

Studies in Computational Intelligence 555

Azah Kamilah Muda  
Yun-Huoy Choo  
Ajith Abraham  
Sargur N. Srihari *Editors*

# Computational Intelligence in Digital Forensics: Forensic Investigation and Applications

 Springer

# **Studies in Computational Intelligence**

Volume 555

*Series editor*

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

For further volumes:

<http://www.springer.com/series/7092>

### *About this Series*

The series “Studies in Computational Intelligence” (SCI) publishes new developments and advances in the various areas of computational intelligence—quickly and with a high quality. The intent is to cover the theory, applications, and design methods of computational intelligence, as embedded in the fields of engineering, computer science, physics and life sciences, as well as the methodologies behind them. The series contains monographs, lecture notes and edited volumes in computational intelligence spanning the areas of neural networks, connectionist systems, genetic algorithms, evolutionary computation, artificial intelligence, cellular automata, self-organizing systems, soft computing, fuzzy systems, and hybrid intelligent systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

Azah Kamilah Muda · Yun-Huoy Choo  
Ajith Abraham · Sargur N. Srihari  
Editors

# Computational Intelligence in Digital Forensics: Forensic Investigation and Applications

 Springer



*Editors*

Azah Kamilah Muda  
Department of Software Engineering  
Faculty of Information and Communication  
Technology  
Universiti Teknikal Malaysia Melaka  
(UTeM)  
Durian Tunggal  
Malaysia

Yun-Huoy Choo  
Department of Software Engineering  
Faculty of Information and Communication  
Technology  
Universiti Teknikal Malaysia Melaka  
(UTeM)  
Durian Tunggal  
Malaysia

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

Sargur N. Srihari  
Department of Computer Science and  
Engineering  
Center of Excellence for Document  
The State University of New York SUNY  
Buffalo, New York  
USA

ISSN 1860-949X

ISBN 978-3-319-05884-9

DOI 10.1007/978-3-319-05885-6

Springer Cham Heidelberg New York Dordrecht London

ISSN 1860-9503 (electronic)

ISBN 978-3-319-05885-6 (eBook)

Library of Congress Control Number: 2014935371

© Springer International Publishing Switzerland 2014

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

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

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

Computational intelligence techniques, which are methods that automate expert human procedures, have been widely explored in various domains including forensics. Analysis in forensics encompasses the study of pattern analysis that answer the question of interest in security, medical, legal, genetic studies etc. However, forensic analysis is usually performed through experiments in lab, which is expensive both in cost and time. Therefore, we seek to explore the progress and advancement of computational intelligence techniques in different areas of forensic studies. This aims to build a stronger connection between computer scientists and forensic domain experts. This edited Volume titled *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*, is the first volume on Digital Forensics in this Book series. The book presents original research results and innovative applications of computational intelligence in digital forensics. This volume contains eighteen chapters and presents the latest state-of-the-art advancement of Computational Intelligence in Digital Forensics; in both theoretical and application papers related to novel discovery in intelligent forensics. The chapters are further organized into three Sections: Introduction, Forensic Discovery and Investigation, and Intelligent Forensic Science Applications.

Section 1 consists of an Introductory chapter by Pratama et al., which introduces the basic principles of forensic science and its prominence, the emergence of digital forensics, the incorporation of computational intelligence towards digital forensics, and finally discusses the leading societies, journals, and conferences related to computational intelligence in digital forensics.

Section 2 comprise of five chapters and focuses on the studies, methodologies, and techniques employed to enhance the existing forensic discovery and investigation. Chapter 2 titled *Digital Forensics 2.0: A Review on Social Networks Forensics* by Keyvanpour et al. introduces the usefulness of social network forensics techniques incorporation for analyzing and surveying social inter actions to detect, predict and prevent all forms of potential criminal activities. In Chapter 3 titled *Impact of Some Biometric Modalities on Forensic Science*, by Awad and Hassanien, present a study of the impacts of using some biometric modalities in forensic applications, and sheds light on the positive and the negative impacts of using some biometric modalities in forensic science.

Chapter 4 titled *Incorporating Language Identification in Digital Forensics Investigation Framework*, by Akosu and Selamat, discusses the incorporation of language identification in digital forensics investigation (DFI) models in order to help law enforcement to be a step ahead of criminals, outlines issues of language identification in DFI frameworks, and proposes a new framework with language identification component. Mitra and Kundu in Chapter 5 titled *Cost Optimized Random Sampling in Cellular Automata for Digital Forensic Investigations* report about an efficient design methodology to facilitate random sampling procedure to be used in digital forensic investigations. Chapter 6 titled *Building Multi-modal Crime Profiles with Growing Self Organising Maps* by Boo and Alahakoon propose the fusion of multiple sources of crime data to populate a holistic crime profile through the use of Growing Self Organising Maps.

Section 3, begins with the Chapter titled *Anthropometric Measurement of North-East Indian Faces for Forensic Face Analysis* by Saha et al. and discusses a study of the facial structural differences between the various tribes and non-tribes of the northeastern region of India. Bera et al. in Chapter 8 titled *Hand Biometrics in Digital Forensics* investigate the potential benefits and scope of hand-based modes in forensics with an illustration of hand geometry verification method. In Chapter 9 *A Review on Age Determination Techniques for Non-Human in Forensic Anthropology*, Sahadun et al. present a review of techniques in identifying age for non-human cases regardless the specimen of data, which focuses on age determination. Mata et al. in Chapter 10 titled *Integrating Computational Methods for Forensic Identification of Drugs by TLC, GC and UV Techniques* combine the computational methods and the techniques of Thin Layer Chromatography (TLC), Ultraviolet (UV) and Gas Chromatography (GC), which brings significant improvements in the speed and accuracy of the analysis and identification of drugs of abuse. In Chapter 11 titled *Detecting Counterfeit RFID Tags Using Digital Forensic*, by Khor et al., present the electronic fingerprint matching method in the digital forensic investigation model. Medeiros et al. in Chapter 12 titled *Learning Remote Computer Fingerprinting* present some advances and surveys the use of computational intelligence for remote identification of computers and its applications to network forensics. In the sequel Pal et al. focus on *Signature-based Biometric Authentication*. Bagchi et al. in Chapter 14 titled *Registration of Three-dimensional Human Face Images across Pose and their Applications in Digital Forensic* analyze the registration methods for face recognition across different poses from 0 to 90°. In Chapter 15 titled *Computational Methods for the Analysis of Footwear Impression Evidence*, Srihari and Tang, propose new algorithms to improve image quality, computing features for comparison, measuring the degree of similarity, and retrieval of closest prints from a database, which determines the degree of uncertainty in identification. Pratama et al. in Chapter 16 titled *A New Swarm-based Framework for Handwritten Authorship Identification in Forensic Document Analysis* focus on identifying the unique individual significant features of word shape by using feature selection method prior the identification task. In Chapter 17 titled *Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning*, Tallón-Ballesteros and Riquelme perform an

experimental study on a forensics data task for multi-class classification including several types of methods such as decision trees, Bayes classifiers, expert systems, neural networks and based on nearest neighbors. In the last Chapter, *Speech Quality Enhancement in Digital Forensic Voice Analysis*, Ekpenyong and Obot improves the integrity, vis-à-vis the intelligibility of speech signals.

Editors

Azah Kamilah Muda  
Yun-Huoy Choo  
Ajith Abraham  
Sargur N. Srihari

# Reviewers

Alberto Ochoa O. Zezzatti	Universidad Autónoma de Ciudad Juárez, Mexico
Aliismail Awad	Al Azhar University, Qena, Egypt
Anirban Kundu	Kuang-Chi Institute of Advanced Technology, China
Antonio Silva	Universidad Venezuela Central, University City, Venezuela
Ajith Abraham	Machine Intelligence Research Labs, USA
Antonioj Tallón-Ballesteros	University of Seville, Spain
Arnab Mitra	Adamas Institute of Technology, India
Asish Bera	Haldia Institute of Technology, India
Ayan Seal	Jadavpur University, India
Azah Kamilah Muda	UniversitiTeknikal Malaysia Melaka, Malaysia
Choo Yun-Huoy	UniversitiTeknikal Malaysia Melaka, Malaysia
Dania Porro	Advanced Technologies Application Center (CENATAV), Cuba
Dewi Nasien	Universiti Teknologi Malaysia, Malaysia
Eduardo Garea	Advanced Technologies Application Center (CENATAV), Cuba
Francisco José Silva Mata	Advanced Technologies Application Center (CENATAV), Havana, Cuba
Francisco Ornelas	Universidad Autonoma de Aguascalientes, Mexico
Habibollah Harun	Universiti Teknologi Malaysia, Malaysia
Isneri Talavera-Bustamante	Advanced Technologies Application Center (CENATAV), Havana, Cuba
Joao Batista Borges Neto	Federal University of Rio Grande do Norte, Brazil
Julio de Carvalho Ponce	Universidad Autonoma de Aguascalientes, Mexico
José Alberto Hernández Aguilar	Universidad Autónoma del Estado de Morelos, Mexico
Lázaro Bustio Martínez	Advanced Technologies Application Center (CENATAV), Havana, Cuba
Moses Ekpenyong	University of Uyo, Nigeria

Mrinal Kanti Bhowmik	Tripura University, India
Marjan Kuchaki Rafsanjani	Shahid Bahonar University of Kerman, Iran
João Paulo de Souza Medeiros	Federal University of Rio Grande do Norte, Brazil
Nicholas Akosu	Universiti Teknologi Malaysia, Malaysia
Mohammad Ghulam Rahman	Universiti Sains Malaysia, Malaysia
Okure Obot	University of Uyo, Nigeria
Paulo S. Motta Pires	Federal University of Rio Grande do Norte, Brazil
Parama Bagchi	MCKV Institute of Engineering, India
Sargur Srihari	University at Buffalo, USA
Srikanta Pal	Griffith University, Australia
Yeeling Boo	Deakin University, Australia
Yenisel Plasencia Calana	Delft University of Technology, The Netherlands
Yoanna Martínez-Díaz	Advanced Technologies Application Center (CENATAV), Havana, Cuba

# Contents

## Section I: Introduction

<b>Computational Intelligence in Digital Forensics</b> .....	1
<i>Satrya Fajri Pratama, Lustiana Pratiwi, Ajith Abraham, Azah Kamilah Muda</i>	

## Section II: Forensic Discovery and Investigation

<b>Digital Forensics 2.0: A Review on Social Networks Forensics</b> .....	17
<i>MohammadReza Keyvanpour, Mohammad Moradi, Faranak Hasanzadeh</i>	

<b>Impact of Some Biometric Modalities on Forensic Science</b> .....	47
<i>Ali Ismail Awad, Aboul Ella Hassanien</i>	

<b>Incorporating Language Identification in Digital Forensics Investigation Framework</b> .....	63
<i>Nicholas Akosu, Ali Selamat</i>	

<b>Cost Optimized Random Sampling in Cellular Automata for Digital Forensic Investigations</b> .....	79
<i>Arnab Mitra, Anirban Kundu</i>	

<b>Building Multi-modal Crime Profiles with Growing Self Organising Maps</b> .....	97
<i>Yee Ling Boo, Dammina Alahakoon</i>	

## Section III: Intelligent Forensic Science Applications

<b>Anthropometric Measurement of North-East Indian Faces for Forensic Face Analysis</b> .....	125
<i>Kankan Saha, Mrinal Kanti Bhowmik, Debotosh Bhattacharjee</i>	

<b>Hand Biometrics in Digital Forensics</b> .....	145
<i>Asish Bera, Debotosh Bhattacharjee, Mita Nasipuri</i>	

<b>A Review on Age Identification Techniques for Non-human in Forensic Anthropology</b> .....	165
<i>Nur A. Sahadun, Mohammed R.A. Kadir, Habibollah Haron</i>	
<b>Integrating Computational Methods for Forensic Identification of Drugs by TLC, GC and UV Techniques</b> .....	187
<i>Francisco José Silva Mata, Dania Porro Muñoz, Diana Porro Muñoz, Noslen Hernández, Isneri Talavera Bustamante, Yoanna Martínez-Díaz, Lázaro Bustio Martínez</i>	
<b>Detecting Counterfeit RFID Tags Using Digital Forensic</b> .....	211
<i>JingHuey Khor, Widad Ismail, Mohammad Ghulam Rahman</i>	
<b>Learning Remote Computer Fingerprinting</b> .....	253
<i>João P. Souza Medeiros, João B. Borges Neto, Agostinho M. Brito Júnior, Paulo S. Motta Pires</i>	
<b>Signature-Based Biometric Authentication</b> .....	285
<i>Srikanta Pal, Umapada Pal, Michael Blumenstein</i>	
<b>Registration of Three Dimensional Human Face Images across Pose and Their Applications in Digital Forensic</b> .....	315
<i>Parama Bagchi, Debotosh Bhattacharjee, Mita Nasipuri, Dipak Kumar Basu</i>	
<b>Computational Methods for the Analysis of Footwear Impression Evidence</b> .....	333
<i>Sargur N. Srihari, Yi Tang</i>	
<b>A New Swarm-Based Framework for Handwritten Authorship Identification in Forensic Document Analysis</b> .....	385
<i>Satrya Fajri Pratama, Azah Kamilah Muda, Yun-Huoy Choo, Noor Azilah Muda</i>	
<b>Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning</b> .....	413
<i>Antonio J. Tallón-Ballesteros, José C. Riquelme</i>	
<b>Speech Quality Enhancement in Digital Forensic Voice Analysis</b> .....	429
<i>Moses Ekpenyong, Okure Obot</i>	
<b>Erratum</b>	
<b>Impact of Some Biometric Modalities on Forensic Science</b> .....	E1
<i>Ali Ismail Awad, Aboul Ella Hassanien</i>	
<b>Author Index</b> .....	453



# Computational Intelligence in Digital Forensics

Satrya Fajri Pratama<sup>1</sup>, Lustiana Pratiwi<sup>1</sup>, Ajith Abraham<sup>1,2</sup>, and Azah Kamilah Muda<sup>1</sup>

<sup>1</sup> Computational Intelligence and Technologies (CIT) Research Group,  
Center of Advanced Computing and Technologies,  
Faculty of Information and Communication Technology,  
Universiti Teknikal Malaysia Melaka  
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia  
satrya@student.utem.edu.my, lustiana@gmail.com,  
ajith.abraham@ieee.org, azah@utem.edu.my

<sup>2</sup> Machine Intelligence Research Labs (MIR Labs)  
Scientific Network for Innovation and Research Excellence, Auburn, WA, USA

**Abstract.** Forensic Science has been around for quite some time. Although various forensic methods have been proved for their reliability and credibility in the criminal justice system, their main problem lies in the necessity of highly qualified forensic investigators. In the course of analysis of evidences, forensic investigators must be thorough and rigorous, hence time consuming. Digital Forensic techniques have been introduced to aid the forensic investigators to acquire as reliable and credible results as manual labor to be presented in the criminal court system. In order to perform the forensic investigation using Digital Forensic techniques accurately and efficiently, computational intelligence oftentimes employed in the implementation of Digital Forensic techniques, which has been proven to reduce the time consumption, while maintaining the reliability and credibility of the result, moreover in some cases, it is producing the results with higher accuracy. The introduction of computational intelligence in Digital Forensic has attracted a vast amount of researchers to work in, and leads to emergence of numerous new forensic investigation domains.

**Keywords:** computational intelligence, digital forensics, forensic science, computational forensics.

## 1 Introduction to Forensic Science

Since the beginning of the nineteenth century, the use of science during the observation and interpretation of evidences has been the major key in solving a wide array of criminal cases. It allows the evidence to be provided to legal investigations, and justifies the validity of conclusions drawn by the forensic investigation authorities. Henceforth, the most logical attempts to be made are to organize the areas for processing the evidence, and known as Forensic Science [1]. Thus, it is required for the investigating authorities to obtain the scientific information from knowledgeable scientists or technical instruments of academic institution, most prominently from chemistry or pharmacology departments [2].

Before the introduction of Forensic Science, the accuracy of the investigation depends on the familiarity and awareness of the investigating authorities towards the felons and their connection and involvements based on their signature in a particular case [1]. Originally, the forensic investigation relies heavily on the anthropologic measurements supported by photographic documentation, which is then substituted by fingerprint identification, due to its significantly more accurate and reliable method. The introduction of fingerprint necessitates greater responsibilities for handling physical evidence, and subsequently opens a gate for new identification methods, such as the identification of human byproducts, soil, and other materials found at the crime scenes.

Sherlock Holmes, an extremely popular consulting detective from novel series made by Sir Arthur Conan Doyle, is the best example of excellent forensic investigator. He is capable of observing the minute details of objects, crime scenes, evidences, or people which is oftentimes neglected by ordinary people, and draws accurate inferences from his observations by using deductive reasoning. He is also capable to differentiate, which events and evidences are actually related to the crime or which are mere coincidence. Having organized the evidences and facts, he will use his inference capability to construct the events, motives, and suspects of the crime, and thus solve it. Sir Arthur Conan Doyle has shown the utmost importance of observation and analysis of evidences, and therefore set it as the primary qualities of every forensic investigator must possess.

Forensic Science started to grow steadily in the end of nineteenth century. Some fields are still in its early stage of inception and thus require special equipment operated by competent and proficient scientists, such as DNA testing and drugs identification. On the other hand, some fields can be conducted by less qualified forensic investigator, such as fingerprint detection and identification [3]. Although initially Forensic Science has never been widely understood by public, it has recently attracted the general attention, largely due to the popularity of *Crime Scene Investigation (CSI)* television show and its successors.

While the show is technically and scientifically flawed and oftentimes inaccurately depicts the investigation process, they serve to convey the prominence of Forensic Science for the public consumption by concealing the complexity of investigation process and replacing it with simpler and faster imaginary process. Several fields have gained more public awareness due to the popularity of CSI, most notably is DNA testing. Although DNA testing has been recognized earlier due to its successes, its reputation is mostly limited to medical and legal community. DNA testing is powerful due to its capability to acquit the wrongly convicted and correctly identify previously unidentified suspects that is still roaming freely since mid-1980s [2].

Numerous people contributed to the advancement of Forensic Science field, most notably the people that made earliest contributions and thus formulated the various disciplines that currently serve as the foundations of Forensic Science. The earliest documented contribution to the Forensic Science is in 1814, where Mathieu Orfila published the first scientific treatise with title "*Traité des Poisons*" or

“*Toxicologie Générale*” on the detection of poisons and their effects on animals, allowing the endorsement of forensic toxicology as a valid scientific discipline, and therefore named him as the father of forensic toxicology [3].

Subsequently, Alphonse Bertillon devised the first scientific system of personal identification by developing the science of anthropometry as a systematic procedure of taking a series of body measurements as a means of distinguishing one individual from another in 1879. Adjacent to Bertillon’s personal identification, Francis Galton undertook the first definitive study of fingerprints, developed a methodology of classifying them for filing, and published a book titled “Finger Prints” in 1892, which contained the first statistical proof supporting the uniqueness of his method of personal identification. This work serves as the foundation of modern fingerprint identification system [2, 4].

Successively, Dr. Karl Landsteiner discovered that blood could be grouped into different categories in 1901. These blood groups or types are now recognized as A, B, AB, and O. The possibility of blood grouping could be a useful characteristic for the identification of an individual as intrigued by Dr. Leones Lattes, a professor at the Institute of forensic medicine at the University of Turin in Italy. In 1915, he devised a relatively simple procedure for determining the blood group of a dried bloodstain, a technique that he immediately applied to criminal investigations [3].

In the meantime, Albert S. Osborn’s development of the fundamental principles of document examination was responsible for the acceptance of documents as scientific evidence by the courts. Osborn authored the first significant text in this field, “*Questioned Documents*” in 1910, which is still considered as a primary reference for document examiners [2, 3].

