihood (ML) estimator is used to provide the following MS position estimate:

$$\hat{r}_{TOA} = r0 + c(F^T N_e^{-1} F)^{-1} N_e^{-1}(t - d_0/c) \tag{21.2}$$

with d$_0$ = $[d_{01}d_{02}...d_{0N}]^T$, d$_{0i}$ = |r$_0$ − s$_i$| for $i = 1, 2, ..., N$ and F the matrix of the equation system after linearizing Eq. (21.1):

$$F = \begin{bmatrix} (r_0 - s_1/d_{01}) \\ \cdots \\ (r_0 - s_N/d_{0N}) \end{bmatrix} \qquad d_{0i} = ||r_0 - s_i||$$

The bias of this estimation depends on the linearization error and on the mean value of the TOA measurement error. The practical use of Eq. (21.2) assumes the knowledge of a more or less accurate initial position estimation vector r_o, which may be available if the UE is being tracked or if previous short term positions have been recorded.

Mixed TOA and TDOA measurements: This method is directly driven by Friedlander's TDOA equations [20]. Let r_{ij} be the difference between two ranges (which are proportional to the TOA measured):

$$r_{ij} = d_i - d_j$$

for $(i, j)\,\varepsilon\{1, N\}^2$ where N is the number of BS, then:

$$d_i^2 = r_{i1}^2 + 2d_1 r_{i1} + d_1^2$$

On the other hand,

$$d_i^2 = ||s_i||^2 - 2s_i^T r + ||r||^2$$

where ||s$_i$|| is the Euclidean norm of s_i. Using these expressions, the linear system of equations can be written as:

$$s_r = u - d_1 p \tag{21.3}$$

$$S = \begin{bmatrix} (x_2 - x_1)(y_2 - y_1) \\ \cdots \\ (x_N - x_1)(y_N - y_1) \end{bmatrix} \qquad u = \frac{1}{2}\begin{bmatrix} ||s_2||^2 - ||s_1||^2 - r_{21}^2 \\ \cdots \\ ||s_N||^2 - ||s_1||^2 - r_{N1}^2 \end{bmatrix} \qquad p = \begin{bmatrix} r_{21} \\ \cdots \\ r_{2N} \end{bmatrix}$$

Thus, Eq. (21.3) is neither a real TDOA equation nor a TOA one: it uses differences of distances as TDOA but also needs the knowledge of one range measurement, d_1. The linear method proposed as above is a linear and closed form solution and it does not require that the BS be synchronized, whereas linearization method as proposed by Torrieri [19] requires an initialization and the linearization procedure is only valid when the initialization is set close enough to the true solution.

TDOA Measurements: If the BS are not synchronized, the clock offset between BS is an ambiguity to be resolved. The use of TDOAs instead of TOAs removes this uncertainty. The TDOA approach is qualitatively different in network-based positioning systems. In network-based systems, the TDOAs are the time difference of arrival of a signal sent by the UE and received by several BSs. In either case, a hyperbola, with foci at the BS is defined, on which the UE must lie. This hyperbola is defined by the constant time difference of two BSs. The intersection of two hyperbolic loci defines the 2-dimensional position of the UE.

An important issue for TDOA systems is the need for synchronicity of the BSs; otherwise, the TDOA measurements will introduce a bias error in resulting hyperbolic locus. In hyperbolic systems, the relative arrival times are defined as:

$$t_i - t_{i+1} = \frac{(d_i - d_{i+1})}{c} + n_i \qquad i = 1, 2, 3, \ldots \ldots \ldots, N-1$$

$$\text{and}\, n_i = e_i - e_{i+1} \qquad i = 1, 2, 3, \ldots \ldots \ldots, N-1$$

Thus, in matrix form, the following non-linear equation system can be arrived at:

$$Ht = \frac{Hd}{c} + n \tag{21.4}$$

with matrix H equal to

$$H = \begin{bmatrix} 1 - 1 \;\; 0 \;\ldots\; 0 \;\; 0 \\ 0 \;\; 1 - \;\; 1 \;\ldots\; 0 \;\; 0 \\ \ldots \\ 0 \;\; 0 \;\; 0 \;\ldots\; 1 \;\; - 1 \end{bmatrix}$$

In order to linearize Eq. (21.4), the method proposed by Torrieri [19] to solve the position estimation by the maximum likelihood estimator:

$$\hat{r}_{TDOA} = r_0 + c(F^T H^T N^{-1} FH)^{-1} F^T H^T N^{-1} (Ht - Hd_0/c) \tag{21.5}$$

The solution (21.5) can be iterated until the position estimation becomes stable. In order to avoid convergence problems, a closed form procedure must be used prior to iterative equations.

AOA measurements: Relatively, direction-finding systems carry some advantages with respect to other techniques. It needs to listen to at least two BSs to estimate the 2-D position. However, it has some disadvantages.The important to note is that the accuracy of the AOA method decreases with increasing distance between the UE and BS due to a fundamental limitation of the location estimation function, and depends strongly on propagating conditions. This loss of accuracy can be partially compensated if more BS are used. Direction finding systems,typically, utilize antenna arrays and Direction-Of-Arrival (DOA) techniques to determine direction of the signal of interest. This measurement restricts the source location along a line in the estimated AOA, which is called Line Of Bearing (LOB). When multiple AOA measurements from multiple BSs are used in a triangulation configuration, the 2-dimensional location estimate of the source may be obtained as the intersection of the two LOBs.

In order to determine a closed form solution, a linear relationship between the UE unknown position and an angle measurement function may be derived. It is derived for two BS and can easily be extended to more BSs.

Let us denote the ith base station (for $i = 1, 2$) by BS^i, located at $r_i^T = [\mathrm{x}_i\,\mathrm{y}_i]$ and the AOA measurement of a given user by θ_i' (with the unitary vector k_i in the broadside direction). Let us denote the unitary vector in the ith LOB direction by:

$$v_i^T = \lfloor \cos\theta_i' \quad \sin\theta_i' \rfloor$$

and the user position co-ordinates by the vector r_u:

$$r_u^T = \lfloor x_u \quad y_u \rfloor$$

Generally, using geometry of a user, two base stations, and an origin, the following relation shall always hold:

$$r^u = r_i + p_i v_i \tag{21.6}$$

with ρ_{i} being the range from the user to the BS^i. Since it is assumed that no knowledge about the range is known, relation (21.6) can be restated as:

$$-x_i \sin\theta_t' + y_i \cos\theta_t' = -x_u \sin\theta_t' + y_u \cos\theta_t'$$

with the unknown position as (x_u, y_u). In case of having n BSs listening to the user mobile, the following over-conditioned system is obtained:

$$\begin{bmatrix} -x_1 \sin\theta_1' + y_1 \cos\theta_1' \\ \cdots \\ -x_n \sin\theta_n' + y_n \cos\theta_n' \end{bmatrix} = \begin{bmatrix} -\sin\theta_1' & \cos\theta_1' \\ \cdots & \\ -\sin\theta_n' & \cos\theta_n' \end{bmatrix} \begin{bmatrix} x_x \\ y_u \end{bmatrix}$$

that leads to the linear matrix-vector notation:

$$b(\theta_i') = H(\theta_i')x \tag{21.7}$$

Hybrid AOA and TOA measurements: Hybrid positioning methods combine timing and delay (like AOA and TOA round trip time (RTT)) measurements to estimate the UE position. As an example, the Eq. (21.1) for TOA and the following equation in matrix form for measured bearing angle φ_i with measurement error η_i:

$$\emptyset = f(r) + n \tag{21.8}$$

can be linearized jointly (leading to a closed form solution). It can be easily verified that, for a single BS$_i$, the position of the UE is obtained as:

$$r_u^{(i)} = r_i + ct_i v_i \qquad v_i^T = [\cos \emptyset_i \quad \sin \emptyset_i] \tag{21.9}$$

where c is the speed of light, r_i is the position of BS$_i$, t_{i} is the RTT measured, and φ_i is the AOA of the UE. Furthermore, for N BS, the least square solution of $2N$ equations is given by the average of the positioning obtained for every BS (provided that the measurements have the same variance at every BS):

$$r_u = \frac{1}{N} \sum_{i=1}^{N} r_u^{(i)}$$

In the previous paragraphs, the equations for existing positioning techniques were presented to see to an extent, which linearized methods perform iteratively, and that which mechanisms could be used to provide the initial guess. It seemed obvious that the linear techniques provide accurate initial guess in the region close to the center of the coverage area, and can be used jointly with linearized optimal estimators. Comparing the TOA-based, TDOA-based and AOA-based Torrieri's methods, it may be concluded as follows:

- The TOA method outperforms the TDOA, both in bias and variance of the position estimate and also with respect to the convergence behavior.
- For the three methods, the geometric dilution of position and measurement error standard deviation product (GDOP$\times\sigma$) of the estimated position can be well approximated by the theoretical one, provided that Torrieri's equations are fed with a good initialization.

Since all these results depend on the quality of the angle or delay measurements, it is difficult to compare between time-based methods and angle-based methods because the variances of the measurements are of different nature.

21.6 Operational Considerations and Privacy

Generally, there are situations in operational area like drilling, mining, defence, sub-surface or any other indoor location where global positioning devices are either in-efficient or expensive to use, and typical local/indoor or private area networking is only available option. The tracking and navigation in such situations is the key to the development of user specific services and applications. Each track is a set of entries regarding a person's time, location, and application-specific data. Technically, each track entry is a tuple comprising location, time, and (in some cases) application specific metadata in the form of an XML document with arbitrary contents. Thus, a track is supposed to capture path taken by a mobile device or, more importantly, a person in possession of that mobile device. As an example, Star Track [21] provides applications with a required set of operations for storing, comparing, clustering and querying tracks.

The key consideration with this collection of tracks is its misuse. The misuse is caused due to availability of these tracks at a central place. Users usually visit a small set of places through foreseeable routes as part of their routine. For example, they travel between home and school or work; they go shopping, they walk the pet, etc. Through track clustering, the system enables applications to eliminate near duplicate tracks and assemble tracks into a smaller set of representative tracks.

Another point, worth noting, is related to battery of the mobile devices. As discussed before, GPS receivers are power-hungry [2] and when the mobile device's battery is low, an application may perform under a lower sampling rate. Lower sampling rate results in a sparser set of track entries. Differences in the speed of motion (e.g., caused by walking at 3 mph or cycling at 15 mph or driving at 40 mph) also result in track entry variations. Furthermore, traffic congestion or terrain (e.g., a steep hill) can also cause speed variation. Thus, even in the presence of a fixed sampling rate, one can find dissimilar track entries for a path.

In these systems, privacy concerns arise from four sources: Content, Location, Tracks, and Metadata. Additionally, privacy threats also surface from the tracks that a mobile tagging application can record. As a summary, it may be safely said that respective solutions found to date are not mature enough to address privacy during tracking. In the next section, an experimental setup is presented that not only addresses tracking, but also addresses privacy and data security.