The recent advancement and maturity of technologies employed in the crime laboratories and other aspects involved in the criminal justice system has put superfluous burden to forensic investigation authorities as well as academic researchers to contribute in critically relevant methods to the integrity of criminal justice system, which may prove to be more challenging in the future. This is due to the sophistication of science and technologies as a double-edged sword, since both local and international criminal individuals and groups also exploit it. Therefore, it is necessary to commit exceptional amount of resources to the advancement science and technology in the field of Forensic Science for the sake of entire human being [3, 5].

Imminent dangers to the national and international security due to the advancement of science and technology are, for instance, the more intricate schemes engaged by local and international criminal entities, the unimpeded flow of money due to sophisticated communication systems, and most distressing of all, the maturation of criminal enterprises to terrorist organizations, and thus provide fecund prospect to gain illegal financial revenues. Nevertheless, forensic investigation authorities equivalently consume the benefits of advanced science and technology and employing it to perform extraordinary measures to unravel the unsolved case files, and often achieve great success. These successes exemplify outstanding motivations to making continuous advancement, keeping pace, and investing in the betterment science and technology in the field of Forensic Science [6].

## 2 The Prominence of Forensic Science

The use of Forensic Science has significantly improved the justice system. Without Forensic Science, it is impossible to prosecute a criminal where eyewitnesses are unavailable. The criminal will be freely roaming and further commit crimes without having to worry of the consequences of their actions. They just simply need to ensure that no one is around to witness during their act of crimes. Fortunately, it is not the case, as the evidence left from criminal acts are collected and analyzed, and science is used as a means to solve the crimes. Forensic Science is used to draw the inferences based on the analysis of evidences which is properly collected and uncontaminated. Therefore, it is necessary for individuals from law agencies and authorities to undertake certain training and education before they can be certified as forensic investigators and perform frequent contacts with the crime scene [7].

The prominence of Forensic Science is growing by day; therefore it is necessary for various agencies and authorities to prevent the use of inadequate methodology during their forensic investigations, whether from the lack of funding, outdated methodologies, inferior scientific standards, and the most dangerous of all, incompetent forensic investigators, along with the flaws and complexities of proper procedures to efficiently collect and analyze the forensic evidence. Although it is not unheard of, these inadequacies still pose a grave danger towards the process of upholding the justice system, where it is possible to cause innocent people convicted or guilty criminal acquitted, or even worse, it might destroy the credibility of the forensic investigation authorities, and possibly cause the past cases solved by said authorities questioned and brought into dispute [6].

Despite the fact of these possibilities are always present regardless, the forensic investigation authorities always remain devoted to hold every aspect of forensic analysis to the utmost standards, as well as ensuring the use of modern processes, the availability of proper resources and equipment, personnel training, and maintaining the integrity benchmarks by periodically performing the accreditation of their forensic laboratories. These fundamental prerequisites must be engaged to prevent the compromise of the criminal justice system; otherwise it might bring the disrepute to the system and its administrators [2]. Every procedure must be performed flawlessly and rigorously thorough the investigation process, in order to ensure the reliability of the results produced can endure the vigorous contests and the conclusion drawn from the results are unquestionably valid. The stake it poses should be more than enough reason for every forensic investigator to perform total and holistic approach in uncompromised fashion and ensure the absolute integrity [5, 6].

Ever since its conception, Forensic Science has contributed to massive amount of successful prosecution by producing significant evidence which is capable to exonerate innocent people and convict guilty criminals [5, 8]. During its early time, the number of flawed Forensic Science cases is continuously increasing. Recent science and technology advancement and knowledge-based environment has allowed higher threshold of accuracy to exist, compared to the previous era. The flaws of previous sciences have provided valuable lessons along the way, and sanction the field to mature substantially. However, ample amount of effort must still be conducted to ensure

the past mistakes are not to be repeated [4]. Moreover, the advances of the discipline possess additional benefits, which are to allow the forensic investigators to identify perpetrators with higher confidence, and to further diminish the occurrence of wrongful convictions, and thus minimize the possibility of true perpetrators to commit crimes while the innocents are held accountable [5].

### 3 CSI Effect and Digital Forensic Techniques

The popularity of CSI and its similar television shows has raised a new phenomenon, termed as *CSI effect*. The negative impact of CSI effect in the context of court trial which is employing jury is that it may affect the process of decision-making. The jurors cannot differentiate the real and fictional forensic investigation methods, since they expected the scientific advancement of Forensic Science shown in these shows are actually existed and implemented in real life. As the result, the unavailability of the fictional evidence or inadequately fictionally processed evidence may result of wrongfully acquittal of the defendants by the jurors. Moreover, it is suggested that CSI effect has taught the criminals various methods to destroy evidences and thus circumvent and obstruct the forensic investigations [9-11].

The rising of the juror expectations originates from the fact of rapid advances in science, particularly in the field of computer science and information technology, and their effects in the popular culture. These advances, however, are highly dramatized, fictionalized, and conveyed via criminal television shows. Since the jurors have been familiar with the technological advances through their frequent use of computers, smartphones, and tablets, they expect the Forensic Science is also as sophisticated as these gadgets [10].

However, this rising expectations also serve as the motivation for researchers to convert the heavily fictionalized Forensic Science procedures into real life implementations in the form of computer systems. Similar to the case of Martin Cooper, whose invents the first handheld mobile phone that is inspired by Star Trek, CSI effect has drove various researchers to create and develop computer systems as tools to perform forensic investigations, also known as Digital Forensic techniques or Computational Forensics. For example, in an episode of CSI, the investigators found the pieces of shredded document in the crime scene. By using sophisticated Digital Forensic technique, they are able to piece it together and thus acquire the original document. While it is scientifically possible, it is cumbersome and tedious process, since it is done manually, and therefore requiring a lot of focus and also time-consuming. Constructing shredded documents is not like piecing the jig-saw puzzle, where we can find the edges and work from there. The number of possibilities is too high to be done manually, and thus a computer system theoretically is capable to construct it. It should be noted that Digital Forensic techniques is different with Digital Forensic Science.

Digital Forensic Science is defined as the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital

sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations [12]. Simply put, Digital Forensic Science is the science to collect, preserve, analyze and present evidence from computers that are sufficiently reliable to stand up in court and convincing [13].

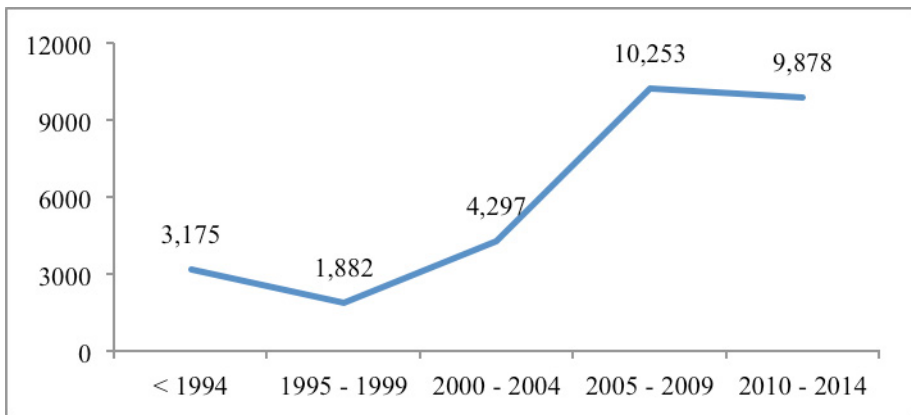
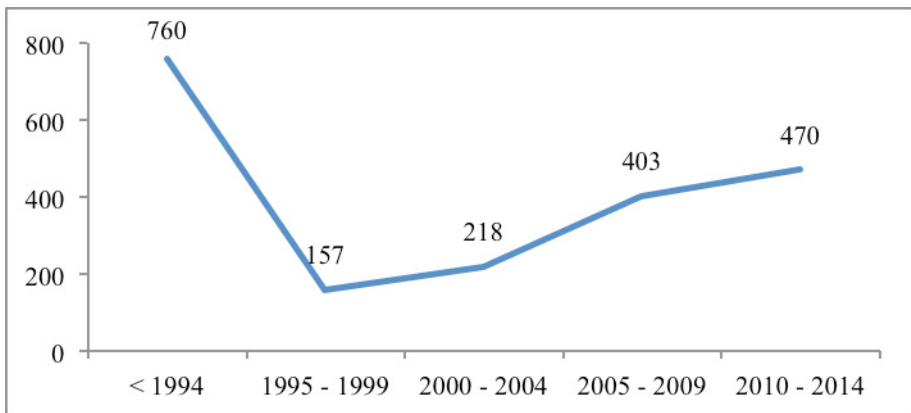
A great deal of researches on Digital Forensic has been conducted by forensic investigation authorities and academic researchers, and thus allows the inception a wide range of new knowledge and methodologies to collect and analyze forensic evidences. It is crucial to collect and analyze properly in order to successfully conduct a prosecution. Digital Forensic is a hybrid of computer science and law, and thus it is necessary to have the knowledge from both disciplines. Oftentimes, the results produced by academic researchers do not comply to the law regulations, which are commonly caused by their lack of knowledge to the relevant laws and their unfamiliarity with the real-world problems and constraints which commonly faced by the forensic investigators. Moreover, improper use of methodologies to analyze the evidence may even suppress the evidence that is properly collected, and thus said evidence is omitted in the court of law. Therefore, academic researchers must consider whether a new methodology they invent can be used practically and legally by the forensic investigators [13]. Therefore, these requirements are the primary reason that the use of Digital Forensic techniques is often disregarded and inadmissible in court.

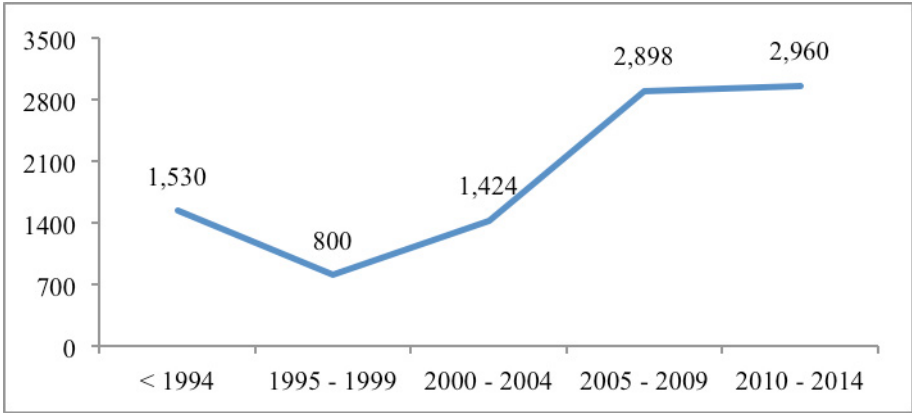
Nevertheless, these restrictions should not be the reason to hinder the technological advancement of Digital Forensic techniques. While its results may inadmissible in court, in certain cases Digital Forensic techniques could produce faster than manual labor, such as fingerprint lookup. Imagine there are  $N$  suspects, and thus forensic investigators must perform  $N$  number of matching to narrow the list of suspects. By using computer systems, the matching can be performed faster, and the results can be verified by the forensic investigators. In essence, there are several Digital Forensic techniques which attract researchers to commit on. These popular techniques, or domains, are fingerprint analysis, bloodstain analysis, questioned documents examination, ballistics examination, shoeprints analysis, surveillance image enhancement, surveillance image noise removal, surveillance image restoration, surveilled object tracking, 3D scene reconstruction, and image integrity analysis.

While several domains are found on the grounds of traditional forensics method, such as fingerprint analysis, bloodstain analysis, questioned documents examination (QDE), ballistics examination, and shoeprints analysis, some domains are new and uniquely surfaced upon the use of modern day technologies, such as surveillance image enhancement, surveillance image noise removal, surveillance image restoration, surveilled object tracking, 3D scene reconstruction, and image integrity analysis, which are obtained from surveillance devices and digital cameras. The rapid growth due to increased interest to these domains in terms of number of publications acquired from IEEE Xplore, ACM Digital Library, and ScienceDirect is shown in Table 1 and depicted in Figures 1-11. The summary of these trends is shown in Figures 12-13.

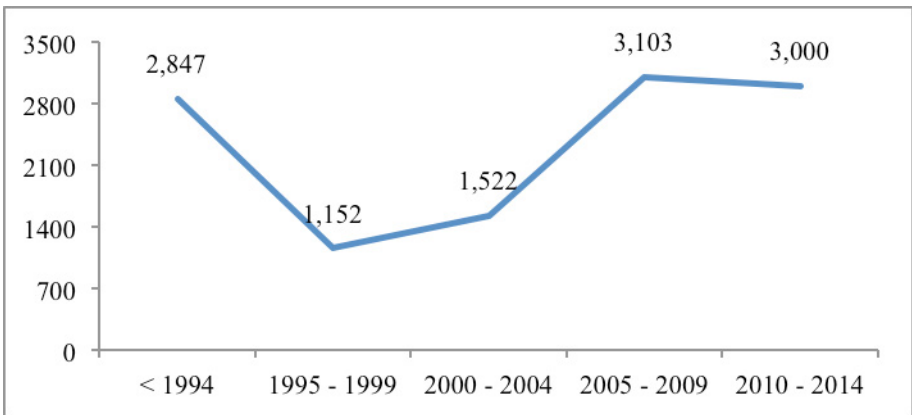
**Table 1.** A quick overview of the Digital Forensic trends

Digital Forensic Domains	< 1994	1995 - 1999	2000 - 2004	2005 - 2009	2010 - 2014
Fingerprint analysis	3,175	1,882	4,297	10,253	9,878
Bloodstain analysis	760	157	218	403	470
QDE	1,530	800	1,424	2,898	2,960
Ballistics examination	2,847	1,152	1,522	3,103	3,000
Shoeprints analysis	33	14	20	63	45
Surveillance image enhancement	736	623	1,383	3,624	4,590
Surveillance image noise removal	655	265	576	1,340	1,784
Surveillance image restoration	337	228	511	1,281	1,515
Surveilled object tracking	1,590	1,556	3,772	12,094	13,811
3D scene reconstruction	895	1,285	2,876	6,374	6,308
Image integrity analysis	13,728	10,608	19,630	40,378	53,174

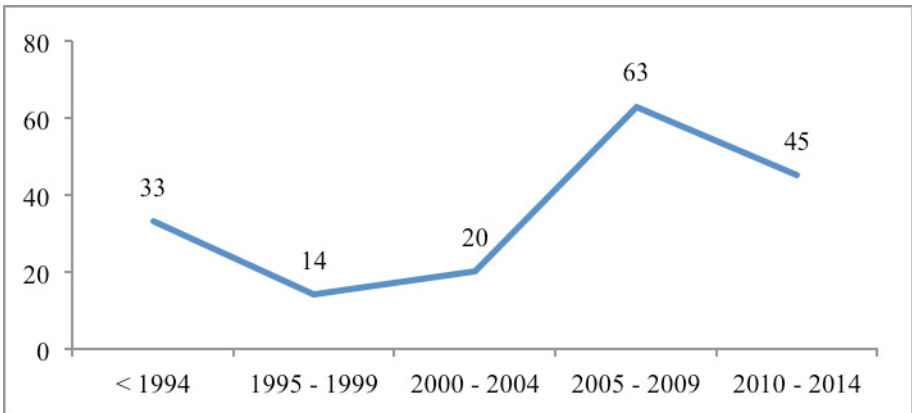
**Fig. 1.** Digital Forensic trends in fingerprint analysis domain**Fig. 2.** Digital Forensic trends in bloodstain analysis domain



**Fig. 3.** Digital Forensic trends in questioned documents examination domain

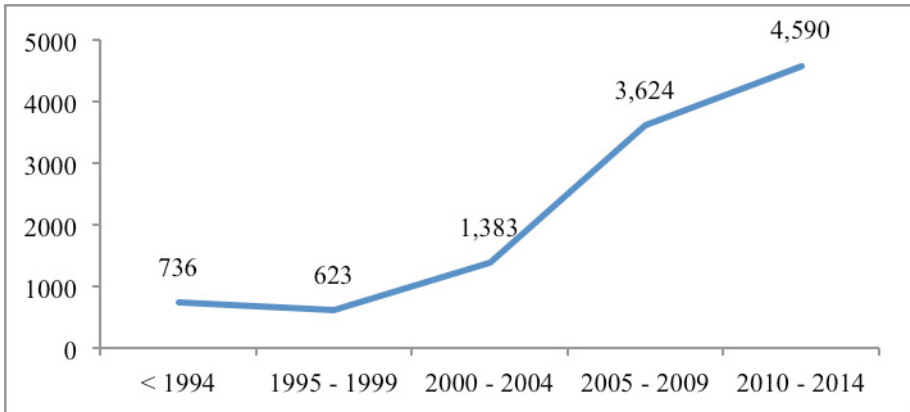


**Fig. 4.** Digital Forensic trends in ballistics examination domain

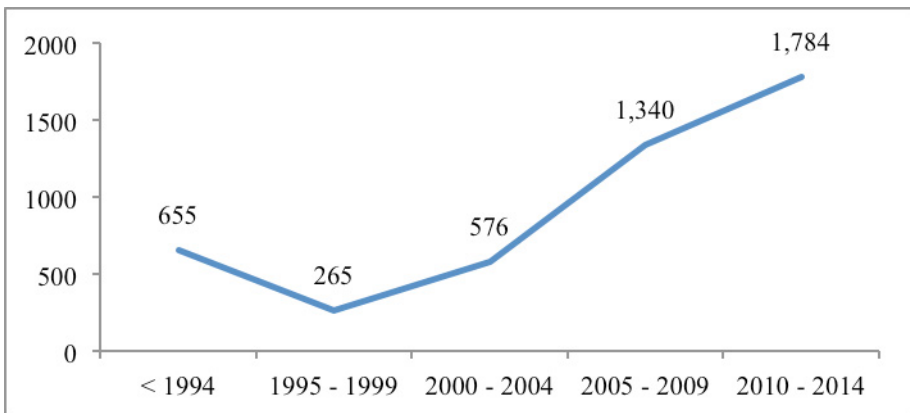


**Fig. 5.** Digital Forensic trends in shoeprints analysis domain

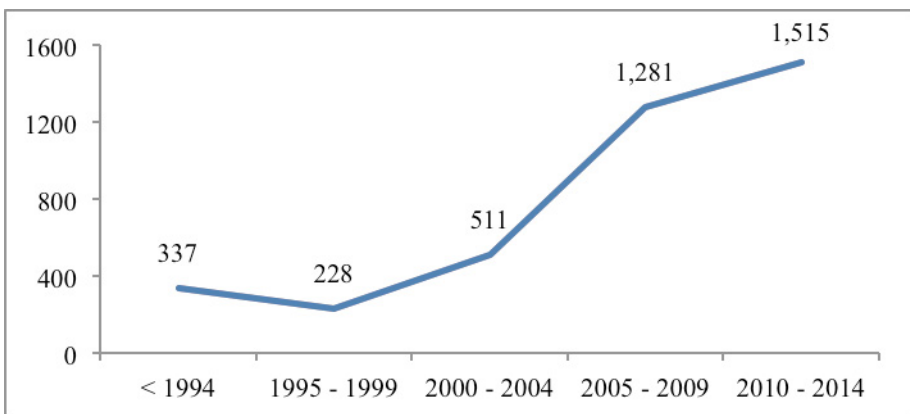




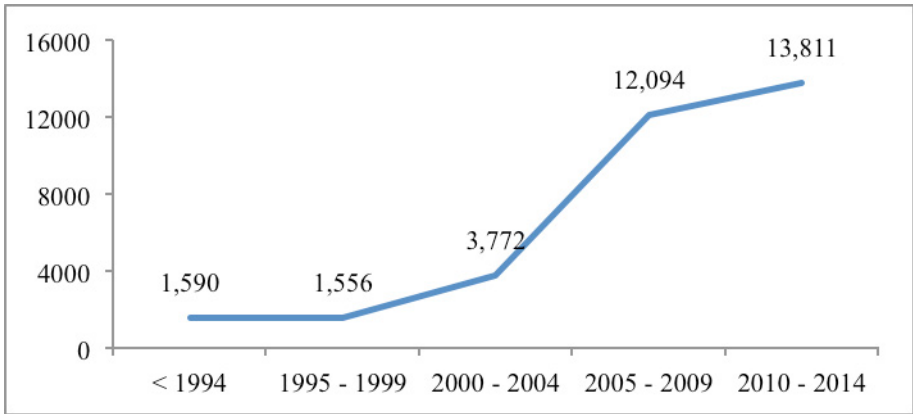
**Fig. 6.** Digital Forensic trends in surveillance image enhancement domain



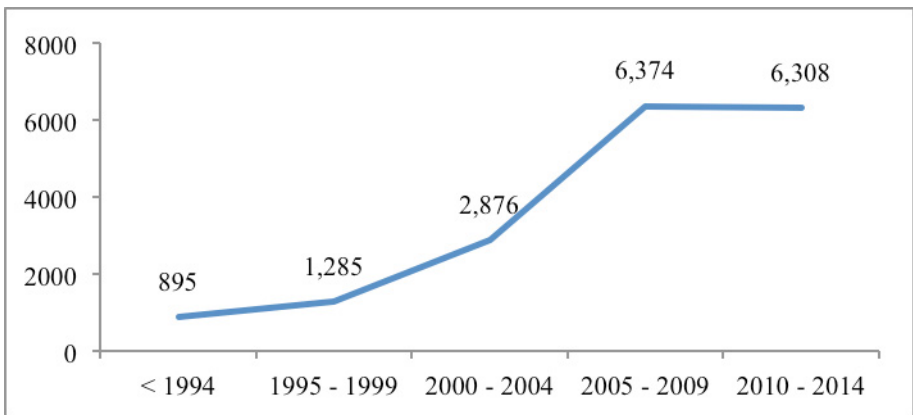
**Fig. 7.** Digital Forensic trends in surveillance image noise removal domain



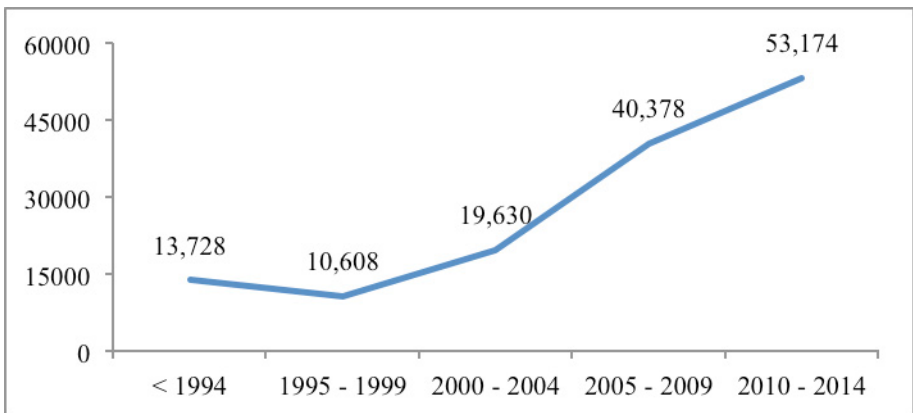
**Fig. 8.** Digital Forensic trends in surveillance image restoration domain



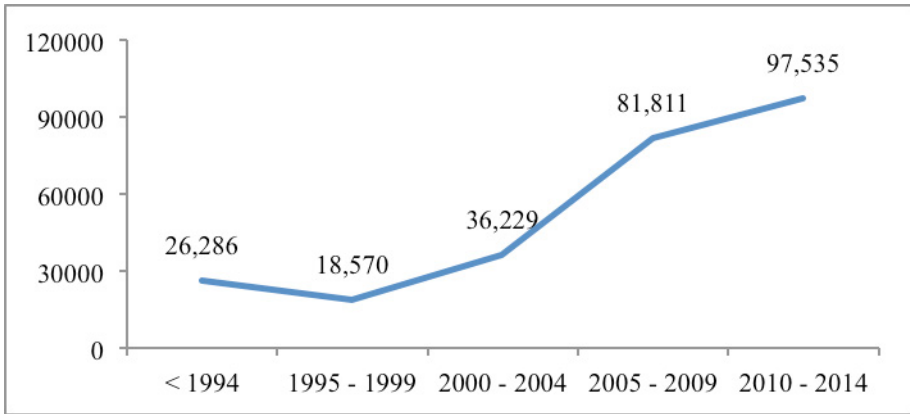
**Fig. 9.** Digital Forensic trends in surveilled object tracking domain



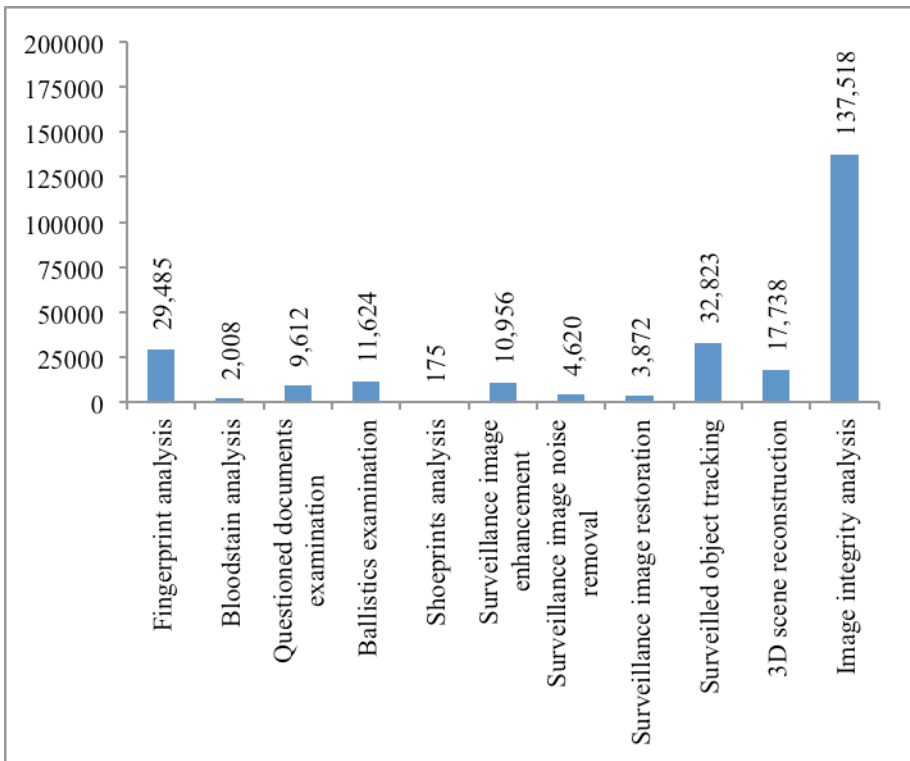
**Fig. 10.** Digital Forensic trends in 3D scene reconstruction domain



**Fig. 11.** Digital Forensic trends in image integrity analysis domain



**Fig. 12.** Digital Forensic trends in all domains



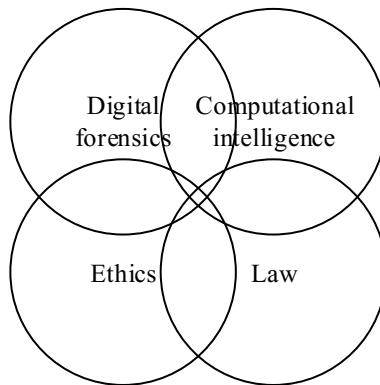
**Fig. 13.** Total number of publications for each domain throughout the years

As summarized in Fig. 12, the number of publications for Digital Forensic domain is continuously growing throughout the year. While it seems the number of research publications pre-1994 is greater than 1995 to 1999 period, it should be noted that this number is the accumulation of publications, which oftentimes dated back to 1970s.

On the other hand, Fig. 13 shows the popularity of Digital Forensic domain. It is shown that image integrity analysis, surveilled object tracking, and fingerprint analysis are the top domains that primarily attract the researchers to work in, while shoe-prints analysis, bloodstain analysis, and surveillance image restoration are the domains where the researchers are struggled in. While these numbers are quite promising, it is difficult to determine their usability in the everyday forensic investigations.

#### 4 Computational Intelligence in Forensic Investigations

Forensic-related technologies are potentially capable to improve the lives of great people. Computational intelligence, a recently growing field of computer science pose a prodigious opportunity to improve Forensic Science. Computational intelligence techniques have been widely used in the domain of computer forensics [14], which has been successfully used in many real world applications on a variety of engineering problems [15]. However, law (legal) and ethical aspects must be taken into consideration when employing computational intelligence in forensic investigations [14, 15]. The relationship between the different fields is illustrated in Fig. 14.



**Fig. 14.** Relationship between various fields [14]

Computational intelligence is based on human intelligence, and therefore it is expected to accomplish the task equal to or even beyond human proficiency [14]. It relies on several key paradigms, such as evolutionary algorithms, neural networks, fuzzy systems, and multi-agent systems [16]. As an example, we take on multi-agent systems. Multi-agent systems is a system composed of many interacting intelligent agents; each one is in itself simple and apparently acts only in its own interest, yet by collaborating and/or competing with each other, it can be used to solve problems which would entirely defeat an individual agent or a monolithic system. Generally, multi-agent systems are flexible and they are easily maintained or modified without the need for drastic rewriting or restructuring, and tends to be robust and recover easily from a breakdown, due to built-in duplication and redundancy of components [16]. The example of multi-agent systems applications in forensic investigations are enormous, as demonstrated in [17-22].

## 5 Leading Societies and Related International Journals and Conferences

The emergence of computational intelligence in Digital Forensic has attracted a vast array of researchers to work on it. These researchers oftentimes form a society and/or working group, where they can collaborate and cooperate. Most notably is the IAPR-TC6, which is the Technical Committee of Computational Forensics (<https://sites.google.com/site/compforgroup/>) under the auspices of International Association for Pattern Recognition (IAPR) and Center of Excellence for Document Analysis and Recognition (CEDAR) of University at Buffalo, State University of New York (<http://www.cedar.buffalo.edu/forensics/>). Meanwhile, many other societies are publicly available on the internet and allow general users to join, such as Computational Forensics Group (<https://groups.google.com/forum/#!forum/compfor>).

Other than these societies, individual researchers have committed researches and publish the results in various journals and conferences. Some of the most celebrated journals related to Digital Forensic are:

- Advances in Digital Forensics:

(<http://www.springer.com/computer/book/978-0-387-30012-2>)

- Forensics in Telecommunications, Information and Multimedia:

(<http://www.springer.com/computer/general+issues/book/978-3-642-02311-8>)

- Digital Forensics and Cyber Crime:

(<http://www.springer.com/computer/general+issues/book/978-3-642-19512-9>)

- CyberForensics:

(<http://www.springer.com/biomed/book/978-1-60761-771-6>)

- Digital Forensics and Cyber Crime:

(<http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-642-11533-2>)

- Digital Forensics and Watermarking:

(<http://www.springer.com/computer/security+and+cryptology/book/978-3-642-40098-8>)

- Information Security and Digital Forensics:

(<http://www.springer.com/computer/security+and+cryptology/book/978-3-642-11529-5>)

- Digital Image Forensics:

(<http://www.springer.com/engineering/signals/book/978-1-4614-0756-0>)

- Forensic Speaker Recognition:  
(<http://www.springer.com/engineering/signals/book/978-1-4614-0262-6>)
- Handbook of Research on Computational Forensics, Digital Crime, and Investigation: Methods and Solutions:  
(<http://www.igi-global.com/book/handbook-research-computational-forensics-digital/449>)

However, the most related to computational intelligence in Digital Forensic is Computational Forensics:

(<http://www.springer.com/computer/image+processing/book/978-3-642-19375-0>).

On the other hand, IAPR has continually organized International Workshop on Computational Forensics (IWCF). More information on the 6<sup>th</sup> IWCF can be found on <http://www.isical.ac.in/~iwcf2014/>.

## 6 Conclusions

Forensic Science has been around for quite some time and played a very important role in the justice system. Without Forensic Science, criminals roam freely without having to fear the consequences of their actions. It is important for forensic investigation authorities to uphold the implementation of Forensic Science to the utmost standards, due to the dangers it pose if it is performed substandard. The advances of Forensic Science are possible due to the flaws of past mistakes and present betterment of justice system. It is essential to continuously advance the discipline itself, since the advancement of science and technology is also abused by the criminal entities. In order to perform proper forensic investigations, forensic investigators are required to undertake adequate training and education.

Forensic Science is indebted to the popularity of CSI television show in gaining its public awareness. Although scientific inaccuracies shown in CSI may pose a great danger to the administration of justice system, it also serve as the motivation for academic researchers from various discipline to realize the fictional technologies presented in the television show, especially from computer science, which lead to the conception of Digital Forensic. However, a great care must be observed when Digital Forensic technique is performed to ensure said Digital Forensic technique complies with the law regulation, and thus allow the evidence collected and analyzed using said technique to be admitted in the court of law.

The introduction of computational intelligence in the Digital Forensic allows the investigation process to be performed in shorter amount of time and with higher reliability and credibility. Computational intelligence is modeled after human intelligence, and it is meant to assist the manual process conducted and support the decision made by human counterpart. Massive amount of research is conducted and the discipline is continuously growing by year. Due to its popularity, numerous societies are founded for the researchers to discuss and collaborate. Moreover, to facilitate the dissemination of the knowledge of computational intelligence in Digital Forensic, a number of journals and conferences have been published and organized periodically.

## References

- [1] Eckert, W.G.: Introduction to the Forensic Sciences. In: Eckert, W.G. (ed.) Introduction to Forensic Science Second Edition, p. 10. CRC Press, Inc., United States of America (1997)
- [2] Tilstone, W.J., Savage, K.A., Clark, L.A.: Forensic Science: An Encyclopedia of History, Methods, and Techniques. ABC-CLIO, Inc., Santa Barbara (2006)
- [3] Saferstein, R.: Criminalistics: An Introduction to Forensic Science. Prentice Hall, New York (2011)
- [4] Eckert, W.G.: Historical Development of Forensic Sciences. In: Eckert, W.G. (ed.) Introduction to Forensic Science Second Edition, p. 20. CRC Press, Inc., United States of America (1997)
- [5] Connors, E., Lundregan, T., Miller, N., McEwen, T.: Convicted by Juries, Exonerated by Science: Case Studies in the Use of DNA Evidence to Establish Innocence After Trial, vol. NCJ 161258, U.S. Department of Justice (1996)
- [6] Fantino, J.: The Police Chief. In: Forensic Science: A Fundamental Perspective, vol. 74(11), p. 1. International Association of Chiefs of Police, USA (2007)
- [7] Muriuki, P.N.: Pathological Truth: The Use of Forensic Science in Kenya's Criminal Justice System. World Academy of Science 78, 2089–2099 (2013)
- [8] Bertino, A.J., Bertino, P.N.: Forensic Science: Fundamentals and Investigations, 1st edn. South-Western Cengage Learning, USA (2008)
- [9] Lawson, T.F.: Before the Verdict and Beyond the Verdict: The CSI Infection Within Modern Criminal Jury Trials. Loyola University Chicago Law Journal 41(132), 119–173 (2009)
- [10] Shelton, D.E., Kim, Y.S., Barak, G.: An Indirect-Effects Model of Mediated Adjudication: The CSI Myth, the Tech Effect, and Metropolitan Jurors' Expectations for Scientific Evidence. Vanderbilt Journal of Entertainment and Technology Law 12(1), 1–43 (2009)
- [11] Podlas, K.: The CSI Effect and Other Forensic Fictions. Loyola of Los Angeles Entertainment Law Review 27(2), 87–125 (2006)
- [12] Reith, M., Carr, C., Gunsch, G.: An Examination of Digital Forensic Models. International Journal of Digital Evidence 1(3), 1–12 (2002)
- [13] Huang, J., Ling, Z., Xiang, T., Wang, J., Fu, X.: When Digital Forensic Research Meets Laws. In: 32nd International Conference on Distributed Computing Systems Workshops, Macau 2012. IEEE (2012)
- [14] Stahl, B., Carroll-Mayer, M., Elizondo, D., Wakunuma, K., Zheng, Y.: Intelligence Techniques in Computer Security and Forensics: At the Boundaries of Ethics and Law. In: Elizondo, D.A., Solanas, A., Martinez, A. (eds.) Computational Intelligence for Privacy and Security. SCI, vol. 394, pp. 237–258. Springer, Heidelberg (2012)
- [15] Elizondo, D.A., Solanas, A., Martínez-Ballesté, A.: Computational Intelligence for Privacy and Security: Introduction. In: Elizondo, D.A., Solanas, A., Martinez, A. (eds.) Computational Intelligence for Privacy and Security. SCI, vol. 394, pp. 1–4. Springer, Heidelberg (2012)
- [16] Mumford, C.: Synergy in Computational Intelligence. In: Mumford, C., Jain, L. (eds.) Computational Intelligence. Intelligent Systems Reference Library, pp. 3–21. Springer, Heidelberg (2009)
- [17] Deguang, W., Hua, S., Haibo, M.: Application of Adaptive Particle Swarm Optimization in Computer Forensics. In: 2010 WASE International Conference on Information Engineering (ICIE), August 14–15, pp. 147–149 (2010)
- [18] Feng, L., Cong, D., Shu, H., Liu, B.: Adaptive Halftone Watermarking Algorithm Based on Particle Swarm Optimization 8(3) (2013)

- [19] Pratama, S.F., Muda, A.K., Choo, Y.-H., Muda, N.A.: SOCIFS Feature Selection Framework for Handwritten Authorship. *International Journal of Hybrid Intelligent Systems* 10(2), 83–91 (2013), doi:10.3233/HIS-130167
- [20] Prandtstetter, M., Raidl, G.R.: Meta-heuristics for reconstructing cross cut shredded text documents. In: *Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation*, Montreal, Canada, pp. 349–356. ACM, 1569950 (2009)
- [21] Hanumantharaju, M.C., Aradhya, V.N.M., Ravishankar, M., Mamatha, A.: A particle swarm optimization method for tuning the parameters of multiscale retinex based color image enhancement. In: *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, Chennai, India, pp. 721–727. ACM, 2345514 (2012)
- [22] Nemati, S., Basiri, M.E.: Particle swarm optimization for feature selection in speaker verification. In: Di Chio, C., et al. (eds.) *EvoApplications 2010, Part I*. LNCS, vol. 6024, pp. 371–380. Springer, Heidelberg (2010)



# Digital Forensics 2.0

## A Review on Social Networks Forensics

MohammadReza Keyvanpour<sup>1</sup>, Mohammad Moradi<sup>2</sup>, and Farnak Hasanzadeh<sup>3</sup>

<sup>1</sup> Department of Computer Engineering, Alzahra University, Vanak, Tehran, Iran  
keyvanpour@alzahra.ac.ir

<sup>2</sup> Faculty of Computer and Information Technology Engineering,  
Qazvin Branch, Islamic Azad University, Qazvin, Iran  
Mhd.moradi@qiau.ac.ir

<sup>3</sup> Department of Software Engineering and Artificial Intelligence,  
Science and Research Branch, Islamic Azad University, Qazvin, Iran  
F.hasanzadeh@qiau.ac.ir

**Abstract.** Nowadays, Social Networks (SNs) are penetrating into all areas of human life including relationships, shopping, education and so on and this growing expansion is inevitable. In addition to their invaluable benefits, due to the plethora of confidential private/corporate information in SNs, these places become the potential target for criminal/illegal activities such as identity theft, fraud, organized crimes and even terrorist attacks. To cope with such issues, it is useful to incorporate social network forensics (SNF) techniques for analyzing and surveying social interactions to detect, predict and prevent all forms of potential criminal activities. This chapter is organized in two main parts. First, SNs, their security and privacy issues are introduced and analyzed. Then, as a reference point for future studies in the field, forensics methods within SNs are explained and classified; then the related literature is reviewed.