21.7 A Case Study Experiment

In this work, an effort is made to provide location detection in order to enable rapid and secured emergency response by developing a system that supplements wireless communications distributed throughout the user operational area. The users are expected to be issued a device that will attach a key chain, for example, as a transmitter. The pressing of the button enables transmission of user (ID) data. The signal propagates throughout the network, and reaches receiver where location of the user is calculated. It is assumed that location detection is calculated in absence of GPS systems. The receiver application runs on a computer. The objective set in this work is privacy, safety and security to be felt by employees at work in return for willingness and productivity of the employee.

Three options were investigated to develop user device. First, Pocket PC was considered for an application development to be used by a user to send a signal to wireless access points/receivers once it is operated. Since, it is highly likely that Pocket PC would be ON all the time hence location information is always available (like GPS system). This creates privacy concerns. The second option investigated was use of a Wireless USB (WUSB) device connected to a Micro controller Unit (MCU) and with a button to control it for sending a signal. In this option, it turned out that it may not be easy or preferable to interface WUSB to MCU. This raised a customization issue, as commercial WUSB's come with set constraints. Third option investigated was use of RF devices. In this case, the chip coverage may be extended if desired. But it requires a whole new system to be built up from scratch, and the concern that it would not be compatible with IEEE 802.11 standard based systems.

Any of the previous options has its own difficulties. The choice opted was a user device simpler than WUSB but compatible with IEEE 802.11 standards as these are widely deployed nowadays. The receiver application has a Visual Basic interface that receives user signal and displays calculated user coordinates onto a map.

(a) Location Calculation Method

Two methods were investigated to calculate the location of the user. One method was triangulation method and the other was GPS based position method. The related equations used for location calculation are shown below:

$$
\begin{aligned}
d_1 &= c(t_{t,1} - t_{r,1} + t_c) = \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + \sqrt{(z_1 - z)^2}} \\
d_2 &= c(t_{t,2} - t_{r,2} + t_c) = \sqrt{(x_2 - x)^2 + (y_2 - y)^2 + \sqrt{(z_2 - z)^2}} \\
d_3 &= c(t_{t,3} - t_{r,3} + t_c) = \sqrt{(x_3 - x)^2 + (y_3 - y)^2 + \sqrt{(z_3 - z)^2}} \\
d_4 &= c(t_{t,4} - t_{r,4} + t_c) = \sqrt{(x_4 - x)^2 + (y_4 - y)^2 + \sqrt{(z_4 - z)^2}}
\end{aligned}
\tag{21.10}
$$

The following Fig. 21.1 shows how these equations will be used to estimate location. Each equation represents an access point that receives a signal. Each access point should have the same receiver and transmitter target. It was determined that in order to find the location of the users, at least three access points should receive the signal or (four to get accurate location). The push button is represented by the push button of the user and the receiver side is represented by the access points.

For the transmitter, there are three unknowns X, Y, Z. The X, Y, Z parameters represent the user coordinates (i.e., push button), which are to be calculated after all variables are substituted in Eq. (21.10). There is a fourth unknown 'time correction (t_c)'—access point receiver clock—which is to be calculated after all variables are substituted in the Eq. (21.10). Thus, the requirement is to find four unknowns (X, Y, Z, tc). The variable 'c' in the equations represents the speed of signal that propagates through access points. The variables (X_1, Y_1, Z_1) represent access point 1 coordinates; (X_2, Y_2, Z_2) represent access point 2 coordinates, and so on. So, for the receiver to calculate user location, the coordinates of all access points should be known.

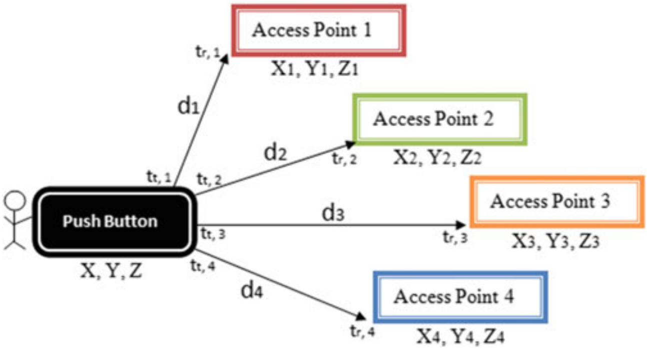


Fig. 21.1 Location estimation

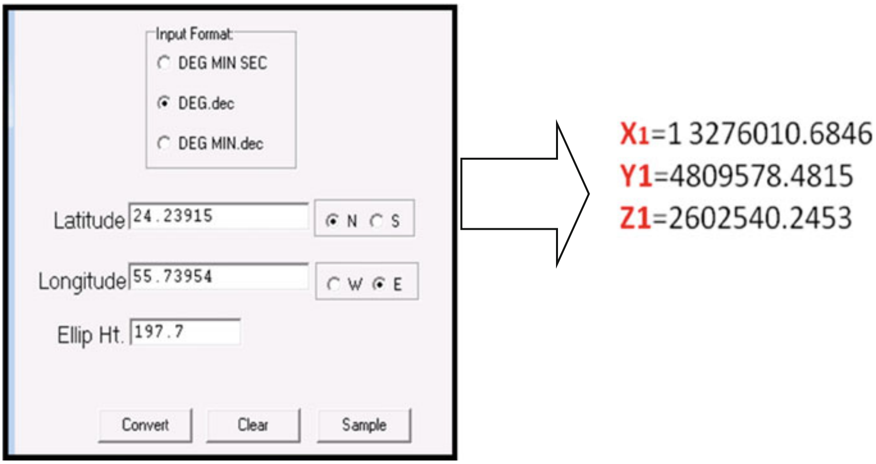
(b) Access points coordinates

In order to measure all access points' coordinates, commonly available GPS method was used and verified by Google earth software program to determine coordinates. It turned out that measuring the access point coordinates by GPS device is not that easy as GPS devices require open area to measure coordinates as opposed to access points that are inside the building. Thus, Google earth software was used to get readings in degree and decimal format for coordinates of all access points.The readings consisted of three points, which are:

(i) Longitude
(ii) Latitude
(iii) Height

The equations need only X, Y, Z format for access point coordinates in order to do calculation and find user location. A conversion program, as shown in Fig. 21.2 was used to convert degree and decimal format to X, Y, Z format. The IP address of the user device, transmit time and receive time of data packet are recorded in the receiver, where as IP address of each access point is stored in receiver application. A code written in Matlab solves the multiple non-linear equations to calculate the location of a person. Solving these non-linear equations gives the exact location X, Y, Z of the sender. In order to solve the four equations "if solve" function in Matlab was used. After an initial value is inserted for each unknown, then "if solve" function iterates till it reaches the correct value of each unknown.

To simulate this setup, the following devices were used: Microchip Explorer 16 board along with Wi-Fi PICtail, MPLAB ICD3, MPLAB IDE, PC, and Linksys Wireless-G router as access point. Additionally, an interface program was developed between Visual Basic (VB) and Matlab in order to pass parameters between them.A Visual Basic code was also written to convert real x, y values to VB x, y values. An overall view of graphical user interface for two area maps is shown in Fig. 21.3. In Fig. 21.3, '1', '2', '3', '4', '5', '6', '7', and '8' represent map view,



Input Format

○ DEG MIN SEC

● DEG.dec

○ DEG MIN.dec

Latitude 24.23915 ● N ○ S

Longitude 55.73954 ○ W ● E

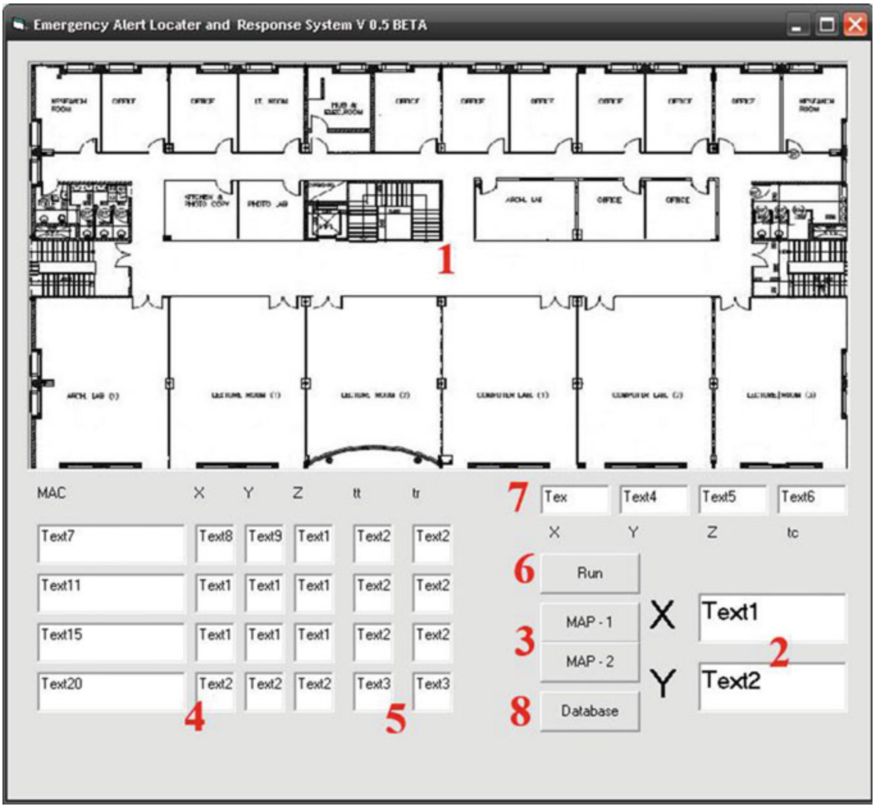
Ellip Ht. 197.7

X1=1 3276010.6846

Y1=4809578.4815

Z1=2602540.2453

Fig. 21.2 Coordinate conversion



Emergency Alert Locater and Response System V 0.5 BETA

MAC	X	Y	Z	tt	tr
Text7	Text8	Text9	Text1	Text2	Text2
Text11	Text1	Text1	Text1	Text2	Text2
Text15	Text1	Text1	Text1	Text2	Text2
Text20	Text2	Text2	Text2	Text3	Text3

Tex	Text4	Text5	Text6
X	Y	Z	tc

X Text1

Y Text2

Fig. 21.3 A sample view of graphical user interface

x-*y* coordinates, map selection, MAC address of access points, transmission and reception time, run application, showing results, and database respectively.

(c) Determining the location of the user device

If only one access point receives the signal from the holder of the device, then the location of the holder of the device will be within the range of the access point itself. But, if more than one access point received the signal, then the location of the holder of the device may be the intersection of these access points range. So, as more access points receive the signal, the more accurate and precise location of the user can be defined by the system equations, as depicted in Fig. 21.3. To make application run faster, it was determined that the Matlab code and database may be translated to VB so that the application uses only one tool to do all tasks in the receiver.