## 1 Introduction

Nowadays Social Networks are an integral part of a large amount of people's lives [1] in the broad range including students, athletes, artists and even politicians [2]. As is predictable, everybody uses SNs in his/ her own ways and interests. This fact shows flexibility and high-level potentiality of SNs, which make it adaptable to different situations and applications. In other words, declaring the situation of current web, social web – to bold the role and position of social media in general and social networks in particular- is the most appropriate and rational term. Due to the plethora of people, sensitive information and numerous SNs in different forms, these places turn into potential targets for attackers and become fertile fields for abusers. There are many threats to SNs which take advantages of their vulnerabilities and security breaches to attack privacy and exposure of confidential information. Fraud, espionage and scamming are only some of these criminal activities.

To alleviate such (mostly, privacy-related) issues, the most straightforward solution is to make the most of security mechanisms and configurations from SNs.

Moreover, as an influential factor, the ball is in the users' court since they should take care of their private information. Frankly speaking, in practical terms, there is not any completely secure and invulnerable place on the web.

In addition to user-centered solutions, another supervisory way is to use (social networks) forensics techniques [3,4] and tools for analyzing and surveying social interactions to detect, analyze, predict and prevent all forms of potential criminal activities. Despite their total higher costs, they are generally efficient. Of course, in contrast to security and privacy preserving mechanisms and because of the nature of results, forensics methods are used by private sectors and organizations rather than regular users. Even, in most of the cases, there is a need for legal authorizations. In the field of SNF as a subset of digital forensics, due to specific features of SNs, in addition to standard forensics techniques, several context-specific ones have been also proposed that will be considered in detail in the rest of this chapter.

In this chapter, security issues of SNs will be considered with a focus on forensics tasks.

In fact, this chapter is divided into two main parts:

- Security Issues of Social Networks  
which includes a short background on social networks as the context, their different types and applications, negative aspects and security and privacy issues of SNs.
- Social Networks Forensics  
Including brief introduction of digital forensics, approaches towards social networks forensics and related topics and issues such as Social Networks Analysis (SNA) and Social Networks Mining (SNM). Moreover, a review on the literature of SNF will be performed.

The structure of this chapter is as follows: in Section 2, we take an overview on Social Networks, their history and different types. Section 3 introduces drawbacks and problems of Social Networking Sites. Security issues with Social Networking sites as well as their different aspects are explained in Section 4. Section 5 provides an introduction to Social Networks Forensics and its differences with Computer forensics as well as proposing a conceptual architecture of a typical SNF system. The literature of the topic is reviewed in Section 6.

## 2 Overview of Social Networks

In this section, as the context of SNF, we take a general overview on social networks - better to say, Social Networking Sites (SNSs) - and their different types, applications and related issues.

With the advent of the World Wide Web in the early 90s, contribution of digital media has been greatly changed due to facilitation of communications, access to resources, information sharing and so on. Despite its numerous benefits, there was a substantial drawback in the old web structure which was degree of collaboration. In fact, during those days, information was produced by owners and used by users. In this model, the only way through which users could participate was their comments

and feedbacks if such mechanisms were provided. In other words, the information model was one-to-many. This model however had good reliability and degree of trust, could not leverage most of users' potentiality. By introducing web 2.0 [5], everything dramatically changed and users turned into most influential players of information creation/management process. In fact, web 2.0 introduced a paradigm shift by empowering the information model through involving users and evolved the model into many-to-many or by users for users. This (r)evolutionary shift has had a great impact on different aspects of users' lives and communications as well as affecting business models, education and even politics. Web 2.0's changes are mainly based on the concept of collaboration and major tools for achieving this concept are social media (SM). That is why some have called web 2.0 as social web.

Within the last decade, social media – in its different forms and specifically SNSs – have drastically gained more popularity among web users. As an example of such growth, LinkedIn - a professional SNS - in 2008 had over 25 million users and it has more than 225 million ones now (2013).

## 2.1 A Bit of History

In fact, there is not exact and definite history for social media and there are several different possible definitions for them. Nonetheless, regarding social media as the tools that humankind uses for communication and interaction with each other, paintings of prehistory cavemen are probably the first examples.

In addition to telegraph, telephone and radio [6] as early SM of modern days, in the computer era, SM was further developed during the 1970s by delivering first email in 1971 and creating MUD and BBS in 1978 as systems for interaction and message exchanging [7]. In 1979, Usenet was an early bulletin board that connected Duke University and the University of North Carolina. In 1992, Tripod was opened as an online community for college students and young adults. 1993 became a new milestone in the history of communications by inventing WWW technology at CERN by Tim Berners-Lee. Generally, the SM and SNSs that we know today appeared after introduction of the web. As a technical definition, Boyd and Ellison introduced [8] the SNSs as follows:

“we define social networking sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system“.

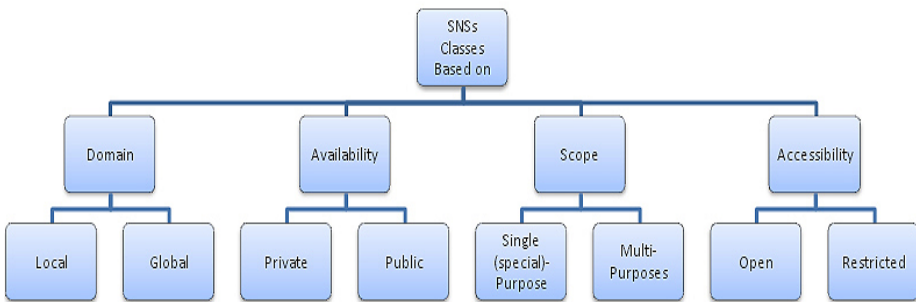
Based on this definition, the first instance was SixDegrees.com (launched in 1997, closed in 2000), which let users create profiles and list friends [8]. After that and until 2003, several SNSs including LiveJournal, MiGente, Cyworld and Fotolog were launched. The major wave of modern SNSs has been prompted since 2003 until now. LinkedIn, Last.Fm and MySpace (2003), Flickr, Orkut and limited (Harvard-only) version of Facebook (2004), YouTube and Yahoo!360 (2005) and Facebook and Twitter (2006) are some of the most famous players in the field of SNSs. Over recent years, SNSs have become global phenomena which are used daily by millions of people around the world as an integral part of their lives, business and education.

## 2.2 Different Types of SNSs

As another definition, an SN is a configuration of people connected to one another through interpersonal means such as friendship, common interests or ideas [9]. Based on this explanation, SNSs found broad meanings and many applications / systems can be fall into their categories.

There are almost no exact statistics on numbers of SNSs and their users; however, needless to say that their increasing growth and applications in different areas have made them the key player of the extensive information society. Due to different applications of SNSs, namely in business [10, 11, 12], medicine and healthcare [13, 14, 15], education [16, 17, 18] and industries [19, 20, 21], there are several hundred SNSs with different aims, scopes, users and applications. Moreover, there are multipurpose SNSs that have users from different folks. Due to these facts, classifying SNSs to appropriate categories could be a not-so-easy and imprecise task. Also, there are new SNSs with (probably) new applications that emerge continuously. Of course, there are several studies in which authors have focused on different types and metrics of SNSs to compare and analyze them based on some features. As some examples, [22, 23, 24, 25] could be mentioned. Even, in [26], authors classified SNSs based on the calculated network indexes and communication patterns.

Despite these difficulties, in this section, we propose a general classification of SNSs based on their most important features, as given in Figure 1. We believe that any SNS could fall into one of the introduced classes. Nevertheless, there are some other criteria like number of users, applications and density that could be regarded while categorizing SNSs.



**Fig. 1.** The Proposed Classification of SNSs (based on four main classes)

- **Domain:** The SNSs based on their underlying concepts and goals may be available around the world or only from specific geographical places, e.g. within a town or even a corporation.
- **Availability:** Level of availability for a given SNS depends on multiple criteria such as degree of sensitiveness, goal of SN and type of applications. For example, it is obvious that an intra-corporation SNS should be only accessible by its personnel, not others. Intra-university SNSs (like early days of Facebook), as an example, are among them. A local SNS is a private one; but, it is not true in the reverse

direction since a private SN may spread around the world. Yammer [27], as an example, is one of the most important and popular commercial private social networks solution. Also, as a scholarly work, in [28], authors proposed – as claimed – the first complete architecture and implementation of virtual private SNSs for Facebook.

- **Scope:** Most of the social networking sites out there are multipurpose since users could do everything they want ranging from posting images and videos to arranging meetings and also scientific conversations. Generally, such SNSs have users from different folks; however, special purpose SNSs have users with a specific interest or aim. This type of SNSs is useful for special interest groups and could work for their users as a bulletin board or discussion desk. In this fashion, [29] proposed a special-purpose SNS that would allow people to communicate their status with friends and family when finding themselves caught up in a large disaster.
- **Accessibility:** Based on the content and theme, SNSs could be restricted for usage of only, for example, adults or women. Of course, in contrast to other classes, checking the requirements of claimed users in this category is harder and costly. In such SNSs, some authorization mechanisms like users' unique identity and so forth are needed. Anyway, there is not any absolute way for verifying users regarding their privacy.

All in all, despite the numerous and undeniable benefits and applications, due to the users' carefulness and plethora of information and connections, SNSs cause several drawbacks that make them an insecure and, in some cases, dangerous space for users. In the following section, these issues are pointed out.

### 3 Dark Sides of Social Networks

Due to their useful features and capabilities that provide communities with unprecedented opportunities, drawbacks of SNSs are usually neglected or, in some cases, underestimated. Such inadvertences cause essential problems and issues in most of the cases that may affect users' personal life or even threaten national security.

People that are used to staying connected in SNSs usually like to do the same in their workplaces. Although this issue allows employees to share information with one another, in some cases, it could be regarded as an obstacle for employees' efficiency. This way, according to a study by Nucleus Research [30], companies that allow employees to use Facebook during the work day lose 1.5 percent of their productivity. Moreover, there are other issues of using SNSs at work such as consuming extra bandwidth [31], wasting time [32] and privacy-related problems. Such issues urge organizations to ban everyone, except managers and the social media team, from using this technology at work.

Another impalpable side-effect of SNSs is their effect on family relationships and divorce. Destructiveness of couple misbehaviors in SNSs causes most of familial problems. There are several reports and statistics about such problems. For example, Dailymail reported [33] that social networking sites were cited as a reason for one third of divorces in 2010, in which unreasonable behavior was a factor, according to

law firm Divorce-Online. Also, based on the Guardian report [34], “A 2010 survey by the American Academy of Matrimonial Lawyers (AAML) found that four out of five lawyers reported an increasing number of divorce cases while citing evidence derived from social networking sites in the past five years, with Facebook being the market leader”.

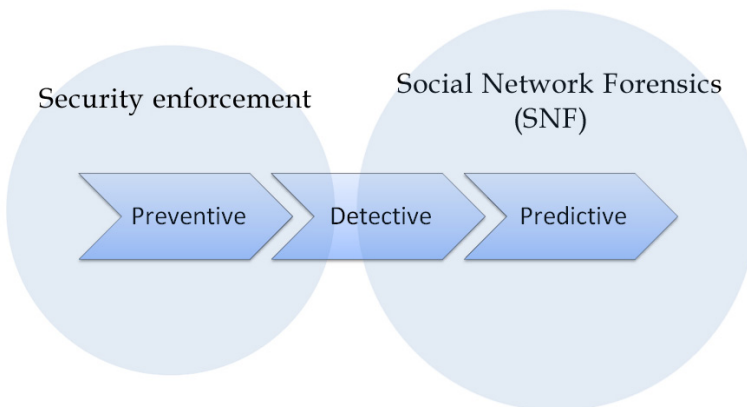
In fact, the most important issue about SNSs is privacy problems, which stem from two main sources. First, users’ carelessness about observing security and privacy guidelines and second intrinsic security breaches of SNSs. As a piece of evidence, it has been reported that 13 million of US Facebook users do not use or are oblivious to privacy controls [35]. Moreover, Facebook has released statistics showing that seemingly there are more than 83 million fake accounts on its social network [36].

Furthermore, during recent years, SNSs have become an applicable environment for terrorist and criminal activities, bullying, scamming, stalking, spamming, espionage and so on. National security could be also threatened via SNSs. All such issues totally enforce the governments and security authorities to think about monitoring and analyzing SNSs in order to avoid and control the consequences. As one such precautionary proceedings, the FBI is asking the industry for help in developing a far-reaching data-mining application that can gather and analyze intelligence from social media sites [37].

Generally, there are two major approaches for coping with SNS-initiated threats.

The first one is preventive approach, which is mainly based on security and privacy preserving techniques and guidelines. The key player in this approach is the user him/herself that must follow instructions and be careful about his/her private and confidential information.

Another one, diagnostic approach, is mainly an after-incident approach to analyze, survey and detect reasons and violations. Organizations, private sectors and legal authorities rather than regular users are in charge of performing such techniques. From another perspective, this approach could be regarded as a predictive one, since through the analysis of current status and statistics, authorities could predict possible / probable criminal acts in future. The Figure 2 illustrates positions of the mentioned approaches.



**Fig. 2.** Positions of Security and Forensics Mechanisms

## 4 Security Issues of SNSs

The increasing popularity and usage of SNSs as well as stashing the large amounts of sensitive (private, corporate, etc) information turn them into potential targets for abuse in its different forms. Based on the concept of “know to survive”, in coping with these threats, the first and foremost step is to know about them. Doing so, in this section, we take a general look at different types and aspects of SNSs’ security issues.

### 4.1 Security Aspects

Security issues of SNSs could be considered from two viewpoints: SNSs and users. Because of the centralized administration of SNSs and existence of intrusion detection, recovery and backup mechanisms, the sites themselves are not the main target of abuses. Since SNSs have millions of users from different folks, certainly many of them do not enough know the basics of security and privacy preservation. Moreover, signing up into SNSs is very simple and often without any minimum security requirements. Unaware users that share their private information without thinking about the consequences, those who accept every friendship request, those who click on any tiny URL and many others are the potential victims of criminal activities and abuses of wicked (ab)users.

Of course, from a supervisory view, there are two main aspects for security in SNSs:

1. SNS-related issues that refer to common security and privacy issues related to them.
2. SNS-based issues that point to the leveraging SNS platform for criminal and illegal acts such as organized crimes, terrorist activities and so forth.

Below, security threats of SNSs will be considered.

### 4.2 Security Threats

Since SNSs are used by people in different positions and places, their risks and threats have several different aspects. Most of the time, users' privacy and personal information are targets of attacks. Sometimes, based on the users’ roles and positions within an organization or company, they will be attacked for what they know or what they have access to. These types of attacks usually are of organized crimes with the intention of corporate espionage, threatening national security, etc. There are several studies on these issues in the literature such as [38, 39] and [40], which argue SNSs and national security and an agent-based model on a social network in the case of defense industrial base, respectively. Unfortunately, despite the importance and prevalence of SNS-originated risks, these topics have not been well-studied and there are many open issues.

According to the Symantec Internet Security Threat Report [41], top 5 social media attacks in 2012 were as shown in Figure 3. Each of these attacks explained [41] as follows:.

**Fake Offering:** These scams invite social network users to join a fake event or group with incentives such as free gift cards. Joining often requires the user to share credentials with the attacker or send a text to a premium rate number.

**Manual Sharing Scams:** These rely on victims to actually do the hard work of sharing the scam by presenting them with intriguing videos, fake offers or messages that they share with their friends.

**Likejacking:** Using fake “Like” buttons, attackers trick users into clicking website buttons that install malware and may post updates on a user’s newsfeed, spreading the attack.

**Fake Plug-in Scams:** Users are tricked into downloading fake browser extensions on their machines. Rogue browser extensions can pose like legitimate extensions but when installed can steal sensitive information from the infected machine.

**Copy and Paste Scams:** Users are invited to paste malicious JavaScript code directly into their browser’s address bar in the hope of receiving a gift coupon in return.”

The key factors that make SNSs fertile for such attacks, laid on the nature of the environment (SNS) and its users’ behaviors. As a real world example, sharing is the most commonplace activity within SNSs. Therefore, it is possible to easily spread a malware throughout the networks. Of course, prerequisite of such widespread outbreak is trust among users, specifically friends.

The literature and security community have proposed several lists and classes for introducing SNSs threats and risks from different perspectives; however, as mentioned earlier, we look at them from users’ standpoint, based on which we classify the threats as follows:

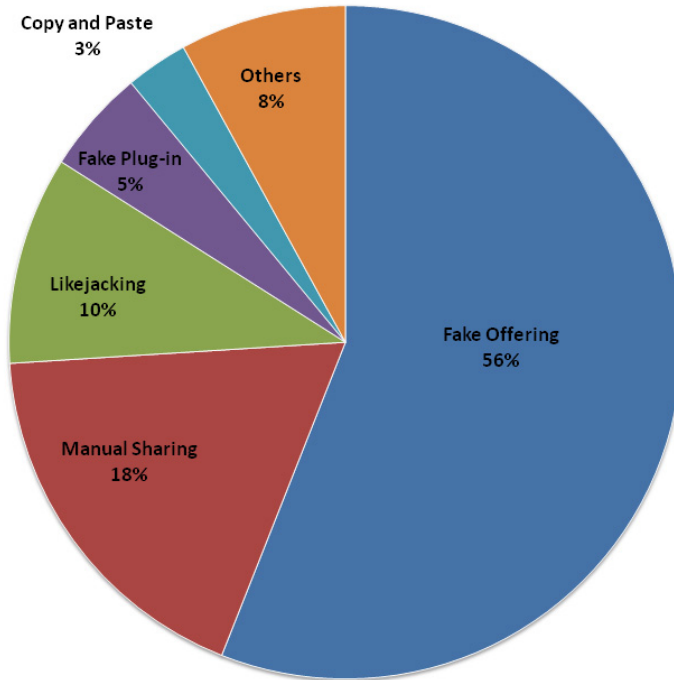
- **Propagation:** Based on what attackers want from users to do
- **Infiltration:** Based on the position of users
- **Disclosure:** Based on what users have

#### 4.2.1 Propagation

One of the popular types of attacks is propagation by leveraging users’ connections and capabilities through their relationships. The main goals of such activities are viral marketing and advertising, spamming, roorbacking and spreading malicious malwares.

Spamming as one of the low-cost (and usually effective) attempts for viral marketing could be also regarded as a tool for spreading organized messages through the population. As noted in [42], there are two main categories for spamming activities: context-aware [43] and broadcast spamming. The former provides spammers with





**Fig. 3.** Top 5 Social Media attacks in 2012 (statistics obtained from Symantec Internet Security Threat Report [41])

high click-through rate by taking advantage of the shared context among friends on social networks. In fact, this class of spamming is the targeted one that takes advantage from the trust among related users within SNSs. The latter class does not have any specific targets, but rather abuses public interaction mechanisms to disseminate information [42].

Another form of abuse in propagation class is (organized) rumor spreading. This action is a usual one in daily life of most people and takes benefits from word-of-mouth power of social (human) communities. However, when it comes to organized purposes, it could affect society, business and even politics, for example, by defaming a politician. Nonetheless, this approach could be useful in some applications such as disaster outbreak and shaping public thought [44]. There are many studies on different aspects of this topic in the literature, namely [45, 46, 47, 48].

Shortened URLs (known as tiny URLs) are phenomena of SNSs –specifically Twitter– that beside their benefits, could be used as a deception tool to trick (tempt) users into visiting malicious sites. They can extract personal (and corporate) information, specifically if accessed through a workplace computer. In other words, they hide the true link to web sites. Twitter is especially vulnerable to this method because it is easy to retweet a post so that it could be eventually seen by hundreds of thousands of people. Decoding such links before clicking should be the first action to do.

Spreading malicious software via tiny URLs, profile, interaction and third-party applications within SNSs is another form of propagation-based abuse. Extracting users' information and damaging victims' systems are two of the goals of propagators. One of the most famous worms which has successfully propagated through SNSs is the Koobface worm [49].

There are several instances of attacks that target great SNSs by malwares such as worms, information stealers and password stealers: Grey Goo targeting Second Life, JS/SpaceFlash targeting MySpace, Kut Wormer and Scrapkut targeting Orkut and Secret Crush targeting Facebook [50]. Cross Site Scripting (XSS) and Cross Site Request Forgery (CSRF) are other threats from this category.

#### **4.2.2 Infiltration**

Based on the position of SNSs users in their organization, corporation and so on, they usually have access to the sensitive information that could be invaluable and even critical, especially in business and political contexts. More often than not, such users are tempting targets. Information leakage and corporate espionage are potential threats of such cases. As reported in [51], during six-month monitoring of 20 companies via an active social media by Cyberoam researchers, leakage of information was found from all of them.

Of course, severity of security mechanisms directly depends on degree of sensitivity. This way, there are specific intra-organizational rules that managers specify for using SNSs within their organizations. In other words, to avoid such threats, companies should specify their comprehensive policies against SNSs. In the first step, companies should determine how they want to use SNSs and why. The response to these questions shapes their approach to SNSs. Moreover, such policies should include user guidelines, content standards, monitoring schedule, limitation and so on. It seems that total banning of users from using SNSs is not an efficient approach. To promote security level of system (and avoid insecure connections), most organizations allow their employees to use SNSs under their specified policies and conditions.

#### **4.2.3 Disclosure**

Undoubtedly, a major target of SNSs attacks is users' privacy related assets including their private (sensitive) information, passwords, relationships, identity, etc. Such information could be used for different purposes like blackmailing, targeted spamming, scamming, extortion, selling to the third parties, defaming and Indignity. Also, the data aggregator applications could collect the same users' information from different sites like Facebook, LinkedIn and Twitter to inference their hidden info, relationships and interests.

Since users voluntarily reveal their private information in SNSs, it is possible to breach their privacy from service providers (e.g. for advertising purposes), from other users and third parties [42] (such as social games).

Because of the importance of privacy preservation in SNSs, there are many researches that have considered this topic from different perspectives. As some examples: in [52], authors surveyed the literature on privacy in social networks with focus on both online social networks and online affiliation networks. Also, they considered

two scenarios for privacy in social networks: privacy breaches and data anonymization. In [53], solution to security and privacy in Mobile Social Networking was studied. [54] Presented a Digital Rights Management (DRM) approach to privacy issues in online social networks.

The privacy issues of Facebook as the most prominent SNS were studied in [55]. [56, 57, 58] are other examples of such studies. There are different methods for acquiring users' private information; however, the most important attacks are social engineering (e.g. by profile impersonation) and phishing attacks.

#### 4.2.4 SNSs as Coordinators

In addition to using SNSs as a platform for taking illegal and criminal measures, based on its intrinsic nature, SNSs could be used as communication tools for arranging criminal activities and meetings (specifically terrorist acts and organized crimes). Moreover, through leveraging capabilities of SNSs, terrorists could spread their ideas and propaganda, recruit new members and intimidate others. As the inclination of the research community to this phenomenon in recent years, some examples are [59, 60, 61]. Further, Cyberbanging [62] (presence of street gangs on SNSs) is another notable issue in the context of SNS-enabled organized crimes. This subject has been also well-studied over the recent years [63].

Although there are several techniques for securing SNSs and their related activities, it is a fact that, in the cyberspace, security is a relative concept. Moreover, security is non-functional and there is not a specific and absolute measure for its evaluation. All in all, to complete the security cycle, the post-incident phase has an influential position in analyzing, tracking and predicting attacks and violations. Such forensics (related) activities are often considered in the case of legal issues (law enforcement and investigation), national security, fraud and other sensitive cases. In the next section, we explain the Social Network Forensics (SNF) and consider its different aspects and methods.

## 5 Social Networks Forensics

Only relying on security and privacy-preserving mechanisms – because of intrinsic security problems of sites, unaware and careless users, huge amount of records, etc – is not definitely sufficient for protecting sites, users and their information.

To be able to track, analyze, prevent and predict attacks and abuse, there is a need for post-incident mechanisms. Such mechanisms that are known as Digital Forensics (DF) have been around for years in the context of digital devices. In fact, DF is a subset of a more-general class, Computer Forensics. Formally speaking, DF is defined [64] as

“The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitation or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations”.

Although the term Social Network Forensics (SNF) - could be defined as web 2.0's version of DF - is a newcomer to the forensics community, it is the rational and predictable continuation of DF. In other words, currently, web 2.0 (better to say, Social web) is the predominant type of media which is used by people for storing their files, communicating with each other, etc. Also, based on the paradigm shift of web 2.0 flourished by evolving SNSs, the term Digital Forensics 2.0 (DF 2.0) could be truly used for describing SNF.

## 5.1 Scope and Challenges

The promise of SNF is that of its ancestor, DF; however, there are several differences due to broad scope of SNF in which large amount of information, connections and relationships are the subject of investigation and tracking. This broad range is simultaneously one of the most important challenges of SNF in acquiring information, evidence and monitoring.

To name some of the tasks in SNF, the following can be mentioned:

- Proving whether a person is cyber-bullied or threatened by another
- Establishing whether a subject is associated with another person of interest
- Detecting pieces of evidence from a convict's posts, tweets
- Finding sources of roorbacking and national security-related rumors
- Finding huge-spammers
- Finding disclosure sources of vital (or private) information
- Finding the person posting the offending content
- Uncovering terrorists and criminal networks
- Predicting organized crimes
- etc.

In fact, SNSs could be considered in forensics analysis from two viewpoints; first and commonly, by tracking and detecting the anomaly, abuses and so on and, second, through gathering the images users uploaded, comments their posted or any things that could be ascribed to an individual, e.g. his/her threatening statements, as pieces of evidences.

From the users' perspective, there are two main directions for SNF analysis and investigations:

- **Their attributions:** that could be used as pieces of evidence against them
- **Their complaints:** that usually take place after facing any forms of violations or threats

Of course, as stated before, most of the times, SNF analysis is performed by private sectors, intelligence agencies and legal authorities.

Since the content of SNSs is directly (or indirectly) related to individuals and legal entities, there is a need for legal authorization in order to be able to check, trace and monitor them. Moreover, as social networks are very dynamic (continuously changing) in nature and the amount of data is rather big, data acquisition for forensics

analysis is a great challenge. All in all, despite its benefits, SNF is a costly and relatively hard-to-perform task.

Due to these facts, practical experiences of using SNF are related to legal authorities – especially police and intelligence agencies. For example, as Mashable reported [65] the efforts currently underway by police in Vancouver, BC provides an excellent example of investigation leveraging social media for identifying those responsible for the Stanley Cup riots in 2011.

## 5.2 SNF vs. CF

To underline fundamental differences between traditional computer (digital) forensics and brand new SNF, the most important features of each one are summarized and compared from two different (major) perspectives in the following tables. This comparison also illustrates the challenges each approach has to face with.

The first and foremost aspect for discriminating CF from SNF is the data to be investigated. In fact, type of data each of the process has to deal with is the core of their differences.

**Table 1.** Comparing SNF and CF based on type of data

Measure 1 : Type of Data	Computer Forensics (CF)	Social Network Forensics (SNF)
Context (scope)	Digital devices (limited)	SNSs (widespread and interconnected)
Nature of data	(usually) Static	(most often) Dynamic
Complexity	(depending on the situation) Low to high	(based on the underlying structure) High
Data quantity	(depending on the situation) Small to large	(based on the underlying structure) Large and increasing

As noted in Table 1, due to specific features of datasets (context), SNF process was far different from CF. SNS data, because of their dynamic nature as well as large number of users interacting within them, continuously grew both in size (quantity) and domain (interconnections). Therefore, processing such datasets is a more challenging task than (usually) stationary datasets collected from digital devices, e.g. hard disk.

The other important aspect of difference between CF and SNF, which is in close relationship with the context (datasets), is quality of processing task or simply operation.

**Table 2.** Comparing SNF and CF based on quality of operation

Measure 2 : Operation	Computer Forensics (CF)	Social Network Forensics (SNF)
Data acquisition	(depending on the situation) Easy to hard	(most often) Hard and challenging
Executability	(depending on the situation) Easy to hard	(most often) Hard
Accuracy	More precise	(currently) less precise
Cost	(depending on the situation) Low to high	(most often) High

It is needless to say that, the more complex the context, the harder the operation for implementation. As in Table 2, due to complicated processing task of SNF, overall cost (including time, complexity and resource utilization) is usually more than CF. Also, because of several essential problems with acquiring SNSs data as well as applying appropriate algorithms – on their specific, graph-based structure-, level of accuracy for such methods is less than similar tasks in CF. However, depending on the special situations that may occur, for example in the case of investigating encrypted data, CF process could be hard to execute and less precise.

Last but not least, another substantial difference between these two approaches towards forensics investigations is privacy concerns. Although privacy preserving is a challenging issue for both approaches, since SNSs are stockpile of a large amount of confidential information and interconnections, consequences of privacy policy violation may be more disastrous than such cases in other contexts. Of course, privacy related issues have another aspect in reverse direction. Assume a case that the process should keep track of an anonymized criminal within a given SNS. It is as arduous as finding a needle in a haystack.

### 5.3 SNF Approaches

Since SNF is a new subject matter, there are not so many works about it. Of course, there are several studies that have focused on tasks such as monitoring SNS, community detection and so on. In fact, such activities are of SNF process and could be implemented for forensics purposes.

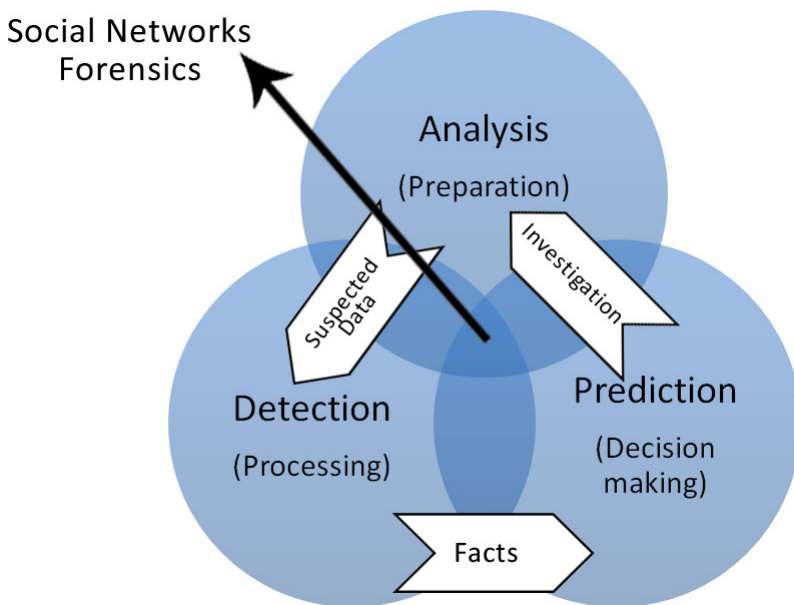
Based on these topics, the literature of SNF could be divided into two categories of implicit and explicit studies; the former refers to the forensic-related researches and the latter points to the studies that indirectly consider forensics tasks.

In addition to standard forensics techniques, in the context of SNF, Social Networks Analysis (SNA) and Social Networks Mining (SNM) techniques are also used.

Before reviewing the literature of SNF in order to organize researches as a framework for future studies and based on different tasks performed in this process, we classify SNF approaches to three classes as follows. Generally, SNF is mainly based on SNA in its different forms, like SNM.

- **Analysis:** in which SNs will be the subject of analysis for different purposes, like monitoring, detecting anomaly, visualizing and so on.
- **Detection:** that is the next step after analysis for detecting the anomaly, crime, etc.
- **Prediction:** after analysis and detection of patterns and trends, it is possible to predict future crimes and illegal actions through SNSs.

It is needless to say that the mentioned classes have no solid borders and have overlaps with each other. In fact, our intention is to organize SNF studies into the most appropriate classes based on the selected approaches and applications as well. The following figure (Figure 4) illustrates the relationships among above mentioned classes.



**Fig. 4.** Relationships among Different SNF Phases and Steps

### 5.3.1 Conceptual Framework

Based on the mentioned approaches, the general conceptual framework of SNF process is depicted in Figure 5.

This framework is composed of three main phases as follows:

**Preparation:** The first phase is an important step towards a successful SNF procedure. Since this step deals with the data (set), it has two essential tasks:

- *Scope definition:* As SNSs have various types and applications, selecting scope that the process should apply to in fact specifies the boundary of process. This is because the context defines which types of algorithms should be used. For example, static and dynamic social networks pose different requirements. Moreover, level of the required details to be investigated which will be determined in this step specifies the amount of data that should be obtained and its depth. All in all, this step is application-driven.
- *Data acquisition:* Probably, one of the most important and hard-to-implement steps of SNF process is data acquisition. Since this step provides materials of investigation procedure, any deficiency within that definitely affects the results and process in general. Main concerns of this step are perhaps validity of the gathered data, especially from temporal aspect.

**Processing:** This phase as the heart of SNF procedure involves three main tasks of analysis, detection and prediction. Details of these tasks will be reviewed in the next section. Generally, the most costly phase of the procedure is processing phase. The bidirectional arrow between preparation and processing phase means that, based on the situation and its requirements, the algorithms may need more data or details to be obtained.

**Reporting:** After the investigation process is completed, the results should be generated. Such results provide discovered facts for further operations, namely decision making or presenting to the legal authorities. There are several main tasks within this phase as follows:

- *Classifying results:* Organizing results for further processes such as comparison and analysis tasks.
- *Visualizing facts:* Representing facts in a user-friendly manner as well as helping to discover hidden knowledge.
- *Documentation of the procedure:* Similar to every software project, this phase helps others (analysts, investigators, project managers and researchers) to know about quality of the procedure implementation and technical reports.

Another important component of this framework is evidence repository. The underlying concept of existence of such component is to store the discovered/extracted pieces of evidence from each of the phases. This invaluable information could be used for different purposes such as statistical analysis of patterns, workflow tuning, process optimization and so on. Moreover, such information may serve as training data for intelligent methods.



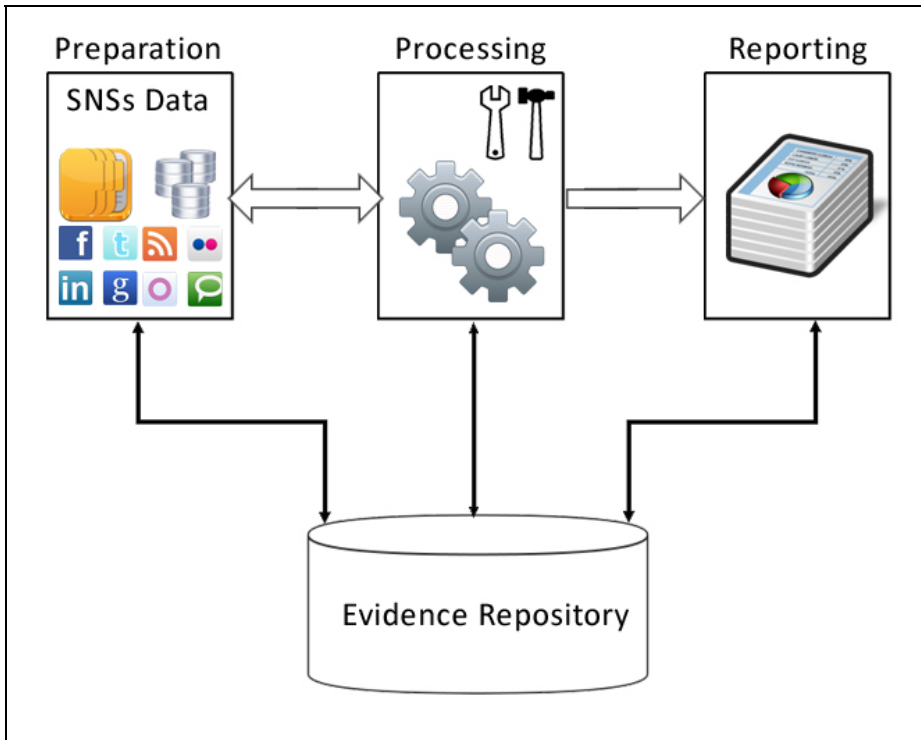


Fig. 5. General conceptual framework of SNF process

## 6 Literature Review

In this section, we have attempted to perform a comprehensive review on the SNF literature as a reference point for future studies. The structure of this review is as follows: first, the researches that have directly considered SNF will be reviewed. Then, other implicit works will be surveyed according to the mentioned classifications.

### 6.1 Focused Researches

As stated before, there are not so many studies that have directly pointed to SNF. Moreover, only in recent years with evolving SNSs, this topic has emerged as a new subject in the context of DF. This way, early works of the topics go back to 2009. Accordingly, in [66], subject of behavior modeling and forensics was considered for multimedia social networks. The authors discussed recent advances in the study of human dynamics for multimedia social networks and reviewed a few methodologies to investigate impact of human factors on multimedia security from a signal processing perspective. Also, a framework was also presented for modeling and analyzing user behaviors.

Authors of [3, 4] proposed some SNF tools that could be used to help in investigation of social network site crimes and raise user awareness. These tools could be used to help protect and educate users from the beginning of their social network site interactions, thereby preventing crimes from even occurring.

In the September 2009 issue of Financial Fraud Law Report, the author explored online sources of evidence of communications as well as manner of approaching the new frontier of social networking communities as they were related to information gathering in litigation and investigations [67]. Moreover, to prevent corporate information leakage, the author advised companies to be aware of how current social networking channels were employed within their organizations.

In [68], the risk of cyber crime against a single user who was not sufficiently careful about protecting his/her information was studied through a real criminal investigation. Also, in this paper, it was stated that, although the used tool could be handy for forensics purposes, it could also provide opportunities for cyber criminals to collect personal information about others' everyday web usage. Moreover, this paper provided several pieces of advice for protecting personal information.

In [69], the authors presented a novel method for automated collection of digital evidence from social networking services. As the advantages of their proposed framework (named Social Snapshot), the following could be mentioned: compared with state-of-the-art web crawling techniques, this approach significantly reduced network traffic, was easier to maintain and had access to additional and hidden information.

Following this trend, in [70], a standard model of digital forensic investigation was proposed for Online Social Networks (OSN). This OSN-specific designed model was composed of four processes as: preliminary, investigation, analysis and evaluation.

In [71], the authors identified important data sources and analytical methods for automated forensics analysis on social network user data. Furthermore, they demonstrated how these data sources could be evaluated in an automated fashion and without any need for collaboration from the social network operator. Finally, they showed feasibility of their approach on the basis of Facebook through implementing a proof-of-concept application for creating social interconnection and social interaction graphs.

As a focused research, authors in [72] considered forensics analysis of images on OSNs (specifically, Facebook, Badoo and Google+) by analyzing characteristics of images published on them. The analysis mainly focused on how the OSN processes the uploaded images and changes were made to some of the characteristics, such as JPEG quantization table, pixel resolution and related metadata. As a result of their experiments, it could be inferred whether an image had been downloaded from an OSN or not.