From the study of availability of similar and compatible projects, it was found out that such technologies were being used either for medical purposes or for tracking purposes (using GPS devices). There are commercial companies available, which offer this service to senior citizens or to people who need a continuous monitoring due to some health issues, so the company provides them a 24 h monitoring by letting the user have a small device such as a watch or a key chain to send a signal to another device plugged into a telephone line. One of the main issues with such designs is the concern for privacy. In our implementation, location signal would be available in the air, when user push-button is pressed.

(d) Protecting data over WLAN

Once the pressed-key data is in the air, the location of the device can be calculated. Though various techniques can be used to improve security of data transfer using approach, for example in [22–24], but it can further be boosted by the use of encryption techniques. This encryption process will only start when the user presses the key. There exist various embedded processor vendors such as Atmel, Free scale, Maxim, Microchip, NXP, PalmChip, STMicroelectronics, Texas Instruments and others, which have included random number generators and dedicated encryption/decryption engines inside their processor chips. Furthermore, several vendors offer encryption engine blocks as intellectual property. Such blocks can be embedded in a field programmable gate array along with a processor core, orco-integrated with a processor core on a custom chip.

Microchip offers security-enhanced processors based on its proprietary 8-bit PIC processor core. The examples are PIC12F635/PIC16F636/639. Simply said, these 8-bit processors include a cryptographic module, named as KEELOQ that employs a block-cipher encryption algorithm based on a key length of 64 bits and a block length of 32 bits. In one way, the quality of the encryption can be judged from the implementation of the encryption algorithm. In that, it obscures the information in such a way that if the unencrypted information differs by only one bit, the next coded response will be totally different. Statistically speaking, if only one bit changes in the 32-bit string of information, there will be a 50 % change in the coded transmission. Typically on these processors, with a device utilization of 100 %, a bit rate of about

51 Kbps can be achieved. This type of performance projects the PIC17C42 a price versus performance leader for encryption algorithms.

Based on this investigation, it was decided to embed encryption of pressed-key data (within PIC microcontroller) to be transmitted from the user device. The objective was to encrypt the identification of the device that pressed the key. As data size is small, necessary padding was done to create 16-byte data blocks. For implementation purposes, the key "This is my data" was used. Once key was selected and data partitioned, the encryption cycle was started. For this purpose, the specific Advanced Encryption System (AES) encryption libraries and routines [25] were used for Microchip microcontroller. When the data is received at the receiver, the respective decryption process is started. This process is carried out in a similar microcontroller, as a part of receiver. The parts or subdivisions in the decryption process are similar to those of the encryption, with large subset being the inverse of the other. The main point to be noted is that the decryption key is different from the encryption key, and thus needs to be loaded correctly. Another point is that the key needs to be reset between blocks during encryption process. Otherwise, the decryption key has to be adjusted accordingly.

A number of rounds may be decided for encryption/decryption. During each round of AES decryption, the same key is to be used that was employed to encrypt the data. However, the key for the next iteration is determined from the previous decryption key using the inverse operation in the encryption key schedule. To obtain the decryption key from the encryption key, one has to cycle the appropriate amount of times through the encryption key schedule. The value of the key at the end of an encryption cycle provides the correct decryption key, at that point. For this, many options exist, like saving this value; recalculating later; or pre-calculated and stored in the system ahead of time.

As the PIC microcontroller completes the decryption in real time, the location calculation is started in the system.

In case, absolute security is needed no matter what the cost, code size or speed, then the best choices are XTEA encryptions technique with 32 or more rounds of AES. Lower rounds of AES using XTEA, say 32, can generate a balance between code size, security and execution speed. As a last point in encryption/decryption process, whenever a system needs to be developed to securely talk to other systems, it has to have same encryption standard so that the communication can be deciphered.

21.8 Standardization Efforts

The standards for locating do not refer or talk about any specific method of location calculation, nor any technique for measuring locations. This may be paragraphed in specifications to include triangulation or any hybrid method to trigonometric computation of planar or spherical models for terrestrial area. The basic issues in RTLS are standardized under ISO/IEC 24730 series by the International Organization for Standardization and the International Electrotechnical Commission. The basic standard

ISO/IEC 24730-1, in this series of standards, identifies and describes the terms used by a set of vendors for RTLS. It does not, however, specify the full scope of RTLS technology.

Based on survey conducted, the following notable standards are published or under discussion:

- ISO/IEC FDIS 19762-5 Information technology AIDC techniques—Harmonized vocabulary, Part 5—Locating systems
- ISO/IEC 24730-1:2006 Information technology real-time locating systems (RTLS) Part 1: Application program interface (published).
- ISO/IEC 24730-2:2006 Information technology real-time locating systems (RTLS) Part 2: 2, 4 GHz Air interface protocol (published, where Net/Zebra approach).
- ISO/IEC WD 24730-5 Information technology real-time locating systems (RTLS) Part 5: (drafted ISO/IEC standard out for balloting in 2008, Nanotron approach).
- ANS/INCITS 371 series: Information Technology—Real-Time Locating Systems (RTLS). The Committee approved three new standards in 2003 that define two Air Interface Protocols and a single Application Programming Interface (API) for Real Time Locating Systems (RTLS), especially for use in asset management.

A lot of work for location detection standardization in the domain of 3G mobile phone systems has been carried out. A summary of standardization for 3G, 3GPP, and 3GPP2 enabled mobile systems is provided in [26].

21.9 Conclusions

The real time locating systems were investigated for variety of uses, privacy and data transfer security. As a competitor for the purpose of privacy and data security, a system was developed in laboratory to determine location positioning during emergency duration to secure privacy at operational times. This was enabled by a push button and encryption of data transfer. The accuracy of location detection is however, dependent on the method used to calculate it. If only one access point receives the signal from push button device, then the location range is the coverage area of one access point only. In case two access points receive the signal, then the accuracy is within intersection of these access points' range. Thus, if more access points receive the signal, better accuracy is ensured. In conclusion, the implemented system boosts privacy by (a) no data transfer except during emergencies (b) encrypting data during transfer. Such a deployed system facilitates the safety management departments to address personal safety and security of the user in an operational area.

Acknowledgments The author would like to express his sincere thanks to the Research Affairs unit of the College of Engineering at the United Arab Emirates University for encouragement and financial support of this project under grant code number G00000814.

References

1. Gibson, B., Cory, T.: Portable Navigation and Wireless Tracking: Western Eurpoean Markets and Forecasts 2007–2012, Juniper Research: www.juniperresearch.com. Accessed on November 11, (2010)
2. Mohan, P., et al.: Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In: Proceedings ofthe 6th ACM Conference on Embedded Networked Sensor Systems, US (2008). doi:10.1145/1460412.1460444
3. Brown, B.: 802.11: the security differences between b and I. IEEE Potentials **22**(4), 23–27 (2003)
4. NBS FIPS PUB 46: Data Encryption Standard. National Bureau of Standards, US Department of Commerce (1977)
5. SO DIS 8730: Banking Requirements for Message Authentication (Wholesale). Association for Payment Clearing Services, London (1987)
6. ISO DIS 8732: Banking Key Management (Wholesale). Association for Payment Clearing Services, London (1987)
7. Quercia, D.: Recommending social events from mobile phone location data. IEEE Int. Conf. Data Mining **327**(5971), 971–976 (2010)
8. Hopper, A., Harter, A., Blackie, T.: The Active Badge System. INTERCHI'93, Amsterdam (1993)
9. http://research.microsoft.com/en-us/projects/radar/. Accessed on April 2012
10. LaMarca, A., et al.: Place Lab: Device Positioning System using Radio Beacons in the Wild, Intel Research, IRS-TR-04-016 (2004)
11. Tharp, D., Wallace, L.: Enhanced position location reporting system: legacy system provides new technology for Warfighters. Navigation and Applied Sciences, SSC San Diego Biennial Review, pp. 206–211 (2003)
12. Nissanka, B.P., Anit C., Hari, B.: The cricket location-support system. In: Proceedings of 6th ACM MOBICOM, Boston, MA (2000)
13. Karlsson, R.: Particle Filtering for Positioning and Tracking Applications. PhD Thesis, Linkoping University, SE-581–83, Sweden, Linkoping (2005)
14. Ananthanarayanan, G., et al.: StarTrack: A Framework for Enabling Track-Based Applications, Microsoft Research (2011)
15. Kannan, B., et al.: Predictive Indoor Navigation using Commercial Smart-phones. In: 28th Annual ACM Symposium on Applied Computing (2013)
16. LocataNet Positioning Signal Interface Control Document-2011, Locata Corporation Ltd, 111 Canberra Avenue, GRIFFITH ACT 2607, Australia (2011)
17. Malik, A.: RTLS For Dummies. Wiley, Hoboken (2009)
18. Koppers, J.: Positioning Techniques: A general model: http://www.positioningtechniques.eu/. Accessed on June 5 (2012)
19. Torrieri, D.J.: Statistical theory of passive location systems. IEEE Trans. Aerospace Electron. Syst. **20**(2), 183–198 (1984)
20. Friedlander, B.: A passive localization algorithm and its accuracy analysis. IEEE J. Oceanic Eng. **12**, 234–245 (1987)
21. Ananthanarayanan, G., Haridasan, M., Mohomed, I., Terry, D., Thekkath, C.: StarTrack: a framework for enabling track-based applications. In: Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services, Kraków, Poland (2009)
22. Memon, Q., Kasparis, T.: Transform coding of signals using approximate trigonometric expansions. J. Electron. Imaging **6**(04), 494–503 (1997)
23. Svitek, M., Zelinka, T., Lokaj, Z.: ITS data security in wireless telecommunication solutions. In: IEEE Colombian Intelligent Transportation Systems Symposium, pp. 1, 6, 30–30 (2012). doi:10.1109/CITSS.2012.6336682
24. Memon, Q., Kasparis, T., Tzannes, N.: Approximate fourier expansion with uncorrelated coefficients. In: IEEE Mediterranean Symposium on Advancements in Controls, Cyprus (1995)

25. Flowers, D.: Data Encryption Routines for PIC18 Microcontrollers (2011), http://www.microchip.com/stellent/idcplg?IdcService=SS_GET_PAGE&nodeId=1824&appnote=en022056. Last Accessed on Oct 4 (2012)
26. Zhao, Y.: Standardization of mobile phone positioning for 3G systems. In: IEEE Communications Magazine, pp. 108–116 (2002)

Chapter 22
Towards Cloud Customers Self-Monitoring and Availability-Monitoring

Sameh Hussein and Nashwa Abdelbaki

Abstract As an attractive IT environment, Cloud Computing represents a good enough paradigm which governments, national entities, small/medium/large organizations and companies want to migrate to. In fact, outsourcing IT related services to Cloud technology, needs monitoring and controlling mechanisms. However, Cloud Customers cannot fully rely on the Cloud Providers measurements, reports and figures. In this book chapter, we cover the two Cloud Computing operation sides. For the first operation side, we provide advices and guidelines for Cloud layers which can be under Cloud Customer control, to allow Cloud Customer contributes in Cloud infrastructure monitoring and controlling. For second operation side, we produce our developed monitoring tool, to allow Cloud Customer contributes in service monitoring. It is for Cloud Customers to self-monitor the Availability as a metric of the outsourced IT service.

22.1 Introduction

Network management is one of the areas which is continuously evolving, widely demanded, and appeared with complex/large networks. It was one of the key components that is discovered when scientists were researching the broad subject of managing computer networks. There exist hundreds of software and hardware products that help network system admins to manage networks under their supervision [1]. Also, there is a variety of tools which guarantee full control over these networks [2]. Network management covers a wide spectrum including security, performance, reliability, class of service, etc.

S. Hussein (✉) · N. Abdelbaki
School of Communications and Information Technology, Nile University, Cairo, Egypt
e-mail: sameh.hussein@nileu.edu.eg

N. Abdelbaki
e-mail: nabdelbaki@nileuniversity.edu.eg

A. E. Hassanien et al. (eds.), *Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations*, Intelligent Systems Reference Library 70,
DOI: 10.1007/978-3-662-43616-5_22, © Springer-Verlag Berlin Heidelberg 2014

Network monitoring is more strategic than its name means. It demands watching for problems on 24/7 manner. Moreover, it’s also about optimizing data flow and data access in a complex and changing environment [3]. Services and tools are as numerous and varied as the environment they analyze changes.

In network management world, network monitoring is the proof of concept used to describe the monitoring system. It continuously monitors the network and notifies the network system administrator via messaging system [4]. Usually, notifications are sent in case of a device fails, lack of connectivity, or an outage occurs [5]. Notifications are through E-mails, SMS, warning messages, or alerts. However, network monitoring is performed through the use of tools and software applications [6].

The previous paragraph leads us to a very important question. What can network monitoring systems monitor? Monitoring network will not help, unless we know the right things to be monitored according to service nature, SLA/SLO metrics, and security constrains [7]. Usually, network monitoring is examining bandwidth usage, application performance and server performance. As a fundamental task, traffic monitoring is one of which network maintenance/building tasks are based on [8].

However, network monitoring systems have evolved to oversee an assortment of devices such as, switches, routers, servers, desktops, backbone devices, network nodes, cell phones, and others related. Moreover, network monitoring systems may come with auto-discovery functions, which is able to continuously log and record devices as they are joined, leaved, or undergo of configuration changes [9, 10]. Like such functions, segregate devices dynamically based on rubrics such as IP address, service, type (switch, router, etc.), and physical location [11].

It is an obvious advantage of knowing exactly (and in real time) what has been deployed and what has been automatically discovered to help monitoring. Underused hardware can provide new functions which help pinpoint problems [12]. As an example, if most of the connected devices at a given area are underperforming, then, there might be a resource management problem to be addressed [13].

On the other hand, business ability to link network monitoring with the provided services, moves the strategic interest to service monitoring instead of network monitoring. However, the deep understanding of the service provided leads to determine SLA/SLO characteristics as well as the service metrics that are necessary to be monitored [14]. For example, when we have a website as a service, some metrics are vital to be measured such as, availability, response time, performance, network connectivity, DNS records, database injections, bandwidth, and computer resources like free RAM, CPU load, disk space, and others [15].

Throughout the rest of this chapter, we visit Cloud Computing monitoring and controlling. We examine Cloud layers versus the three basic and main implementation models, address the conflict between Cloud Customer and Cloud Provider, produce recommendations and guidelines for layers under Cloud Customer control, then discuss Cloud service availability to produce our developed Availability Monitoring tool and its flow chart, we examine our tool in test environment. Finally, we conclude and expect the future.

22.2 Monitoring and Controlling Cloud Computing

Measurement climate and monitoring weather get changed once Cloud Computing becomes the atmosphere and the hardware environment of the service to be outsourced. Nothing more than Cloud Computing has different nature compared with ordinary service providers. This difference in nature is steaming from Cloud Computing characteristics like, on demand self-service, elasticity, metered service and ubiquitous access. From business prospective, hosting IT services on public, private, or hybrid cloud is troublesome without appropriate metrics measurements. Where unified visibility, control and awareness of the entire cloud infrastructure is required to monitor cloud operations.

Cloud Computing monitoring has two operational sides. The first is to monitor the core infrastructure of the cloud [16, 17]. It has benefits for the Cloud Provider like increase servers and network equipment availability, fast detection of network outages, and fast detection of Cloud Computing environment problems. The other operational side is to monitor an assortment of service related metrics, to guarantee that the delivered services are matching with the agreed quality levels.

The first operation side can be monitored and controlled via monitoring and controlling Cloud layers. One of its problem is the conflict of interests between Cloud Customers and Cloud Providers. More clarifications of the Cloud layers, conflict of interests problem, and the proposed solution are discussed within the following sections.

The second operation side can be monitored and controlled via monitoring and controlling some selected service metrics. One of its problem is that Cloud Providers sometimes report inaccurate measurements and misleading figures of the Quality of Service metrics. We have developed an Availability Monitoring tool which allows Cloud Customers monitor service availability to compare its results with the ones reported by the Cloud Provider.

However, it is like the flip coin game, as the Cloud Provider flips the Cloud to operate where the Cloud Customer would like to monitor both operation sides. Figure 22.1, represents so.

22.2.1 Monitoring and Controlling Cloud Layers

However, According to Cloud Security Alliances (CSA) work [18], Cloud Computing can be layered into seven layers. Like the rainbow, each color represents a layer in the spectrum. They are Facility (F), Network (N), Hardware (H), OS (O), Middleware (M), Application (A), and User (U). Exactly as the raining weather, a Cloud rains the layers which are allowed for Cloud Customers to monitor and Control. In fact, Cloud atmosphere which represent the implementation model (IaaS, PaaS, SaaS), decides which of these layers are under Cloud Provider control, and which are under Cloud Customer control. In Fig. 22.2, a nature scenario which implements what we were explaining.

Fig. 22.1 Flip Coin Game

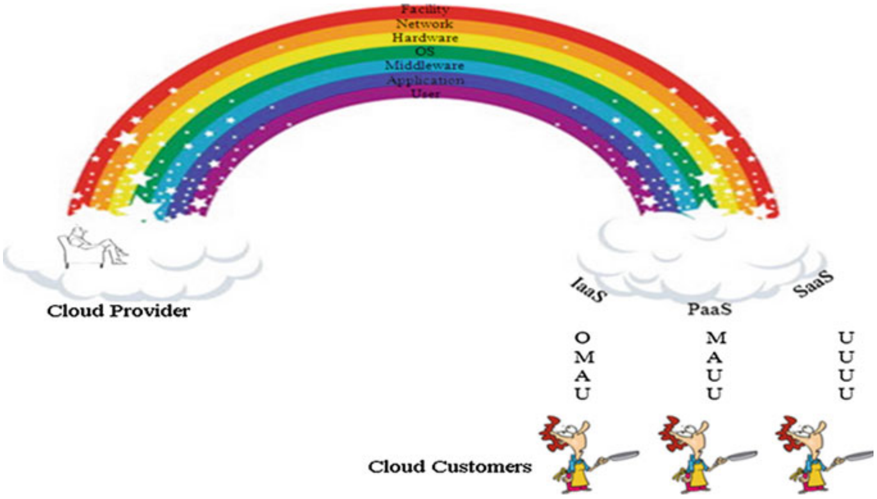


Fig. 22.2 Cloud layers, implementation models, provider versus customer control

As shown in the above Fig. 22.2, Cloud Customers are able to monitor and control the Cloud till a certain depth according to the implementation model. Each layer is linked with its previous and next layer. In SaaS, the Cloud will rain User layer for Cloud Customers to monitor and control. This because other layers are completely managed by the Cloud Provider.

In PaaS, the Cloud will rain User, Application, and Middleware layers for Cloud Customers to monitor and control. For this model, Middleware layer will be a negotiated one, where both Cloud Provider and Cloud Customer should decide who will have hands on it. Usually, whoever will control it, layer operation recommendations should be shared with the other side. Other layers are completely managed by the Cloud Provider.

In IaaS, the Cloud will rain User, Application, Middleware, and OS layers for Cloud Customers to monitor and control. For this model, OS layer will be a negotiated one, where both Cloud Provider and Cloud Customer should decide who will have hands on it. The same as before, whoever will control it, layer operation recommendations should be shared with the other side. Other layers are completely managed by the Cloud Provider.

Regardless which layer is under Cloud Customer supervision, the Cloud Provider always sits away. Not doing nothing, but for overall management of the entire Cloud as well as remote monitoring and controlling for the left layers.

22.2.2 Cloud Customer/Provider Conflict of Interests

Day after day, Cloud Customers discover new traps and new backdoors for the Cloud technology. This pushes them to negotiate more and more with the Cloud Providers, looking for more visibility and more management over the Cloud layers. This might not be possible in the public Clouds, but for sure, it can be achieved in the private Clouds.

However, any Cloud Provider is used to be keen enough to keep as much layers under his control. On the other side, Cloud Customer is afraid having troubles. Then, Cloud Customer seeks more layers for monitoring and controlling, especially when new drawbacks get discovered. Thus, we have a conflict of interests. Usually, new traps and backdoors tumble customers business in terms of availability, accessibility, continuity, and others which will have financial influences.

The shown Fig. 22.3, represents a scenario where conflict of interests takes place. As human being, Cloud Customer will be very happy running away with the Cloud to try to serve his business. To quickly achieve so, Cloud Customer needs to have control over more layers. However, it is not that easy, the Cloud is bounded by the layers under Cloud Provider control. Also, it might be controlled by other Cloud Customers, in case of public Clouds.

Therefore, it depends on the Cloud atmosphere and its consequences, to determine the area which Cloud Customer is allowed to drive the Cloud within it. As shown in Fig. 22.3, Cloud Provider is used to get back the control of the cloud. Cloud Provider tries to bound the cloud by a wire which is fixed in the ground.