As another focused study, [73] regarded facial recognition software as SNF tools and took some practical tests on Facebook for image matching.

Forensics analysis of social networking applications on mobile devices was performed in [74]. This study focused on the recovery of artifacts and traces related to the use of social networking applications on a variety of Smartphones using different operating systems. Also, the study explored forensics acquisition, analysis and

examination of logical backup copies of the three Smartphones (BlackBerry, Android and iPhone).

The main contributions of the work reported in [75] were design, development and evaluation of a novel (Latent Dirichlet Allocation) LDA-based social media analytical model to combat cybercrimes and facilitate cybercrime forensics. Their preliminary experimental result showed that the proposed model could discover semantically rich and relevant latent concepts related to cybercrimes.

In [76], the authors provided a real case review concerning crime scene reconstruction with respect to the previous Facebook session of the victim based on the digital evidence collected and analyzed via live internal data acquisition.

Finally, in his thesis, Son [77] evaluated evidence extraction tools in a systematic and forensically sound manner to measure capability of extracting evidence from SNSs in different test scenarios.

There are many studies in the literature devoted to analysis of SNSs for different purposes that could be established in the forensics analysis, such as spam detection, community detection, etc; however, so far, such activities have not mainly focused on forensics approaches and could fall into the forensics investigation processes. Therefore, for the sake of extending SNF borders and shedding some light for future studies, we review the related works based on the mentioned classifications.

## 6.2 Analysis

Social Network Analysis (SNA) is one of the most interested technologies for studying different aspects of SNSs including criminal and terrorist networks. SNA techniques describe the roles and interaction among the actors within SNs and could reveal hidden facts and approaches as well as detecting subgroups, discovering their patterns of interaction, identifying central individuals [78] and so on. Generally, the analysis is done by collecting data from various incidents and sources related to the case under research (context) and discovering the patterns, structures and flow of information in the context network (in forensics analysis, the context is usually criminal or terrorist networks). Thus, there are several studies that have been devoted to analyzing criminal and terrorist SNs like [79, 80, 81, 82, 83, 84]. The general process of SNA – e.g. for terrorist networks- could be depicted as follows: Interrelationships are displayed through graphs. Graphs are built by analyzing the data including nodes (terrorists) and links (relationships). Then, these relationships are used to understand information about people and group [78]. SNA has different forms such as Social Networks Mining (SNM), monitoring, etc.

### 6.2.1 Monitoring

SN monitoring for observing users' interactions, connections and behaviors is a supervisory task that comes in handy in different applications like checking kids about using SNSs and so on. Of course, the most important applications of monitoring are security and forensics related issues. Most of the time, such tasks need to be legalized; therefore, governmental organizations usually perform them. For example, as NY Times reported [85], "The Department of Homeland Security paid a contractor in

2009 to monitor social networking sites — like Facebook, blogs and reader comments on a news article — to see how the residents of Standish, Mich., were reacting to a proposal to move detainees from Guantánamo Bay, Cuba, to a local prison there, according to newly disclosed documents”. Similarly, according to the BBC report [86], SNSs like Facebook could be monitored by the UK government under proposals to make them keep details of users' contacts.

As a scholarly work, in [87], the authors proposed a framework, SPAM, for monitoring social profile abuse. Also, they proposed a four-class classification model for measuring profile similarity indexing based on fine-grained user similarity features. [88, 89] could be mentioned as less-related works in this field that could be regarded as inspirational examples for future works in forensics context.

### **6.2.2 Mining**

Since digital forensics process often face the challenge of analyzing large volumes of data involved in criminal and terrorist activities, therefore, an appropriate theoretical method for that is to leverage data mining capabilities. Utilizing this idea, there are several works in traditional DF such as [90, 91, 92]. Applying this approach in SNF context introduces the need for using Social Networks Mining (SNM) and Knowledge Discovery (KDD) techniques. [93, 94, 95, 96, 97] are of such researches. Also, finding trends and frequent (interesting) patterns through SNs is another useful task of SNM techniques for analyzing SNs [98, 99]. Mining communities and graph patterns - as two methodologies - in SNs are also could be used in pattern mining to find (repetitive/periodical) behaviors. In a general view, content (text, multimedia), structure and relationships between people (account) could be the subject of mining to extract pieces of evidence and facts for analysis purposes.

### **6.2.3 Visualization**

A good starting point for the analysis process is to map (criminals/illegal/abnormal) activities to the visualized graph that displays associations and relationships between the acts and individuals. Visualization methods prepare alternative means to perform the analysis task by transforming complicated data characteristics into clear patterns to view the relationships uncovered. Data visualization is generally associated with data mining as a post-processing (mining) step. The cliché of “a picture is worth a thousand words” could be realized in such a case. Beside several tools and libraries that are out there for SN visualization like TouchGraph, Google+ Ripples and MentionMap, the literature of this topic has many scholarly examples, namely [100, 101, 102].

## **6.3 Detection**

As the post-analysis phase and for leveraging initial facts discovered from the past step, the case (crime or illegal acts, evidences, etc.) will be deeply investigated to detect and identify sources, relationships and unusual (suspected) behaviors.

One of the most important and well-studied topics in this category is spam detection. Following this trend, in [103] authors proposed a framework for unsupervised spam detection in SNSs. Then, they tested the models on data from a popular social network (the largest Dutch social networking site, Hyves) and compared the models in terms of two baselines, based on message content and raw report counts. The subject of detecting social spam campaigns and spammer was studied in [104, 105], respectively. Also, there are other activities in this field such as [106, 107, 108] that have considered different aspects of spam detection. Furthermore, using data mining techniques –specifically classification ones- such as decisions tree, neural networks, support vector machines, k-nearest neighbors and Bayesian classification for spam and fraud detection was introduced in [109, 110].

In addition to spam, detection of other abnormal behaviors in SNs has been a subject for researches. For example, detecting new trends in terrorist networks was studied in [111]. Moreover, a new case study was examined to show usefulness of the presented techniques.

In fact, unfolding the groups and communities within SNs is usually based on mining approaches; for example, graph matching techniques are recommended for group detection tasks [94]. As a work in this class, [112] considered distributed community detection in SNs using genetic algorithms. Community detection and tracking the evolution of communities in SNs were also studied in [113, 114]. In addition to detection of terrorists and criminal groups in SNs, identifying anomalies and suspected behaviors could provide the forensics process with implicit but invaluable information. The analysis of mobile networks' communication patterns in the presence of some anomalous “real world event” was such an activity which was presented in [115]. The anomaly detection in SNs was also considered in [116, 117, 118, 119]. Since detection topic is very extensive, there are many other works that could be useful for specific cases. For instance, the problem of fake profile identification is another approach. It is notable that a user with a fake profile is a potential instance for being monitored against misbehavior. Concerning this issue, [120] evaluated e implications of fake user profiles on Facebook, as a case. Detection of Cyberbullying [121], fraud detection [122], identifying similar people [123] and identifying events [124] are some of the specific-purpose studies. Employment of Intrusion Detection Systems (IDSs) is another feasible approach within this category. In fact, the finding of the content from the existing sites and known terrorist traffic on social networks could be accomplished by using such systems. The IDS sequentially spies different activities within the network traffic to approximate possible attacks [125].

## 6.4 Prediction

The next rational phase of SNF, after analysis and detection, is to use patterns and findings of suspected and potential criminals/abusers so as to discover the performed crimes for legal investigations, prevent probable ones and predict ongoing crimes and criminal acts through the detected communities. A predictive task (policy) could be defined as “a multi-disciplinary, law enforcement-based strategy that brings together advanced technologies, criminological theory, predictive analysis, and tactical

operations that ultimately lead to results and outcomes -crime reduction, management efficiency, and safer communities.” [126]

In previous years, being able to predict future crimes and illegal actions was only an interesting idea for Sci-Fi movies like *Minority Report*; however, through SNF and by means of SNM and other techniques, this dream came true. As a governmental project, US defense Raytheon has developed some software that uses social networking sites to track users’ movements and is able to predict where a person will be and their future behavior [127]. Such tools come in handy in preserving national security and combat against organized crimes. However, privacy concerns are the most important issues around such techniques.

In SNA process, by utilizing link prediction approach in SNs, it is possible to infer which new interactions among its members are likely to occur in near future. This question was formalized in [128] to develop approaches for link prediction based on measures for analyzing the proximity of nodes in a network. Although the topics of this class are currently less-focused, there are several studies that pay attention to its some aspects. Authors in [129] proposed a mechanism for identifying correlations between spatiotemporal crime patterns and social media trends in order to predict and prevent such actions. Furthermore, a general review of current crime prediction techniques was performed in [130], which could be inspirational for future works in SNF context.

## 7 Conclusion and Future Directions

Flourishing SNs in the age of web 2.0 as the most important medium for interaction and communication among users on the one side and their extensive applications in a broad range of domains on the other side has introduced several privacy and security related issues. Since there are huge amounts of private/corporate information stored in SNs, they are potential targets of growing threats such as espionage, information leakage, identity theft, fraud, etc. Moreover, due to the facilitation of communications via SNSs, terrorist acts and organized crimes have moved into them for planning, meeting, recruiting and spreading their ideas. All in all, some solutions have been proposed for coping with security issues of SNs; however, they are not completely feasible. Thus, there is a need for post-incident mechanisms to track, analyze and detect abuses and misbehaviors. Such approaches are also useful for predicting and preventing future issues. In the field of computer science, Computer Forensics (CF) is responsible for performing such tasks. Digital Forensics (DF) as an extension of CF in the context of SNs is known as SNF. Since this topic is a new one in the community, many studies have not been devoted to it. Of course, over recent years, several methods have been proposed for SNA-related issues that have indirectly done the forensics tasks. Viewing from a general perspective to organize the current studies and as a framework for future researches, we classified SNF tasks to three major categories of analysis, detection and prediction. Then, the literature of this topic was reviewed based on the proposed classification. Since SNF is an emerging subject, there are many open issues about it that should be considered in future research activities.

As an ongoing direction, it is predictable that, for leveraging users' capabilities – as key players of web 2.0 and social web, collaborative SNF methods will be proposed. In fact, users could be the volunteer social police officers by their reports, experiences, investigations and alarms. Moreover, SNs service providers may implement preventive security features like “trusted friend suggestion” that include degree of trust (lack of misbehavior) as an influential factor in their (friend) suggestion algorithms.

## References

1. Hampton, K., et al.: Social networking sites and our lives. Technical report, Pew Internet & American Life Project (2011)
2. Rainie, L., Smith, A.: Politics on Social Networking Sites. press release, Pew Research Center (September 4, 2012), <http://www.pewinternet.org/Press-Releases/2012/Politics-on-Social-Networking-Sites.aspx>
3. Cheng, J., et al.: Forensics Tools for Social Network Security Solutions. In: Proceedings of Student-Faculty Research Day. CSIS, Pace University, NY, USA (2009)
4. Silva, M., et al.: Virtual Forensics: Social Network Security Solutions. In: Proceedings of Student-Faculty Research Day. CSIS, Pace University, NY, USA (2009)
5. O'Reilly, T.: What is Web 2.0-Design Patterns and Business Models for the Next Generation of Software (2005), <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html> (accessed February 2013)
6. Rimskii, V.: The influence of the Internet on active social involvement and the formation and development of identities. *Russian Social Science Review* 52(1), 79–101 (2011)
7. Edosomwan, S., et al.: The History of Social Media and its Impact on Business. *The Journal of Applied Management and Entrepreneurship* 16(3) (2011)
8. Boyd, D.M., Ellison, N.B.: Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13(1), article 11 (2007)
9. Coyle, C.L., Vaughn, H.: Social networking: Communication revolution or evolution? *Bell Labs Technical Journal* 13(2), 13–17 (2008)
10. Geierhos, M., Ebrahim, M.: Customer Interaction Management Goes Social: Getting Business Processes Plugged in Social Networks. In: *Computational Social Networks*, pp. 367–389. Springer, London (2012)
11. Costa, A.A., Tavares, L.V.: Social e-business and the Satellite Network model: Innovative concepts to improve collaboration in construction. *Automation in Construction* 22, 387–397 (2012)
12. Zhu, Z.: Discovering the influential users oriented to viral marketing based on online social networks. *Physica A: Statistical Mechanics and its Applications* 392(16), 3459–3469 (2013)
13. Domingo, M.C.: Managing Healthcare through Social Networks. *Computer* 43(7), 20–25 (2010)
14. Meltzer, D., et al.: Exploring the use of social network methods in designing healthcare quality improvement teams. *Social Science & Medicine* 71(6), 1119–1130 (2010)
15. Li, S., Hao, F., Li, M., Kim, H.-C.: Medicine Rating Prediction and Recommendation in Mobile Social Networks. In: Park, J.J.(J.H.), Arabnia, H.R., Kim, C., Shi, W., Gil, J.-M. (eds.) *GPC 2013. LNCS*, vol. 7861, pp. 216–223. Springer, Heidelberg (2013)

16. Greenhow, C., Schultz, K.: Using online social networks to support underrepresented students' engagement in postsecondary education. In: Chinn, C.A., et al. (eds.) Proceedings of the 8th International Conference on Computer Supported Collaborative Learning (CSCL 2007). International Society of the Learning Sciences, pp. 232–233 (2007)
17. Fardoun, H.M., Alhazzawi, D.M., López, S.R., Penichet, V.M.R., Gallud, J.A.: Online Social Networks Impact in Secondary Education. In: Vittorini, P., Gennari, R., Marenzi, I., de la Prieta, F., Rodríguez, J.M.C. (eds.) International Workshop on Evidence-Based TEL. AISC, vol. 152, pp. 37–45. Springer, Heidelberg (2012)
18. Miloslava, Ć., et al.: Utilization of learning management systems & social networking systems not only in the process of education. In: Mastorakis, N., et al. (eds.) Proceedings of the 10th WSEAS International Conference on Communications, Electrical & Computer Engineering, and 9th WSEAS International Conference on Applied Electromagnetics, Wireless and Optical Communications (ACELAE 2011), pp. 154–159. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point (2011)
19. Potts, J., et al.: Social network markets: a new definition of the creative industries. *Journal of Cultural Economics* 32(3), 167–185 (2008)
20. Dong, Y., et al.: TeleData: data mining, social network analysis and statistics analysis system based on cloud computing in telecommunication industry. In: Proceedings of the Third International Workshop on Cloud Data Management (CloudDB 2011), pp. 41–48. ACM, New York (2011)
21. Felzensztein, C., Gimmon, E.: Industrial clusters and social networking for enhancing inter-firm cooperation: the case of natural resources-based industries in Chile. *Journal of Business Market Management* 2(4), 187–202 (2008)
22. ter Maat, J.: Identifying centrality metrics for different types of social networks. In: The 18th Twente Student Conference, Enschede, Netherlands (2008), <http://referaat.cs.utwente.nl/conference/18/paper>
23. Takaffoli, M., et al.: A framework for analyzing dynamic social networks. In: 7th Conference on Applications of Social Network Analysis (ASNA 2010). University of Zurich (2010)
24. Dwyer, C., et al.: Trust and privacy concerns within social networking sites: A comparison of Facebook and MySpace. In: Proceedings of the 13th Americas Conference on Information Systems. Keystone, Colorado (2007)
25. Subrahmanyam, K., et al.: Online and offline social networks: Use of social networking sites by emerging adults. *Journal of Applied Developmental Psychology* 29(6), 420–433 (2008)
26. Toriumi, F., et al.: Classification of social network sites based on network indexes and communication patterns. In: Proceedings of International Workshop on Social Web Mining Co-located with The Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI) (2011)
27. Yammer, The Enterprise Social Network, <https://www.yammer.com/>
28. Conti, M., Hasani, A., Crispo, B.: Virtual private social networks. In: Proceedings of the first ACM Conference on Data and Application Security and Privacy (CODASPY 2011), pp. 39–50. ACM, New York (2011)
29. Allman, M.: On building special-purpose social networks for emergency communication. *ACM SIGCOMM Computer Communication Review* 40(5), 27–34 (2010)
30. <http://nucleusresearch.com/news/press-releases/facebook-costs-companies-1-dot-5-percent-of-total-productivity/> (accessed May 2013)



31. Isheriff.: Strategies for managing social networking and personal web use in the workplace. White paper. Isheriff Company, Costa Mesa (2010)
32. Rooksby, J., et al.: Social Networking and the Workplace. School of Computer Science, North Haugh, University of St Andrews (2009)
33. <http://www.dailymail.co.uk/femail/article-2080398/Facebook-cited-THIRD-divorces.html> (accessed March 2013)
34. <http://www.guardian.co.uk/technology/2011/mar/08/facebook-us-divorces> (accessed March 2013)
35. <http://nakedsecurity.sophos.com/2012/05/04/13-million-us-facebook-users-not-using-or-oblivious-to-privacy-controls/> (accessed May 2013)
36. <http://nakedsecurity.sophos.com/2012/08/02/fake-facebook-accounts> (accessed May 2013)
37. <http://www.infosecurity-magazine.com/view/23520/outhoover-hoover-fbi-wants-massive-datamining-capability-for-social-media/> (accessed May 2013)
38. Chen, Y.: Research on Social Media Network and National Security. In: Du, W. (ed.) Informatics and Management Science II. LNEE, vol. 205, pp. 593–599. Springer, Heidelberg (2013)
39. Abdulhamid, S., et al.: Privacy and National Security Issues in Social Networks: the Challenges. *International Journal of the Computer, the Internet and Management* 19(3), 14–20 (2011)
40. Hare, F., Goldstein, J.: The interdependent security problem in the defense industrial base: An agent-based model on a social network. *International Journal of Critical Infrastructure Protection* 3(3), 128–139 (2010)
41. Symantec Corporation: Internet Security Threat Report, 2012 Trends (18), 32 (2013)
42. Gao, H., et al.: Security issues in online social networks. *IEEE Internet Computing* 15(4), 56–63 (2011)
43. Brown, G., et al.: Social Networks and Context- Aware Spam. In: Proceedings of ACM Conference on Computer Supported Cooperative Work (CSCW 2008), pp. 403–412. ACM Press (2008)
44. Galam, S.: Modeling rumors: the no plane Pentagon French hoax case. *Physica A: Statistical Mechanics and Its Applications* 320, 571–580 (2003)
45. Kimmel, A.J.: Rumors and the financial marketplace. *The Journal of Behavioral Finance* 5(3), 134–141 (2004)
46. Nekovee, M., Moreno, Y., Bianconi, G., Marsili, M.: Theory of rumor spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications* 374(1), 457–470 (2007)
47. Karp, R., et al.: Randomized rumor spreading. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, pp. 565–574. IEEE (2000)
48. Sauerwald, T., Stauffer, A.: Rumor spreading and vertex expansion on regular graphs. In: Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 462–475. SIAM (2011)
49. Xu, W., Zhang, F., Zhu, S.: Toward Worm Detection in Online Social Networks. In: Proceedings 26th Annual Computer Security Applications Conference (ACSAC 2010), pp. 11–20. ACM Press (2010)
50. Schmugar, C.: The Future of Social Networking Sites. *McAfee Security Journal: Security Vision from McAfee Labs*, 28–30 (2008)