22.3 Cloud Layers Under Cloud Customer Control

As we mentioned before, we always have conflict of interests, where Cloud Customer will have control over some Cloud layers. Regardless the Cloud atmosphere, Cloud Customer will never have a control over deeper layers. At maximum, OS layer, where

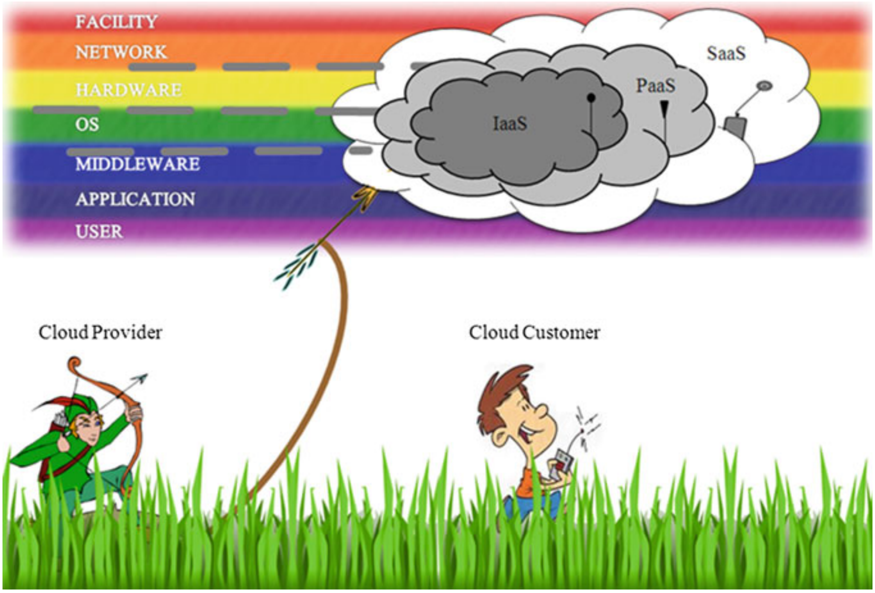


Fig. 22.3 Cloud customer/provider conflict of interests

the agreed recommendations and guidelines will be deployed. In the following we discuss the four Cloud layers which can be under Cloud Customers control within deferent implementation models.

OS, The cloud based OS were evolved to own four mature roles. It acts as well defined interfaces that hide all implementation details. It is responsible for core security services. It manages, hosts, controls, and assign resources for virtualization. It also manages the workloads to ensure quality of service and performance. Therefore, Cloud Customer should be keen to deploy highly secured and controlled OS. Fundamentally, the deployed OS should be cleaned from all additional and non-essential functions. Because only the necessary functions should operate over the OS, these functions should be checked thoroughly for backdoors and vulnerabilities before installing. Also the OS itself must be immune against compromising. However, all system calls between VMs and hardware should be controlled and monitored by the OS. Thus, OS has access to all data passing to or from the VMs, as VMs transferring and processing plaintext data. On the other hand, it also has access to all data stored on VMs, because it is stored on disks which are controlled by the OS, but data can be stored after encryption, where the OS doesn't have its key. Cloud Customer has to ensure monitoring of VMs logs and binary changes, and any offending change has to be returned into a known good state, where monitoring memory dumping and processor over utilize need to be investigated to define and configure new security policies to prevent similar incident. Reports of hardware and software regarding performance should be matched and shared with both customer and provider.

Middleware, as a term, has wide range of definitions. As a simple form, it is a software that connected computers with databases. For Cloud Computing, middleware is a floppy topic that extends from virtualization management tools, to data format conversion. It needs to run security functions for dynamic cloud architectures. Although middleware is important for Cloud Computing, it can be a significant potential weak point for customers and providers when deploying information security assurance mechanisms. Middleware as a concept, still immature layer, especially for cloud computing. However, this layer is the natural place to monitor and secure communication between various system components because it mediates between the applications and the OS, where there are various safeguards to be implemented and pitfalls to be avoided. Then, customers should ensure that middleware will accept and transmit encrypted data. When the customer takes the control over middleware, the provider should protect it against malicious manipulation. As the middleware tends to gain rights to access, manipulate and distribute data, beside specialized functions such as managing access controls, it would be damaging the OS as modifying the OS. To guarantee the avoidance of related concerns, provider should be ready for customer misconfiguration of resources and policies as well as abusing of middlewares functions. For sure, the provider need very intelligent and sophisticated monitoring system for the middleware, but it is very difficult. Also provider needs code inspection tools to scan middleware coding vulnerabilities.

Application, it represent the software hosed by the Cloud Computing. Customers should seek applications in which its source code and business logic have been carefully examined by neutral entrusted third parties for potential flaws and deficiencies. Application must be holding the standards of best practice like sanitizing of all user inputs. In traditional environment, a host based security system can monitor abnormal behaviors in the operating applications. However, in cloud environment, monitoring system should keep track of all violations for each running application. It is difficult to have so, because one instance of an application may serves multiple users simultaneously and doesnt reside on a dedicated host. An application may sit in memory to accomplish multiuser nature of cloud computing. Then, application compromising may lead to memory dump which needs corruption detection mechanisms. When the layer be under customer control, the only different from a traditional computing model is that the monitoring will also be virtualized. Then, it allows for a more costbenefit analysis of monitoring different metrics. This is due to the Cloud architecture inherent scalability and flexibility. In SaaS, providers might develop customized monitoring solutions. However, it should be able to describe those monitoring and remediation strategies in detail.

User, We have two kinds of users. First, is the stand alone users who seek cloud webpage or video services, they have little security impact. Second, users who are members of the customer organization, they should comply with organization security policies. However, both kinds for users access should be monitored and controlled against malicious behaviors. Any aberrant, abnormal or anomaly user of the service should be logged. Like such alerts and notifications should be reported to IT managers of accounts for which their organization is responsible. However, IT security party must add proscribing access to sensitive data in public areas.

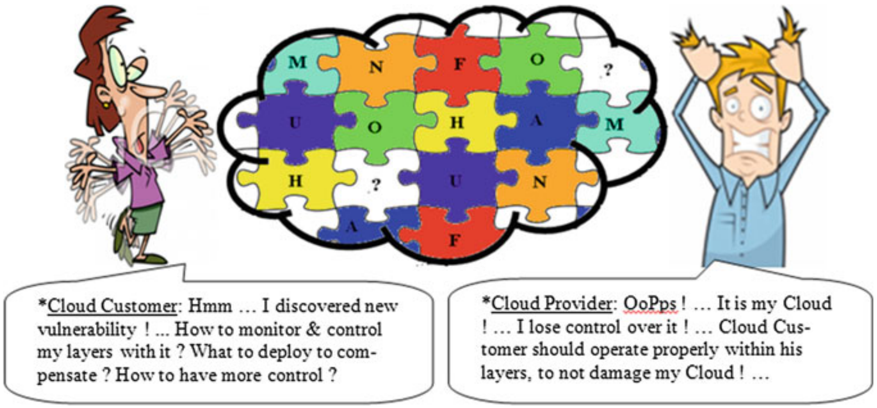

M
N
F
O
?
U
O
H
A
M
H
?
U
N
A
F
*Cloud Customer: Hmm … I discovered new vulnerability ! ... How to monitor & control my layers with it ? What to deploy to compensate ? How to have more control ?
*Cloud Provider: OoPps ! … It is my Cloud ! … I lose control over it ! … Cloud Customer should operate properly within his layers, to not damage my Cloud ! …

Fig. 22.4 Puzzle game (cloud customer versus cloud provider)

However, managing Cloud layers is more or less similar to the puzzle game (Fig. 22.4), where each player wants to lead putting the missing part to own it. Also, each player thinks of how to control and how to monitor, according to the new traps and backdoors. In our case, Cloud Customer is looking for mechanisms and approaches to benefit his business the most. Therefore, Cloud Provider is looking for how to make his Cloud secure and safeguard his Cloud against Cloud Customer abuse. Then, each one is playing the puzzle game based on his experience and business needs.

22.4 Cloud Service Availability

To address Cloud Computing availability, we can say it is the number one Cloud Customer priority. Since Cloud Computing became a great choice for the IT needs of large companies and organizations all over the world, Cloud Providers were faced with many obstacles that threatened to bring the development and expansion of this technology to a halt [19, 20]. The fact of the matter is that making applications highly available is very difficult. It requires highly specialized and sophisticated tools, systems and trained staff. Furthermore, it is very much expensive. Many Cloud Providers are required to run multiple data centers due to high availability requirements (usually for customers business requirements). Some Cloud Providers have data centers in a standby mode, waiting to be used in a case of a failover [21]. Other Cloud Providers are able to achieve a certain level of success with active/active data centers, where all data centers are ready for incoming user requests. Achieving high availability for services is relatively not easy, establishing a highly available database farm is far more complex. Actually, it is very complex for many companies to establish yearly tests to validate failover procedures.

Being not able to keep services available 24/7 is what all providers fear the most, as even the slightest mishap will have painful consequences on their clients business workflow and reflects on the trust. When we think about it, it is like hypothetically buying the services of Google and not being able to perform online searches.

22.4.1 Cloud Availability Notable Comments

Addressing availability as a metric to be measured, is more or less vital for Cloud Customers. As it means whether the outsourced services are alive or not. Usually Cloud Customers are looking for 100.

First is the planned outages, which can be carried out due to maintenance window, software update, equipment upgrade, install new license after renewal, site migration, adoption of new technology, service upgrade or downgrade, delayed paid installments fee, or customers ask for service suspension [22, 23].

Seconds is the unplanned outages, which can be carried out due to power failure, hardware failure, software failure, network failure, authentication failure, bad configuration, wrong setup, external and internal attacks, or security breaches. Although, there are more reasons behind service lack of availability, but monitoring it and reporting its results, should be totally independent on the actual reasons. At the end of the day, there will be a percentage of service availability, in which both parties should be keen enough to pursue [22, 23].

Raising the point of achieving or not achieving the desired availability percentage, will lead the decision makers to allow compensations and penalties terms and conditions take place according to contact clauses and SLA/SLO financial terms.

On the other hand, the measured availability percentage should be multiplied by the event severity. In other words, when a Cloud Customer experiences lack of service availability (whatever the reason is) in weekends and public bank of holydays, they will raising alerting messages to their Cloud Provider with moderate severity. However, when they suffer the same within normal business days (especially rush hours, where heavy transaction are performed), a strong and high management level channel should be held between both Cloud Customers and Cloud Provider (with very high severity) to ensure that service will be restored within a time window according to SLA/SLO/QoS terms and conditions.

Furthermore, Cloud Provider should be ready with alternatives to guarantees that customers still a live with minimum interruptions. However, like such scenarios should trigger SLA monitoring team and legal department to focus on and activate compensations and penalties terms and conditions.

22.4.2 Surveying the Existing Cloud Availability Monitoring Tools

Nowadays, hundreds of powerful tools are available. Some are for specific service metrics, and others are comprehensive. Some are for LANs, and others are for WANs.

Some are generic to operate within any platform, and others are platform dependent. Some are for general purposes, and others are for specific purposes. Some are made using standard/known programming languages, and others are using special programming languages. Some are offering basic functions, and others are offering advanced functions. However, most of the well-known monitoring tools are using PING command. In fact, it is a programmer decision, where other SNMP commands still can be used. PING command is being widely used by programmers and demanded by Cloud customers for its great benefits. We discuss these benefits through the next section.

As an example of PING monitoring tool, Ping Plotter, EMCO Ping Monitor, Ping for life, Kaseya Ping Monitoring, NirSoft Ping Info View, SoftPedia Ping Monitor, etc.

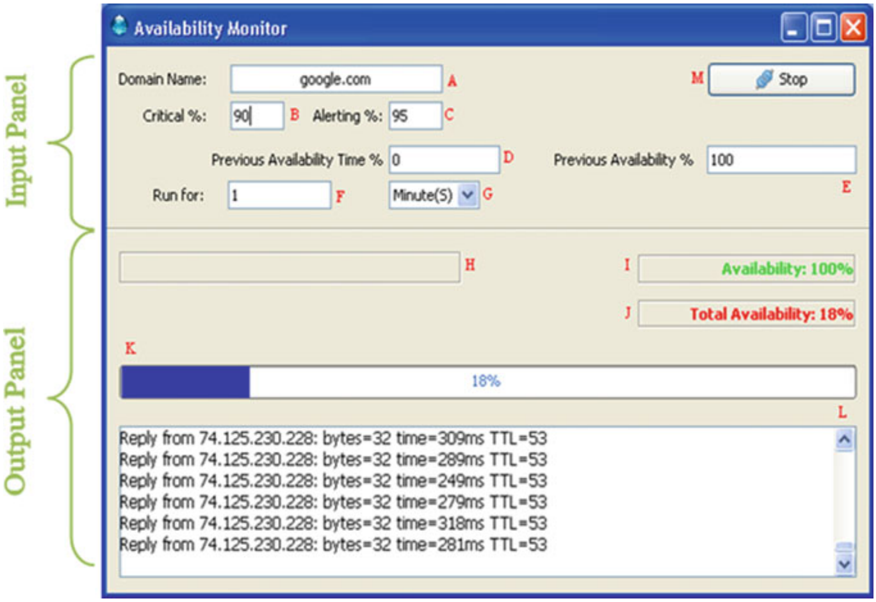
We can say, that most of these tools are using PING commend for monitoring the availability. PING is considered a type of network monitoring tool at the most basic level. Within the commercial context, other software packages can include a network monitoring system that is developed to monitor an entire business or enterprise network. Some tools and software applications are used to monitor network traffic, such as VoIP, video streaming, mail server, and others.

As common features offered by most of the availability monitoring tools, we have Connection Status Tracking, Connection Loss and Recovery Detection, Regular PING Statistics, Connection Quality Report, Configurable Event Handlers, Alerts and Notifications, Custom Event Handlers, Configurable Terminate Actions, E-Mail and SMS Notifications, Pause/Continue Button, etc.

On the other hand, we found how it is badly in need to have monitoring tools which are measuring the accumulative value of service metrics. This is because usually when Cloud Customers start self-monitoring they would need to append the previous findings. Furthermore, Cloud Customers usually seek advanced technical analysis for further investigations. This is in case of sudden or gradual changes in Quality of Service metrics. To do so, a monitoring tool needs to log all sent and received commands or replies. However, although Cloud Customers dream with the idea of having JAVA developed tool for mobility and portability purposes, it is rarely found. Both missing features has been developed to be offered through the using of our developed tool.

22.5 Our Developed Availability Monitoring Tool

Our developed Availability Monitoring tool allows Cloud Customers to have their own view and calculations over the outsourced service availability instead of total dependency on Cloud Providers reports and measurements. However, Cloud Providers still suffer lack of round-the-clock service, this actually results in frequent outages (planned or un-planned). Then, It is important to monitor the service being provided using internal or third-party tools.



Domain Name: google.com
Critical %: 90
Alerting %: 95
Previous Availabilty Time % 0
Previous Availabilty % 100
Run for: 1 Minute(S)

Availability: 100%
Total Availability: 18%
18%

```
Reply from 74.125.230.228: bytes=32 time=309ms TTL=53
Reply from 74.125.230.228: bytes=32 time=289ms TTL=53
Reply from 74.125.230.228: bytes=32 time=249ms TTL=53
Reply from 74.125.230.228: bytes=32 time=279ms TTL=53
Reply from 74.125.230.228: bytes=32 time=318ms TTL=53
Reply from 74.125.230.228: bytes=32 time=281ms TTL=53
```

Fig. 22.5 Availability monitor tool GUI

Our developed tool performs PING command continuously for the given domain name, then it collects all replies for analysis. If the returned values of the PING command is TTL, then we consider the desired equipment/service/network is alive and available. On the other hand, if the returned values of the PING command is Timeout, then we consider the desired equipment/service/network is unavailable.

During the running of this function, the tool logs the replies history for further investigations and deep analysis. It uses some probability and statistics mathematical function to calculate and display the availability and the accumulative availability depending on user inputs.

Because we are targeting of an easy and friendly interface, we used JAVA programming languages. In fact, there are 3 billion devices are using JAVA, this clearly shows us how much our developed tool will be compatible with many operating systems.

The above shown tool GUI, Fig. 22.5, was designed to be simple, friendly and easy to use. It can be run over any platform including the recent devices mobile phones, DPAs, and mobile computers. It also can be converted to operate over smart phones like I-Phone, tablets and mobile computers. It uses standard Java classes, and needs minimum resources, in terms of memory, processing, and bandwidth. Tools GUI has two main panel. The Input Panel, for all input fields, program expects user to modify the default values, and the Output Panel, for all output fields, program show and represent the calculated values in this panel and show the progress percentage. However, the Table 22.1 is defining each field of the tool GUI.

Table 22.1 Tool GUI fields function

Letter	Field name	Function
A	Domain name	To enter either desired IP address, device name or URL
B	Critical %	To enter the defined critical range, which will red color the calculated availabilities
C	Alerting %	To enter the defined alerting range, which will orange color the calculated availabilities
D	Pervious availability time %	To enter how long the previous availability lasts
E	Pervious availability %	To enter percentage of the appended availability
F	Run for time	To enter the specified time for the tool to run
G	Time unit	A drop down list to select run time units
H	Error bar	To display errors due to wrong values entered
I	Availability result	To display the colored calculated current availability
J	Total availability result	To display the colored calculated total availability
K	Progress bar	To show how long has the tool being run
L	History log box	A text box where all returned values and replies logged
M	Start/stop button	A button where user can start and stop the tool any time

22.5.1 Flowchart of Our Tool Operation

Exactly as any developed tool, it is highly recommended to deliver the operation flowchart for tool users. Once the user starts to run the .exe file, GUI will appear and then the tool becomes operational to loop inside the flowchart. It keeps running till the user ends it by closing GUI window. The Flowchart in Fig. 22.6, represents all stages, branches, and possible scenarios of tool operation including invalid input parameters. There is only one process that can be triggered any time during the tool operation (running or idle stages), it is the green one. On the other hand, the button STOP it can be clicked any time, but the button START cannot be clicked unless all parameters are entered. Therefore we set default values for each parameter. Table 22.2, shows the default values for each fields in the input panel.

22.5.2 Tool Examination in Test Environment

In this section we will show an example for analyzing a logged history when we were monitoring its availability using our developed tool. Before we go through that, we have to set a group of assumptions and environmental factors, then show the logged history, show the 2-D graphs, then analyze and comment them:

22.5.2.1 Analysis Assumptions

1. We are measuring the availability and other quality related factors.
2. We assume a live and reachable server to perform our measurements.

Fig. 22.6 Availability monitor tool flowchart

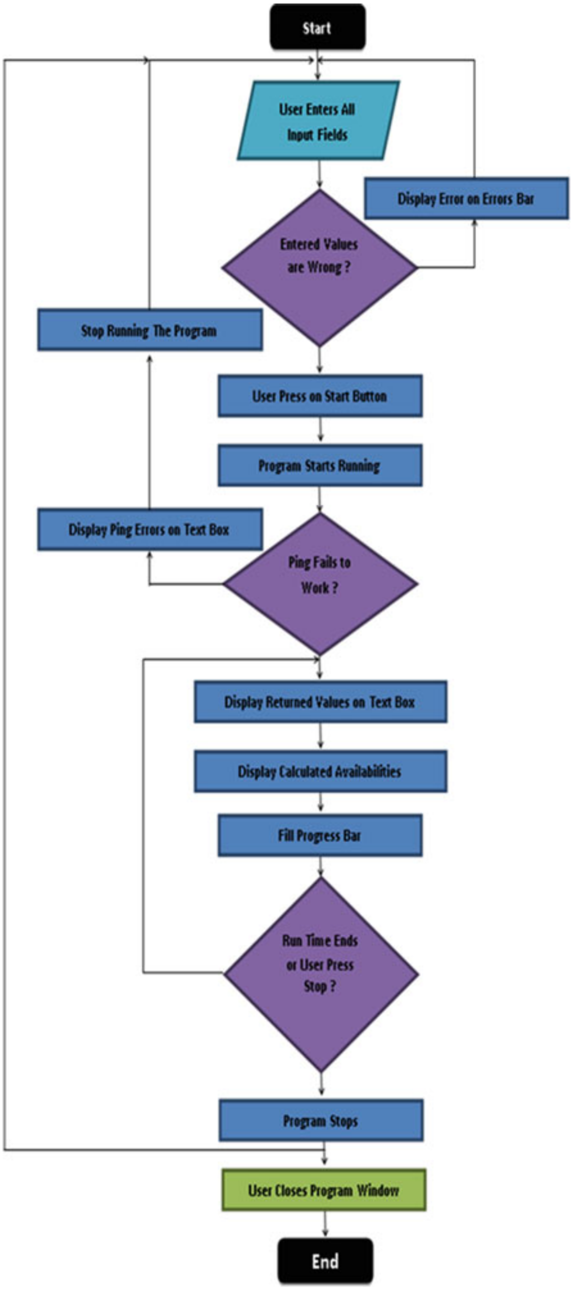

Start
User Enters All Input Fields
Display Error on Errors Bar
Entered Values are Wrong ?
Stop Running The Program
User Press on Start Button
Program Starts Running
Display Ping Errors on Text Box
Ping Fails to Work ?
Display Returned Values on Text Box
Display Calculated Availabilities
Fill Progress Bar
Run Time Ends or User Press Stop ?
Program Stops
User Closes Program Window
End

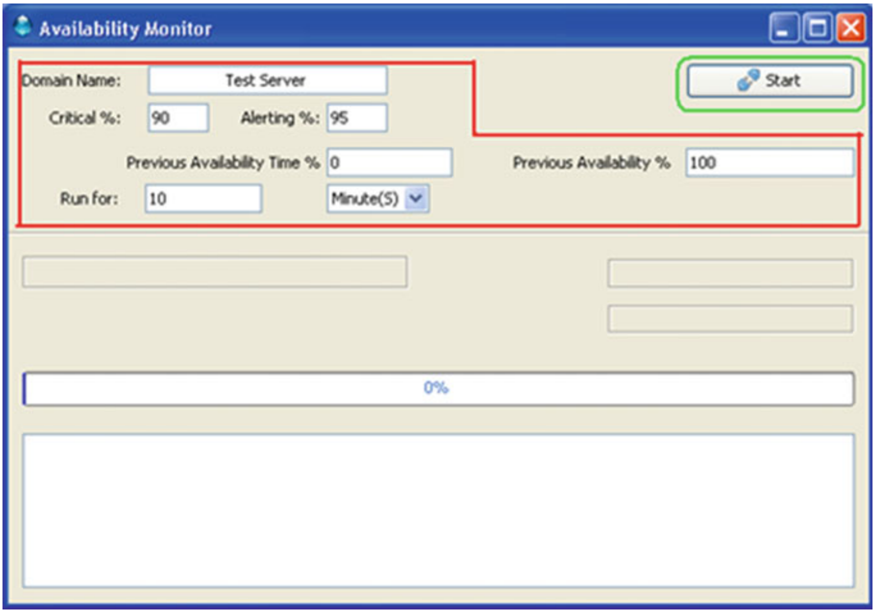
Table 22.2 Defaul values of input fields

Letter	Field name	Function
A	Domain name	127.0.0.1
B	Critical %	90
C	Alerting %	95
D	Pervious availability Time %	0
E	Pervious availability %	100
F	Run for time	1
G	Time unit	Minutes

3. We assume no software, hardware or networks difficulties.
4. We assume PING traffic is permitted between server and monitoring PC.
5. We assume desired server between replying and not replying.
6. We assume monitoring PC is up and running probably.
7. We assume the unaltered and integrity for the returned values.
8. We assume the monitoring tool works in healthy enough environment.
9. We assume DNS and DHCP servers are alive.
10. We assume DNS and DHCP servers are working probably.
11. We assume tool user is able to run it probably.
12. We assume no appended previous availability.
13. We assume availability thresholds are standard.
14. We assume this example as a part of long term monitoring.