51. <http://www.net-security.org/secworld.php?id=10661> (accessed May 2013)
52. Zheleva, E., Getoor, L.: Privacy in social networks: A survey. In: Aggarwal, C.C. (ed.) *Social Network Data Analytics*, pp. 277–306. Springer, US (2011)
53. Beach, A., et al.: Solutions to security and privacy issues in mobile social networking. In: *Proceedings of International Conference on Computational Science and Engineering (CSE 2009)*, vol. 4, pp. 1036–1042. IEEE (2009)
54. Rodríguez, E., et al.: A Digital Rights Management approach to privacy in online social networks. In: *Workshop on Privacy and Protection in Web-based Social Networks (within ICAIL 2009)*, Barcelona (2009)
55. Johnson, M., Egelman, S., Bellovin, S.M.: Facebook and privacy: it's complicated. In: *Proceedings of the Eighth Symposium on Usable Privacy and Security*, Article 9. ACM (2012)
56. Sahinoglu, M., Akkaya, A.D., Ang, D.: Can We Assess and Monitor Privacy and Security Risk for Social Networks? *Procedia-Social and Behavioral Sciences* 57, 163–169 (2012)
57. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: *Proceedings of the 19th International Conference on World Wide Web*, pp. 351–360. ACM (2010)
58. Kafali, O., et al.: PROTOSS: A Run Time Tool for Detecting Privacy Violations in Online Social Networks. In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 429–433. IEEE Computer Society, Washington, DC (2012)
59. Sundsoy, P.R., et al.: The Activation of Core Social Networks in the Wake of the Oslo Bombing. In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 586–590. IEEE Computer Society, Washington, DC (2012)
60. Memon, N., Larsen, H.L., Hicks, D., Harkiolakis, N.: Detecting Hidden Hierarchy in Terrorist Networks: Some Case Studies. In: Yang, C.C., et al. (eds.) *ISI Workshops 2008*. LNCS, vol. 5075, pp. 477–489. Springer, Heidelberg (2008)
61. Memon, N., et al.: Understanding the structure of terrorist networks. *Int. J. Bus. Intell. Data Min.* 2(4), 401–425 (2007)
62. Morselli, C., Décary-Héту, D.: *Crime Facilitation Purposes of Social Networking Sites: A Review and Analysis of the "cyberbanging" Phenomenon*. Public Safety Canada, Ottawa (2010)
63. Womer, S., Bunker, R.J.: Surrños Gangs and Mexican Cartel Use of Social Networking Sites. *Small Wars & Insurgencies* 21(1), 81–94 (2010)
64. Palmer, G.: A road map for digital forensics research. Technical Report from the first Digital Forensics Research Workshop (DFRWS), Utica, New York (2001)
65. <http://www.mashable.com/2012/02/13/social-media-forensics> (accessed March 2013)
66. Zhao, H., et al.: Behavior modeling and forensics for multimedia social networks. *IEEE Signal Processing Magazine* 26(1), 118–139 (2009)
67. Lau, S.: *Uncovering Communications: A Forensic Look at Online Social Networking Communities*. Financial Fraud Law Report (2009)
68. Nagy, Z.: Social media risks from forensic point of view. *International Journal of Computers and Communications* 6(4), 245–253 (2012)
69. Huber, M., et al.: Social snapshots: Digital forensics for online social networks. In: *Proceedings of the 27th Annual Computer Security Applications Conference*, pp. 113–122. ACM (2011)

70. Zainudin, N.M., Merabti, M., Llewellyn-Jones, D.: A Digital Forensic Investigation Model for Online Social Networking. In: Proceedings of The 11th Annual Conference on the Convergence of Telecommunications, Networking & Broadcasting (PGNet 2010), Liverpool, UK, pp. 21–22 (2010)
71. Mulazzani, M., et al.: Social Network Forensics: Tapping the Data Pool of Social Network. In: Eighth Annual IFIP WG 11.9 International Conference on Digital Forensics. University of Pretoria, Pretoria (2012)
72. Castiglione, A., Cattaneo, G., De Santis, A.: A forensic analysis of images on online social networks. In: Proceedings of Third International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 679–684. IEEE (2011)
73. Araujo-Valdez, K., et al.: Social Network Forensic Tools. In: Proceedings of Student-Faculty Research Day. CSIS, Pace University, NY, USA (2012)
74. Al Mutawa, N., Baggili, I., Marrington, A.: Forensic analysis of social networking applications on mobile devices. *Digital Investigation* 9, 24–33 (2012)
75. Lau, R.Y.K., et al.: Social Media Analytics for Cyber Attack Forensic. *International Journal of Research in Engineering and Technology (IJRET)* 1(4), 217–220 (2012)
76. Chu, H.C., Deng, D.J., Park, J.H.: Live data mining concerning social networking forensics based on a Facebook session through aggregation of social data. *IEEE Journal on Selected Areas in Communications* 29(7), 1368–1376 (2011)
77. Son, J.: Social Network Forensics: Evidence Extraction Tool Capabilities. Masters Thesis, AUT University (2012)
78. Alzaidy, R.: Criminal Network Mining and Analysis for Forensic Investigations. Doctoral dissertation, Faculty of Engineering and Computer Science, Concordia University, Canada (2010)
79. Ressler, S.: Social network analysis as an approach to combat terrorism: past, present, and future research. *Homeland Security Affairs* 2(2), 1–10 (2006)
80. Svenson, P., Svensson, P., Tullberg, H.: Social Network Analysis And Information Fusion For Anti-Terrorism. In: Proceedings of the Conference on Civil and Military Readiness 2006 (CIMI 2006), Enköping, Stockholm, Sweden, paper S3.1 (2006)
81. Lu, Y., et al.: Social network analysis of a criminal hacker community. *Journal of Computer Information Systems* 51(2), 31–41 (2010)
82. Yang, C.C., Sageman, M.: Analysis of terrorist social networks with fractal views. *Journal of Information Science* 35(3), 299–320 (2009)
83. Kerschbaum, F., Schaad, A.: Privacy-preserving social network analysis for criminal investigations. In: Proceedings of the 7th ACM Workshop on Privacy in the Electronic Society, pp. 9–14. ACM (2008)
84. L’Huillier, G., et al.: Topic-based social network analysis for virtual communities of interests in the dark web. *SIGKDD Explor. Newsl.* 12(2), 66–73 (2011)
85. [http://www.nytimes.com/2012/01/14/us/federal-security-program-monitored-public-opinion.html?\\_r=0](http://www.nytimes.com/2012/01/14/us/federal-security-program-monitored-public-opinion.html?_r=0) (accessed March 2013)
86. <http://news.bbc.co.uk/2/hi/7962631.stm> (accessed May 2013)
87. Chakraborty, A., et al.: SPAM: A Framework for Social Profile Abuse Monitoring. Course Project, Department of Computer Science. Stony Brook University (2012), <http://www.cs.sunysb.edu/~aychakrabort/courses/cse508/>
88. Corley, C.D., et al.: Monitoring influenza trends through mining social media. In: International Conference on Bioinformatics & Computational Biology, pp. 340–346 (2009)
89. Patrick, A.S.: Monitoring corporate password sharing using social network analysis. In: International Sunbelt Social Network Conference. St. Pete Beach, Florida (2008)

90. Sindhu, K.K., Meshram, B.B.: Digital Forensics and Cyber Crime Datamining. *Journal of Information Security* 3(3), 196–201 (2012)
91. Brown, R., Pham, B., de Vel, O.: Design of a digital forensics image mining system. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) *KES 2005. LNCS (LNAI)*, vol. 3683, pp. 395–404. Springer, Heidelberg (2005)
92. Qin, L.: Data mining method based on computer forensics-based ID3 algorithm. In: *Proceedings of the 2nd IEEE International Conference on Information Management and Engineering (ICIME)*, pp. 340–343. IEEE (2010)
93. Reid, E., Qin, J., Chung, W., Xu, J., Zhou, Y., Schumaker, R., Sageman, M., Chen, H.: Terrorism knowledge discovery project: A knowledge discovery approach to addressing the threats of terrorism. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) *ISI 2004. LNCS*, vol. 3073, pp. 125–145. Springer, Heidelberg (2004)
94. Ozgul, F., Bondy, J., Aksoy, H.: Mining for offender group detection and story of a police operation. In: *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, vol. 70, pp. 189–193. Australian Computer Society, Inc. (2007)
95. Xu, J.J., Chen, H.: CrimeNet explorer: a framework for criminal network knowledge discovery. *ACM Transactions on Information Systems (TOIS)* 23(2), 201–226 (2005)
96. Chaurasia, N., et al.: A Survey on Terrorist Network Mining: Current Trends and Opportunities. *International Journal of Computer Science and Engineering Survey* 3(4) (2012)
97. Hosseinkhani, J., Chaprut, S., Taherdoost, H.: Criminal Network Mining by Web Structure and Content Mining. In: *Proceedings of the 11th WSEAS International Conference on Information Security and Privacy (ISP 2012)*, Prague, Czech Republic, pp. 210–215 (2012)
98. Nohuddin, P.N., et al.: Finding “interesting” trends in social networks using frequent pattern mining and self organizing maps. *Knowledge-Based Systems* 29, 104–113 (2012)
99. Maruhashi, K., Guo, F., Faloutsos, C.: MultiAspectForensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In: *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 203–210. IEEE (2011)
100. Brandes, U., Wagner, D.: Analysis and visualization of social networks. In: *Graph Drawing Software*, pp. 321–340. Springer, Heidelberg (2004)
101. Henry, N., Fekete, J.D., McGuffin, M.J.: NodeTrix: a hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics* 13(6), 1302–1309 (2007)
102. Boman, M., et al.: Social network visualization as a contact tracing tool. In: *Proceedings of the First International Workshop on Agent Technology for Disaster Management*, pp. 131–133 (2006)
103. Bosma, M., Meij, E., Weerkamp, W.: A framework for unsupervised spam detection in social networking sites. In: Baeza-Yates, R., de Vries, A.P., Zaragoza, H., Cambazoglu, B.B., Murdock, V., Lempel, R., Silvestri, F. (eds.) *ECIR 2012. LNCS*, vol. 7224, pp. 364–375. Springer, Heidelberg (2012)
104. Gao, H., et al.: Detecting and characterizing social spam campaigns. In: *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pp. 35–47. ACM (2010)
105. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: *Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 1–9. ACM (2010)
106. Wang, D., Irani, D., Pu, C.: A social-spam detection framework. In: *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, pp. 46–54. ACM (2011)

107. Huber, M., et al.: Exploiting social networking sites for spam. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 693–695 (2010)
108. Lumezanu, C., Feamster, N.: Observing common spam in Twitter and email. In: Proceedings of the 2012 ACM Conference on Internet Measurement Conference (IMC 2012), pp. 461–466. ACM, New York (2012)
109. Wang, A.H.: Detecting spam bots in online social networking sites: A machine learning approach. In: Foresti, S., Jajodia, S. (eds.) Data and Applications Security and Privacy XXIV. LNCS, vol. 6166, pp. 335–342. Springer, Heidelberg (2010)
110. Sharma, A., Panigrahi, P.K.: A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications* 39(1), 37–47 (2012)
111. Wiil, U.K., Memon, N., Karampelas, P.: Detecting new trends in terrorist networks. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 435–440. IEEE (2010)
112. Halalai, R., Lemnaru, C., Potolea, R.: Distributed community detection in social networks with genetic algorithms. In: Proceedings of International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 35–41. IEEE (2010)
113. Blondel, V.D., et al.: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (10), P10008 (2008)
114. Greene, D., Doyle, D., Cunningham, P.: Tracking the evolution of communities in dynamic social networks. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 176–183. IEEE (2010)
115. Altshuler, Y., Fire, M., Shmueli, E., Elovici, Y., Bruckstein, A., Pentland, A(S.), Lazer, D.: Detecting Anomalous Behaviors Using Structural Properties of Social Networks. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 433–440. Springer, Heidelberg (2013)
116. Heard, N.A., et al.: Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics* 4(2), 645–662 (2010)
117. Bilgin, C.C., Yener, B.: Dynamic Network Evolution: Models, Clustering, Anomaly detection. *IEEE Networks* (2006)
118. Gupta, N., Dey, L.: Detection and Characterization of Anomalous Entities in Social Communication Networks. In: Proceedings of 20th International Conference on Pattern Recognition (ICPR), pp. 738–741. IEEE (2010)
119. Chen, Z., Hendrix, W., Samatova, N.F.: Community-based anomaly detection in evolutionary networks. *Journal of Intelligent Information Systems* 39(1), 59–85 (2012)
120. Krombholz, K., Merkl, D., Weippl, E.: Fake identities in social media: A case study on the sustainability of the Facebook business model. *Journal of Service Science Research* 4(2), 175–212 (2012)
121. Dinakar, K., et al.: Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2(3), Article 18 (2012)
122. Ying, X., Wu, X., Barbará, D.: Spectrum based fraud detection in social networks. In: Proceedings of 27th International Conference on Data Engineering (ICDE), pp. 912–923. IEEE (2011)
123. Cetintas, S., et al.: Identifying similar people in professional social networks with discriminative probabilistic models. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1209–1210. ACM (2011)

124. Liu, X., Troney, R., Huet, B.: Using social media to identify events. In: Proceedings of the 3rd ACM SIGMM International Workshop on Social Media, pp. 3–8. ACM (2011)
125. Saraf, P., et al.: Social Media Analysis and Geospatial Crime Report Clustering for Crime Prediction & Prevention. Data Analytics Course Project, Department of Computer Science. Virginia Tech University (2011)
126. Uchida, C.D.: Predictive Policing in Los Angeles: Planning and Development. Justice & Security Strategies, Inc. (2009)
127. <http://rt.com/news/software-tracks-predicts-raytheon-878/> (accessed June 2013)
128. Liben - Nowell, D., Kleinberg, J.: The link - prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58(7), 1019–1031 (2007)
129. Chaurasia, N., et al.: Exploring the Current Trends and Future Prospects in Terrorist Network Mining. In: Wyld, C.D., et al. (eds.) Proceedings of The Second International Conference on Computer Science, Engineering and Applications (CCSEA 2012), Delhi, India, vol. 2(2) (2012)
130. Grover, V., Adderley, R., Bramer, M.: Review of current crime prediction techniques. In: Springer London, Applications and Innovations in Intelligent Systems XIV, pp. 233–237 (2007)

# Impact of Some Biometric Modalities on Forensic Science

Ali Ismail Awad<sup>1</sup> and About Ella Hassanien<sup>2</sup>

<sup>1</sup> Faculty of Engineering, Al Azhar University, Qena, Egypt  
Member of the Scientific Research Group in Egypt (SRGE)

aawad@ieee.org

<sup>2</sup> Faculty of Computers & Information,  
Cairo University, Cairo, Egypt

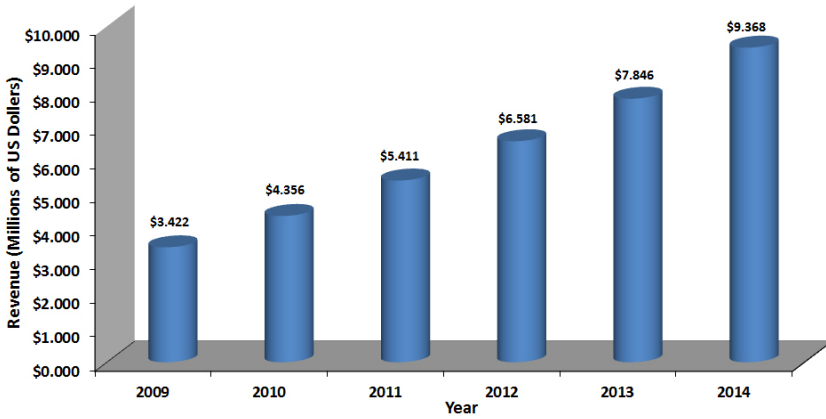
Chairman of Scientific Research Group in Egypt (SRGE)

aboitcairo@gmail.com

**Abstract.** Recently, forensic science has had many challenges in many different types of crimes and crime scenes, vary from physical crimes to cyber or computer crimes. Accurate and efficient human identification or recognition have become crucial for forensic applications due to the large diversity of crime scenes, and because of the increasing need to accurately identify criminals from the available crime evidences. Biometrics is an emerging technology that provides accurate and highly secure personal identification and verification systems for civilian and forensic applications. The positive impact of biometric modalities on forensic science began with the rapid developments in computer science, computational intelligence, and computing approaches. These advancements have been reflected in the biometric modality capturing process, feature extraction, feature robustness, and features matching. A complete and automatic biometric identification or recognition systems have been built accordingly. This chapter presents a study of the impacts of using some biometric modalities in forensic applications. Although biometrics identification replaces human work with computerized and automatic systems in order to achieve better performance, new challenges have arisen. These challenges lie in biometric system reliability and accuracy, system response time, data mining and classification, and protecting user privacy. This chapter sheds light on the positive and the negative impacts of using some biometric modalities in forensic science. In particular, the impacts of fingerprint image, facial image, and iris patterns are considered. The selected modalities are covered preliminarily before tackling their impact on forensic applications. Furthermore, an extensive look at the future of biometric modalities deployment in forensic applications is covered as the last part of the chapter.

## 1 Introduction

Undoubtedly, the unprecedented size of the global human population, modern networked society, and computerized transactions are crucial to contemporary



**Fig. 1.** Biometrics technology revenue from 2009–2014, measured in millions of U.S. Dollars. The figure is redrawn according to the data provided in [11].

human civilization. On the other hand, they open a door to wide a diversity of crimes, including both physical and computer crimes. Driven by the aforementioned factors, forensic science has become a modern necessity. Forensic science is understood generally as the use of some scientific knowledge to solve crimes (or other legal problems), and more precisely as a scientific analysis of physical evidence from a crime scene [1]. Forensic science faces plenty of challenges in terms of criminal detection accuracy, processing efficiency, and productivity. Given the recent advances in computing approaches, computational intelligence techniques, and the large storage facilities, replacing manual forensic techniques or applications with computerized systems has become a vital requirement.

Biometrics technology is key fundamental security mechanism that assigns a unique identity to an individual according to some physiological or behavioral features [2], [3], [4]. These features are sometimes called as biometric modalities, identifiers, traits, or characteristics. Extensive biometric identifiers are grouped into biological traits (e.g., DNA, EEG analysis, and ECG analysis), behavioral traits (e.g., signature dynamic, human gait, and voice signal) and morphological traits (e.g., fingerprints, facial image, and iris patterns) [5], [6], continuing the development of biometric-based human identification begun by Herschel in 1858 [7]. Additionally, soft biometric characteristics such as skin color and body length could be used for coarse-level suspect classification processes [8]. Due to its related civilian and forensic deployments, biometrics technology is presently attracting researchers from private and academic institutions [9], [10]. Fig. 1 represents the amount of expected investment in biometrics technology during the period from 2009 to 2014 measured in millions of U.S. Dollars [11].

Obviously, there is a link between biometrics technology and forensic science in the processing of human modalities as identifiers in biometrics-based identification or recognition systems, and as criminal evidence in forensic applications. Some governmental authorities have already started using some human



biometric traits for accurate and efficient forensics examinations [8], [12]. Although using human identifiers as biometric traits is a cooperative process, it presents a challenge as a forensic evidence since the criminals obviously do not want to be convicted [1].

Recent advancements in computing technology, parallel processing, computational intelligence, and processing power open doors for an extensive deployment of biometric modalities in forensic applications. Undoubtedly, these advances have had a positive impact on biometrics technology itself, not only by reducing the processing time, but also by enhancing the accuracy and the reliability of the biometrics-based identification systems [13]. Therefore, the usage of biometric modalities as forensic evidence has accordingly become consolidated. It is worth noting that the advances and the challenges of biometric modalities will be reflected on the forensic applications when both are coupled.

This chapter discusses the link between some biometric modalities and forensic science, and emphasizes the impact of using biometric identifiers in forensic investigations. The contribution of this chapter is twofold. First, it provides an extensive review of biometrics technology, generic biometrics identification systems, and the foremost biometric modalities in forensic applications. Second, it articulates the current challenges of using some biometric modalities in forensic science. Furthermore, this chapter highlights the positive and the negative impacts of coupling biometric technology with crime investigation. The outcomes of this chapter provide help for overcoming some challenges in forensic science, and provide a basis or further conducted researches linking both biometrics technology and forensic science.