22.5.2.2 Analysis Environmental Factors

1. We monitor in test environment.
2. We monitor in LAN network topology.
3. We monitor a server located within the same LAN of monitoring PC.
4. DNS, DHCP, Monitoring PC, and servers are located in same LAN.
5. Desired server has domain name: Test Server.
6. Desired server has as record in the local DNS server.
7. Desired server has an IP address of: 192.168.1.110
8. Monitoring PC has an IP address of: 192.168.1.10
9. We analyze the logged history via Microsoft Excel.
10. We will show 2-D graphs for our analysis.
11. We created a Macro program to do the same analysis in future.
12. The Macro is used to repeat analysis procedures on an Excel file.
13. The created Macro is usable for any Excel edition.
14. The created Macro is usable for any logged history.
15. We set Critical
16. We set Alerting
17. We set Previous Availability Time



Domain Name: Test Server
Critical %: 90
Alerting %: 95
Previous Availability Time % 0
Previous Availability % 100
Run for: 10 Minute(S)

Fig. 22.7 Snapshot of the tool GUI before running

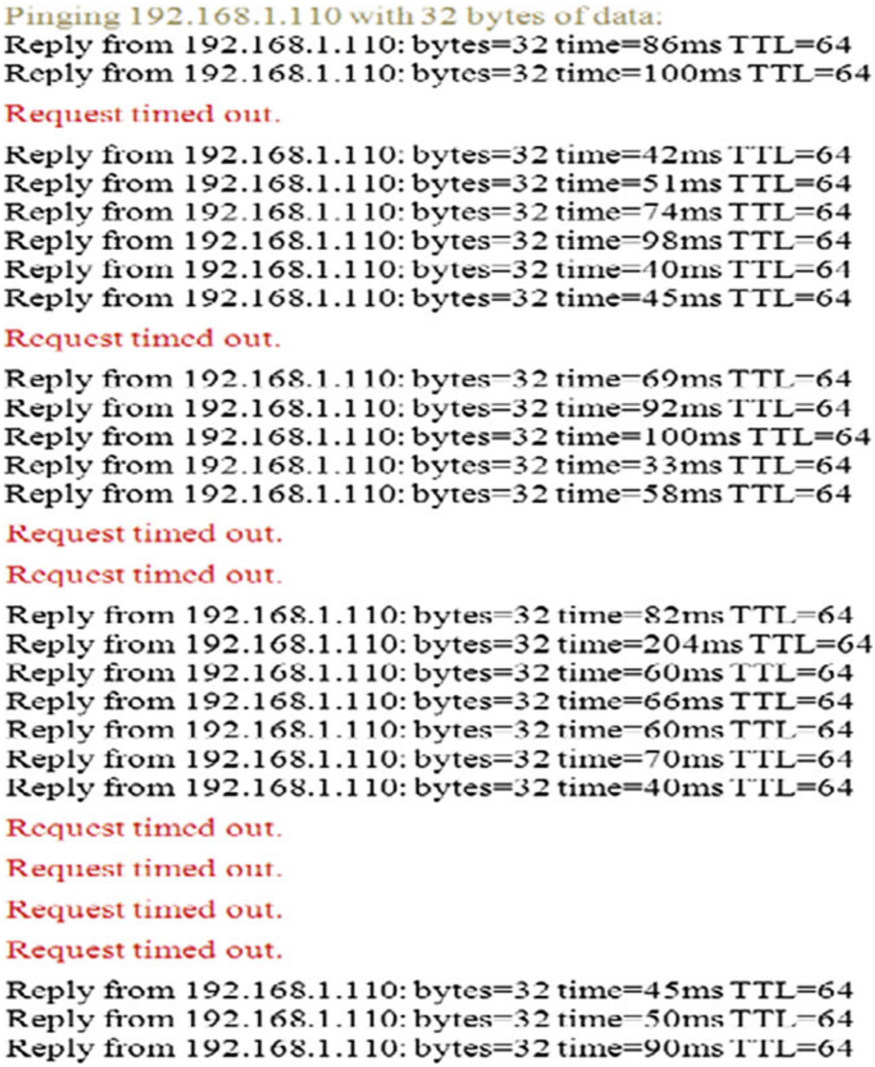
18. We set Previous Availability
19. We set Run For time = 10 min.
20. We analyze a sample of the 10 min logged history.

The above, Fig. 22.7, we show the monitoring tool with values entered according to environmental factors mentioned above, before we start running it. After we have ran the tool with such entered parameters, we will ex-tract the logged history and perform some analysis on it, to conclude some notable comments. Then, we show a sample of the logged history and not all of it, as for 10 minutes, we may have hundreds of lines, thus , we focus on a random time window of it. Figure 22.8, represents the sample.

We will extract all the logged history and perform some basic analysis on it. We opted to analyze some of the returned parameters of the PING command. They are Round Trip Time, Server Availability, Server Total Availability, and Time To Live. However, technical users may make other advanced analysis to reach to deeper concludes.

In the above graph, Fig. 22.9, it shows that round trip time varies from a sample to another. In fact, there are notable differences which indicate (more or less) network instability. These gaps are for the Request timed out replies. It means that for some PING signals, the server was not able to respond.

In the above graph, Fig. 22.10, it shows the availability percentages and how it varies according PING replies, it is crystal clear that monitored desired server suffers



```
Pinging 192.168.1.110 with 32 bytes of data:
Reply from 192.168.1.110: bytes=32 time=86ms TTL=64
Reply from 192.168.1.110: bytes=32 time=100ms TTL=64
Request timed out.
Reply from 192.168.1.110: bytes=32 time=42ms TTL=64
Reply from 192.168.1.110: bytes=32 time=51ms TTL=64
Reply from 192.168.1.110: bytes=32 time=74ms TTL=64
Reply from 192.168.1.110: bytes=32 time=98ms TTL=64
Reply from 192.168.1.110: bytes=32 time=40ms TTL=64
Reply from 192.168.1.110: bytes=32 time=45ms TTL=64
Request timed out.
Reply from 192.168.1.110: bytes=32 time=69ms TTL=64
Reply from 192.168.1.110: bytes=32 time=92ms TTL=64
Reply from 192.168.1.110: bytes=32 time=100ms TTL=64
Reply from 192.168.1.110: bytes=32 time=33ms TTL=64
Reply from 192.168.1.110: bytes=32 time=58ms TTL=64
Request timed out.
Request timed out.
Reply from 192.168.1.110: bytes=32 time=82ms TTL=64
Reply from 192.168.1.110: bytes=32 time=204ms TTL=64
Reply from 192.168.1.110: bytes=32 time=60ms TTL=64
Reply from 192.168.1.110: bytes=32 time=66ms TTL=64
Reply from 192.168.1.110: bytes=32 time=60ms TTL=64
Reply from 192.168.1.110: bytes=32 time=70ms TTL=64
Reply from 192.168.1.110: bytes=32 time=40ms TTL=64
Request timed out.
Request timed out.
Request timed out.
Request timed out.
Reply from 192.168.1.110: bytes=32 time=45ms TTL=64
Reply from 192.168.1.110: bytes=32 time=50ms TTL=64
Reply from 192.168.1.110: bytes=32 time=90ms TTL=64
```

Fig. 22.8 A sample of the 10 min logged history

some problems, as some PING packets are dropped, then it might be over load-ed or suffers connection issues like high IP-Band-Width utilization. Therefore, we have average availability of 81 % which falls in critical range. Thus, a course of corrective and adaptive actions need to be addressed.

In the above graph, Fig. 22.11, it shows the total availability percentages and how it varies according PING replies. In fact, total availability always represents an increasing trend which stemming from its nature (accumulative availability, will be

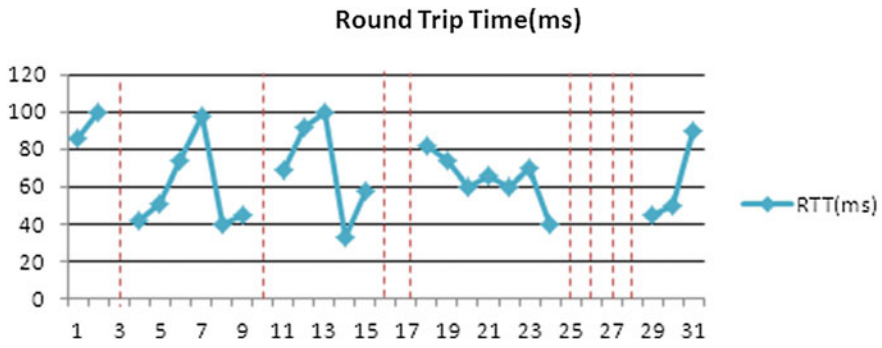


Fig. 22.9 Round trip time graph

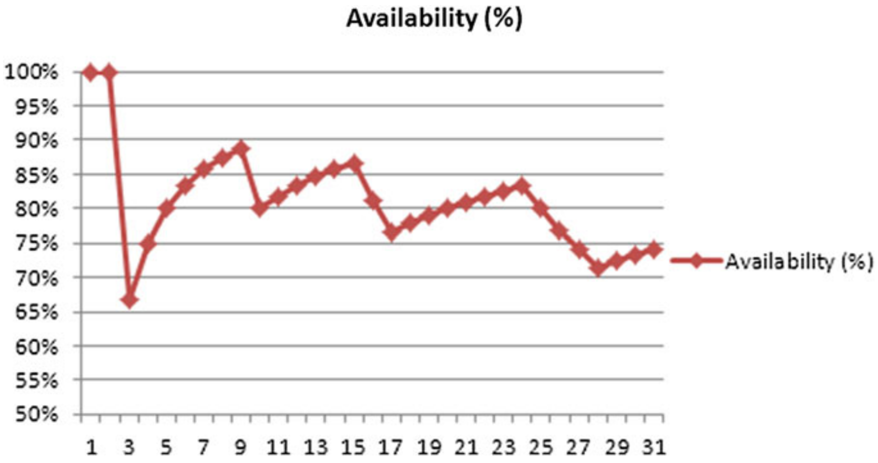


Fig. 22.10 Server availability graph

discussed in the next section). Total availability will be more and more meaningful when we add pervious availability to be appended to the one we are monitoring.

In the above graph, Fig. 22.12, it shows the time to live counts, which means that PING packets still can live 64 hop counts, a hop count means how many nodes a packet can pass through. However, it show a constant straight line because both monitored server and monitoring PC are in the same LAN, but if we going to monitor a WAN server, then networks dynamics will lead to variable TTL.

In real world, there are more and more calculations should be performed (i.e. variance, deviations, standard deviations, upper and lower control limits, upper and lower specification limits, means (averages), medians, and many others).

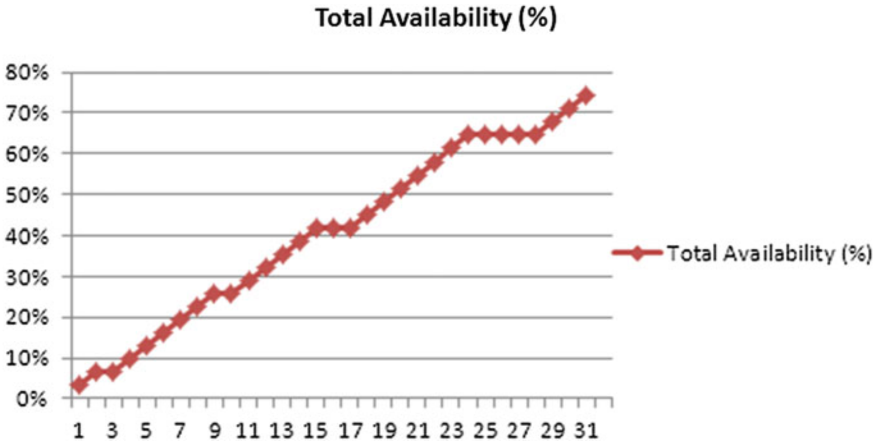


Fig. 22.11 Server total availability graph

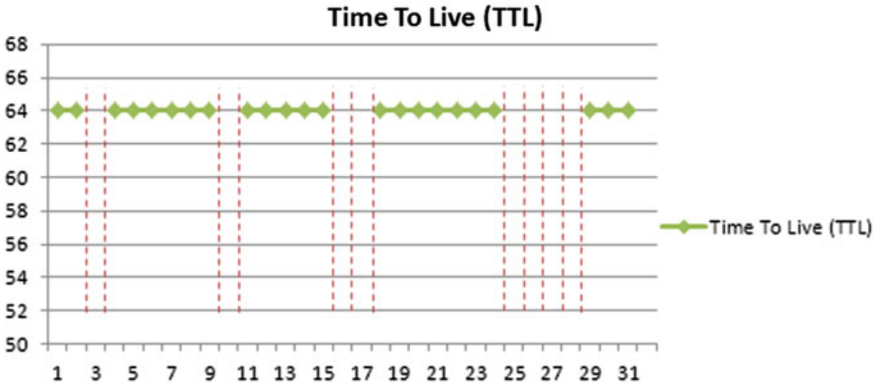


Fig. 22.12 Time to live graph

22.5.3 Tool Accumulative Function

In fact, this tool was not developed only for measuring the current availability, but it also can measure the accumulative availability (total availability). However this accumulative function comes very useful and necessary, to guarantee that at the end of the contract or at SLA termination, user will have the two side of the availability. One is for the period of time the program ran for, and the other for any additional previous availability.

For more clarifications, let us assume that user has SLA contract period of one year (12 months), user was knowing that by the end of September user had 90 % availability (reported by the Cloud Provider), then by the beginning of October user will take the lead and back in-source monitoring of the availability as one of SLO/QoS metrics.

Then, when user starts to run the developed tool, he needs to enter the previous availability = 90 % and the previous availability time = (12 − 3)/12 = 9/12 = 75 %, where the 90 % availability was distributed over 75 % of the one year contract. On the other hand, user also needs to enter run for period = 92 days! where the rest of contract period will be October (31 days) + November (30 days) + December (31 days).

After that, user will press on START button to start monitoring the availability over the remaining interval of the one year contract (last 3 months). During the monitoring, user will have the current availability percentage plus the current total availability which contains both current and previous availabilities.

At the end of the run time (end of the year), the program will stop automatically and will keep all logged history for farther analysis.

If user wants to measure the availability only for certain period of time without appending and previous experiences, then user has to enter previous availability time = 0 and no matters what value user will enter in previous availability. Thus, at the end of the program, user will have the current availability and the total accumulative availability distributed over the needed run period.

22.6 Conclusions and Future Work

As we have seen throughout of this book chapter, contracting with Cloud Provider looks easy, but it is not. It is all about the integration between Cloud Customer and Cloud Provider to achieve the agreed SLAs and meet its objectives. However, monitoring QoS and its related metrics is very much necessary for both Cloud operation sides.

Following the proposed advices, recommendations and guidelines of Cloud layers under Cloud Customer supervision, will resolve the conflict of interests with the Cloud Provider. There should be a sustained effort to allow deep cooperation and collaboration against all Cloud layers. Once held, both parties can guarantee the first Cloud Computing monitoring operational side. Moreover, this work can be intergrated with any related framework. Simply, it can be combined with our proposed IT/Legal framework which has been published before. This integration, enrich the understanding that business need to make its decision towards Cloud Computing.

However, for the seconds Cloud Computing monitoring operational side, we have our tool can be developed to tackle more aspect of the availability being monitored. Also, adding more metrics to be monitored like bandwidth utilization, adding more functions to analyze deeply the PING returned lines, adding an option to monitor more target devices in the same time and in the same program window, allow users to enter more parameters for more precise measurements, add function to draw the metric measured on a 2-D graph, add functions to let the tool send periodic and exceptional reports automatically, add function to let the tool export the deep details to excel files, add functions to provide more calculations options, enhance the GUI to run in the background then pop-up messages in warning cases, and many others which allow cloud computing customer to rely on its own findings to validate and verify the measurements reports provided by the cloud service provider.

References

1. Samaan, N., Karmouch, A.: Towards autonomic network management: an analysis of current and future research directions. Commun. Surv. Tutor. IEEE **11**(3), 22–36 (2009)
2. Chowdhury, K., Boutaba, R.: Network virtualization: state of the art and research challenges. Commun. Mag. IEEE **47**(7), 20–26 (2009)
3. Hyojoon, K., Nick, F.: Improving network management with software defined networking. Commun. Mag. IEEE **51**(2), 114–119 (2013)
4. Timothy, H., Natasha, G., Martin, C., John, M., Scott, S.: Practical declarative network management. In: WREN '09 Proceedings of the 1st ACM Workshop on Research on Enterprise Networking, pp. 1–10 (2009)
5. Jeonghwa, Y., David, H., Keith, W.: Supporting home network management through interactive visual tools. In: UIST '10 Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, pp. 109–118 (2010)
6. Jeonghwa, Y., Keith, W.: A study on network management tools of householders. In: HomeNets '10 Proceedings of the 2010 ACM SIGCOMM Workshop on Home, Networks, pp. 1–6 (2010)
7. Danny, R., Rolf, S., Constantine, E., Mads, D.: In-Network monitoring, In: Algorithms for Next Generation Networks, Series Computer Communications and Networks Springer, pp: 287–317 (2010)
8. Jorge, V., Antonio, G., Vctor, V., Julio, B.: Ontology-based network management: study cases and lessons learned. J. Netw. Syst. Manag. (Springer) **17**(3), 234–254 (2009)
9. Ralf, W., Keith, W., Motivation: the dawn of the age of network-embedded applications. In: Network-Embedded Management and Applications, pp. 3–21, Springer (2013)
10. Johannes, W.: Secret-Sharing Hardware Improves the Privacy of Network Monitoring. Data Privacy Management and Autonomous Spontaneous Security, Series Lecture Notes in Computer Science Springer, vol. 6514, pp. 51–63 (2011)
11. Changzhong, W, Shukun, C., Yu, M.: One kind of remote monitoring network design based on open CNC system. In: Proceedings of the 2012 International Conference on Cybernetics and Informatics Springer, vol. 163, pp. 2007–2013 (2013)
12. Jiantao, G., Yan, W., Zhao, G.: Efficient network monitoring system. In: Information Computing and Applications, pp. 34–40. Springer (2012)
13. Francesco, F., Luca, D.: High speed network traffic analysis with commodity multi-core systems. In: IMC '10 Proceedings of the 10th ACM SIGCOMM Conference on Internet, Measurement, pp. 218–224 (2010)
14. David, R., Fabin, E., Zihui, G.: Crowdsourcing service-level network event monitoring, ACM SIGCOMM Computer Communication Review - SIGCOMM '10, 40(4), pp: 387–398, (2010)
15. Wuhib, F., Dam, M., Stadler, R., Clem, A.: Robust monitoring of network-wide aggregates through gossiping. Netw. Serv. Manag. IEEE Trans. **6**(2), 95–109 (2009)
16. Rad, M., Fouli, K., Fathallah, A., Rusch, A., Maier, M.: Passive optical network monitoring: challenges and requirements. Commun. Mag. IEEE **49**(2), 45–52 (2011)
17. Xi, C., Zheng, X., Hyungjun, K., Gratz, P., Jiang, H., Kishinevsky, M., Ogras, U.: In-network Monitoring and Control Policy for DVFS of CMP Networks-on-Chip and Last Level Caches, Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium, pp. 43–50 (2012)
18. Cloud Security Alliance, http://cloudsecurityalliance.org/, (2013)
19. Dillon, T., Chen, W., Chang, E., Cloud Computing: Issues and Challenges, Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference, pp. 27–33 (2010)
20. Salvatore, V., Rocco, A., Beniamino, D., Massimilano, R., Dana, P.: A Cloud agency for SLA negotiation and management. In: Euro-Par 2010 Parallel Processing Workshops, Series Lecture Notes in Computer Science Springer, vol. 6586, pp. 587–594 (2011)
21. Michael, A., Armando, F., Rean, G., Anthony, D., Randy, K., Andy, K., Gunho, L., David, P., Ariel, R., Ion, S., Matei, Z.: A view of cloud computing. Mag. Commun. ACM **53**(4), 50–58 (2010)

22. Yi, W., Blake, M.: Service-oriented computing and cloud computing: challenges and opportunities. Internet Comput. IEEE **14**(6), 72–75 (2010)
23. Mladen, A., Eric, S., Patrick, D.: Integration of high-performance computing in to cloud computing services. In: Handbook of Cloud Computing, pp. 255–276. Springer (2010)