The rest of this chapter is organized as follows. Section 2 explores biometrics technology, and covers the automatic system for personal identification. Section 3 explains in detail the foremost biometric modalities that are most widely used in forensic applications, including fingerprints, facial imagery, and iris patterns. Section 4 covers the mutual link between biometric modalities and forensic applications, and emphasises the impact of these modalities on forensic investigations. Section 5 describes our vision of the future union between biometrics technology and forensic science. Finally, research conclusions are reported in Section 6.

## 2 Biometrics Technology

Biometric modalities provide a high security level with preserved accuracy and reliability for the automated identification and verification systems. Biometrics-based identification systems compensate some weaknesses of traditional token- and knowledge-based identification approaches by replacing “something you possess” or “something you know” with “something you are” [14], [15]. It offers not only an automatic method of identification, but also a convenience for users as they do not have to remember information or carry a possession [6]. Driven by its merits, biometrics technology deployment is keeping apace with high industrial revenue and investments, though it is also becoming a fundamental technology for future personal, mobile, and governmental applications

**Table 1.** Comparison of different biometric identifiers: 1 = High, 0.5 = Medium, and 0 = Low\*

	Universality	Uniqueness	Performance	Acceptability	Circumvention	Score
Fingerprints	0.5	1.0	1.0	0.5	1.0	<b>4.0</b>
Facial image	1.0	0.0	0.0	1.0	0.0	2.0
Iris patterns	1.0	1.0	1.0	0.0	1.0	<b>4.0</b>
DNA	1.0	1.0	1.0	0.0	0.0	<b>3.0</b>
EEG	1.0	0.0	0.0	0.0	0.0	1.0
Signature	0.0	0.0	0.0	1.0	0.0	1.0
Voice	0.5	0.0	0.0	1.0	0.0	1.5
Gait	0.5	0.0	0.0	1.0	0.5	2.0

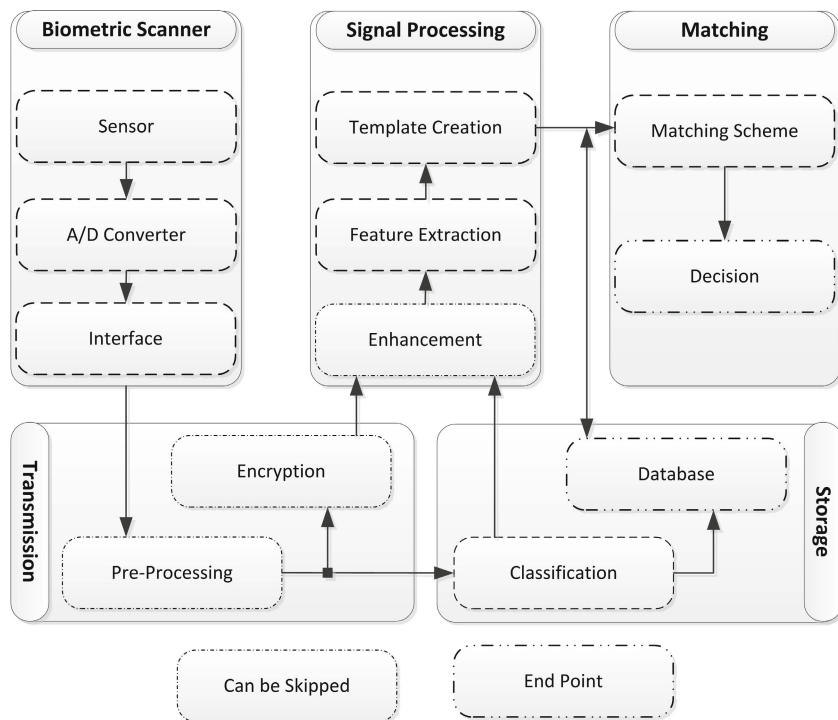
\*The table is adopted and updated from [17], [20].

[6], [11]. See Fig. 1 for the projected investments in the current and future biometrics technology deployments.

Due to the enormous needs for biometrics deployment in civilian and forensic applications, a large number of biometric traits have been discovered by taking advantages of the comprehensive understanding of the human body [16]. A qualified biometric trait must be investigated and filtered through selection criteria. The candidate biometric identifiers should achieve some technical and operational requirements according to the type of application. The competency requirements might be summarized as [10], [17], [18], [19]:

- Acceptability, measuring to what extent the user may accept the biometric trait in terms of acquisition, data representation, and user privacy. User acceptability is determined according to the application obtrusiveness and intrusiveness which are subjected to user agreement.
- Circumvention, is an important parameter that affects the reliability of the system. It refers to how it is easy to fool a system by fraudulent means. According to Table 1, the higher the circumvention value, the better and more suitable the biometric identifier.
- Performance, which refers to achievable identification criteria (such as accuracy, speed, and robustness), as well as to the resources required to achieve an acceptable identification performance.
- Uniqueness, indicating that the selected identifier should contain enough features to differentiate between two persons carrying the same trait. Moreover, the identifier should be immutable over time.
- Universality, such that the selected trait must be available in everyone, and can be measured quantitatively without affecting user privacy or user health.

The research result reported in [17], [20] demonstrates the performance evaluation of individual biometric modalities based on the above qualification criteria, and its impact on coupling the suitable identifier to the appropriate application. Table 1 then compares these biometric modalities according to the

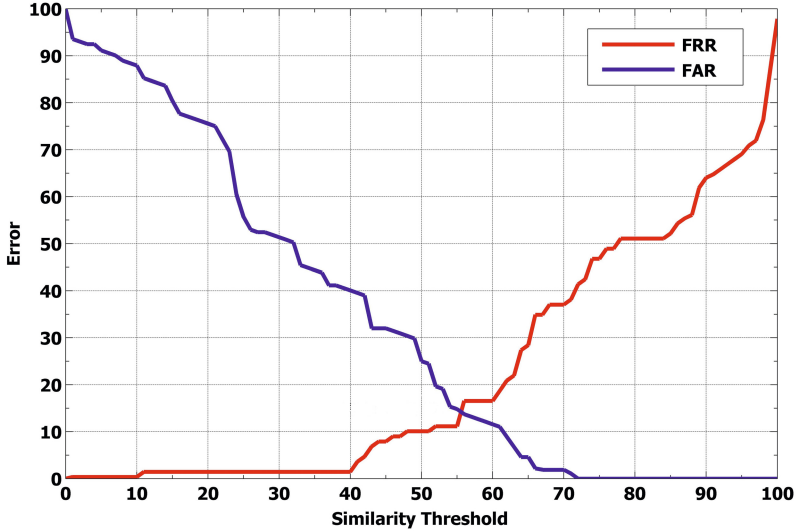


**Fig. 2.** A generic architecture of the Automated Biometrics Identification System (ABIS) from a signal processing stand point. The figure is adopted from [18].

above-mentioned criteria. The reported results indicate the superiority of fingerprints compared to the other biometric identifiers as fingerprints achieve a very high comparison score. Iris patterns come in second due to an acceptability problem (low acceptability score). DNA comes in third due to acceptability and circumvention limitations. It is worth noting that no single biometric identifier can achieve excellent performance for all requirements as explained above [21]. Therefore, two or more traits can be combined to achieve stronger security levels [22], [23], [24]. Fingerprints, facial image, and iris patterns are discussed further as three dominant biometric identifiers, and their impact on forensic science will be demonstrated [9].

## 2.1 Biometrics Identification Systems

Automated Biometrics Identification Systems (ABIS) have replaced human experts in human recognition by a computerized approach. An ABIS consists of two major phases: (i) the enrollment phase, and (ii) the identification phase. During the enrollment phase, identities of individuals are registered in a database for future template matching and identity assignment. Assigning an identity to an



**Fig. 3.** False Acceptance Rate (FAR) and False Rejection Rate (FRR) plotted versus the similarity threshold. The Equal Error Rate (EER) is shown as a cross point between FAR and FRR.

individual person is then performed in the identification phase through presenting the user with a biometric sample [25]. Signal processing analysis provides a comprehensive information about the most of ABIS components, including identifier sensing, transmission to the processing machine, identifier processing, identifier classification, features storage as features template, and template matching. Fig. 2 shows a generic ABIS system architecture from a signal processing point of view [18]. Although, the figure shows a generic and complete ABIS, some components may be removed (skipped) from the processing sequence according to the application requirements without affecting the overall system performance.

An ABIS suffers from an abundance of error sources that deteriorate the performance of the system. Errors are found in: (i) the sensor level, (ii) the processing level, (iii) the enrollment phase, and (iv) the matching phase. A sensor-level error occurs when a biometric sample is presented to the sensor, but the sensor fails to detect its presence due to a hardware problem, which is commonly known as a Failure to Detect (**FtD**) error. If the sensor succeeds in detecting the presence of the biometric sample, but fails to capture it due to user misbehavior, this is defined as a Failure to Capture (**FtC**) error. A noisy and unclear biometric sample leads to failure in feature extraction, known as a Failure to Process (**FtP**) error. These three error types can all be categorized under one major error, known as a Failure to Acquire (**FtA**) error [6], [10].

On the other side, the behavior of the system matcher has a large impact on the system's performance. The matcher can produce two types of errors due to inter-user similarity and intra-user variations factors [3]. These two errors are called False Match Rate (FMR) and False Non-Match Rate (FNMR), they

also known as False Acceptance Rate (FAR) and False Rejection Rate (FRR), respectively. FAR and FRR are common notions for measuring the performance of a verification systems [10]. Fig. 3 shows the FAR and FRR error rates plotted against the similarity threshold between the input biometric sample and the registered template.

Mathematically speaking, suppose one biometric template is donated by  $T$ , and one presented sample (input) is donated by  $I$ . The similarity score ( $s$ ), between the template and the input, is measured by the function  $M(I, T) = s$ . The identity assignment decision is made according to the selected value of the similarity threshold ( $h$ ) [6].

**FMR** is the rate that the decision is made that  $I$  matches  $T$ , while in fact  $I$  and  $T$  come from two different individuals [6].

$$FMR(h) = 1 - \int_{s=h}^{\infty} p_n(s) ds \quad (1)$$

where  $p_n(s)$  is the non-match distribution between two samples as a function of the similarity score  $s$ .

**FNMR** is the rate that the decision is made that  $I$  does not match  $T$ , while  $I$  and  $T$  do in fact come from the same individual [6].

$$FNMR(h) = 1 - \int_{s=-\infty}^h p_m(s) ds \quad (2)$$

where  $p_m(s)$  is the match distribution between two samples as a function of the score  $s$ .

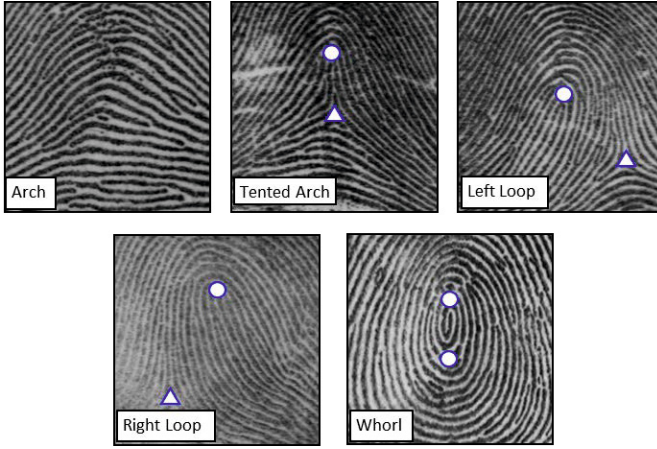
The Equal Error Rate (**EER**) is defined as the value of FMR and FNMR at the point of the threshold ( $h$ ) where the two error rates are identical [6], [13].

$$EER = FMR_{h=EER} = FNMR_{h=EER}. \quad (3)$$

The similarity threshold ( $h$ ) should be picked carefully in the system design phase and according to the security level and the system sensitivity. The similarity threshold should achieve a tradeoff between FMR and FNMR errors. See Fig. 3 for more considerations related to the selection of the system threshold.

### 3 Foremost Biometric Modalities

This section explains the foremost biometric modalities in biometric-based identification and verification systems and in some forensic applications. While iris patterns deployment still receives great research attentions, it is widely used in border and immigration controls. It also provides accurate results for identifying suspects from the collected data from previously recorded or well known criminals. Therefore, we believe that it is important to consider this perspective, and to include iris patterns as one the dominant biometric identifiers in forensic applications. Moreover, facial image provides a faster identification response with lower cost than DNA. Therefore, it is included as a foremost biometric modality.



**Fig. 4.** A sample of fingerprint images represents the fingerprint structures, the singular points, and the most famous five classes

### 3.1 Fingerprints

The skin structure on the palm and the fingers of a human hand in addition to the skin structure on the sole and the toes of the foot have a unique property of being completely individual, which allows them to be used for human identification [26]. An example of fingerprint patterns are shown in Fig. 4. Fingerprints have been used for identification purposes long time ago due to the well understanding of their biological properties and skin structure. Since the beginning of the 20<sup>th</sup> century, fingerprints have been extensively used for manual identification of criminals by various forensics around the world [10], [27].

A fingerprint is concisely defined as a smooth patterns of alternating ridges and valleys constructed on the finger tip. While, fingerprint ridges and furrows are treated as parallel in most fingerprint regions, a plenty of other altering features such as scars, cuts bruises, cracks, and calluses can also be found as a part of the fingerprint structure. These features can also be invested as discriminating factors. Generally, fingerprint structures are grouped into: (i) global, (ii) local, and (iii) low-level structure [10], [17], [26].

Global fingerprint architecture represents the overall structure of the finger. A unique representation is valid for the whole fingerprint, and it is typically determined by an investigation of the entire finger. The most important feature extracted from the global structure is a set of singular points, circles for cores and triangles for deltas shown in Fig. 4 [28]. Singular points are compound with the whole fingerprint structure to produce the distinct fingerprint classes [29], [30]. Local fingerprint architecture, the ridge level structure, is based on analyzing both ridge endings and ridge bifurcations, which maps the global representation into a minutiae structure that is used in most biometric identification or recognition systems [28], [31], [32]. The low-level of fingerprint structure

expresses some hidden features such as sweat pores distributed on the finger tip. The low-level features are sometimes inspected manually when comparing two fingerprint images. However, these features are rarely used in computerized or automated systems due to the high cost required for high-resolution sensors which are needed for a reliable feature extraction process [27], [33].

In spite of the great research efforts related to fingerprint identification systems, there are still some challenges that deteriorate a system's performance. One of the current challenges is identification time. Due to the high demands for fingerprint deployments, fingerprint databases are supposed to contain a huge number of enrolled users. The identification process (or 1:N matching) searches for a person's identity inside the database with size  $N$ . In a large-scale identification deployments, the database size becomes larger, and identification time correspondingly becomes much longer. Fingerprint classification is an available solution that is based on limiting the database search process into sub-classes, shown in Fig. 4. Moreover, parallel processing techniques achieve extremely low processing times based on hardware rendering techniques [34], [35], [36].

Latent fingerprints play an important role in forensic science. It is the most important evidence that may be tracked from a crime scene and compared against extremely large fingerprint databases. A latent fingerprint image is extracted from any surface that a criminal touched using special materials. Its resolution is therefore different from live-scanned fingerprints, and a special care must be given to enhance its quality prior to any feature extraction and matching operations [10].

### 3.2 Facial Image

The image of human face is used intuitively by human being to recognize each other. It is considered as the common method of human recognition in automatic or manual manner [20]. However, considering face image as a biometric identifier with a relatively lower score (see Table 1), it has high acceptability and universality characteristics that make it useful as forensic evidence. Recently, face recognition has become one of the most important applications of image analysis in forensics due to its availability in crime scenes, and the value of the evidence that the captured face image may carry [37], [38].

Automatic facial recognition is conducted in two ways: (i) using the global representation of the full face image, (ii) using the location and shape of facial components such as eyes, nose, lips, and their spatial connections [20], [39]. In forensic applications, using a human face as crime evidence involves many challenges due to the face rigidity, variations in facial pose, facial expression, and illumination condition under different emotions that needs special processing algorithms. Additionally, the difficulty in handling different photometric conditions, such as image quality and image resolution, deteriorates the performance of a facial recognition system. Thus, robust facial recognition is both attractive prospect and a challenging problem [3], [37]. Fig. 5 shows a sample of the face images of an individual person under various emotions, facial poses, and facial expressions.



**Fig. 5.** A sample of face images of an individual under different conditions, including emotions, facial poses, and facial expressions. The image is adopted from [41].

In order to reliably use facial images in forensic investigation, three main issues must be considered a priori. These issues are: (i) facial image immutability over time, (ii) facial image retrieval, and (iii) matching forensic (sketched) images with real captured facial images [40]. The effect of aging is also a great challenge for facial recognition deployment. Plenty of research focuses on mitigating the aging effect problem for reliable recognition systems [38]. While, other techniques have been used for addressing other concerns like the face occlusion problem [41].

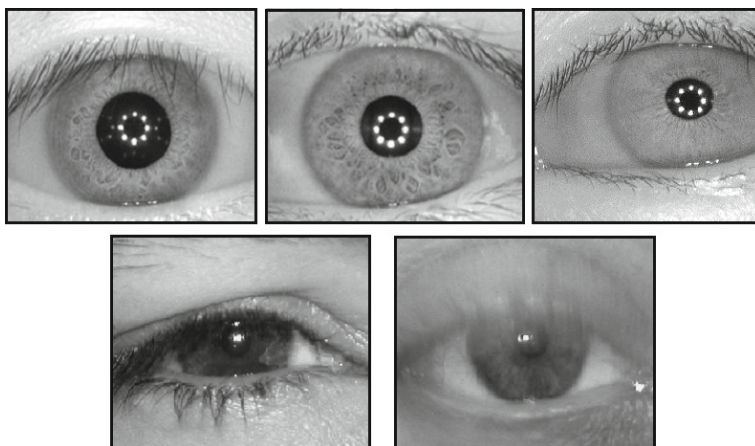
### 3.3 Iris Patterns

Human irises are a valuable source of information that is immutable over time. It has many distinctive features such as crypts, cornea, furrows, and rings. Daugman first presented an accurate and reliable personal recognition using iris patterns with a 2D Gabor filter [42] in order to modulate the iris phase information and construct the iris features code [43]. Subsequently, several approaches have been developed for iris-based individual recognition. These approaches are divided into four categories: phase-based approaches, texture analysis approaches, zero-crossing approach, and intensity variation analysis approaches [44].

The low user acceptability of iris recognition means that iris capturing is becoming a difficult task. Due to user misbehavior, the captured irises suffer mainly from two problems as: blurred irises and irises occluded by eyelids or eyelashes [44]. Problematic iris images are shown in Fig. 6. The research reported in [45] presents an iris quality assessment approach based on the combination of occlusion score and dilation score as two selective iris features. The link between the quality score and the recognition accuracy has been studied in [45]. A contactless iris capturing approach may be considered as an affordable solution to the above mentioned problems [46].

Recently, the immutability of iris features through a person's lifetime has come into question. Some changes in iris texture appearance occur with age, disease, and medication, which lead to deterioration or failure in iris-based recognition systems [47]. The reported information in [48] has refuted the previous claim





**Fig. 6.** Normal iris images taken from Chinese Academy of Sciences (CASIA) iris images databases [44]. Good iris images are shown in the top row. Problematic iris images, blurred and occluded images, are shown in the bottom row.

based upon the weakness of the evaluation algorithm, and the lack of presenting a photographic evidence of iris texture change over time. As a continuation of the scientific sparring, some remarks on the refutations in [48] have been presented in [49]. One new direction of individual identification and authentication is to use saccadic eye movement as a new biometric modality [50].

## 4 Biometric Modalities and Forensics

Because of the huge progress in computing, hardware, and storage media, traditional identification techniques have been supplemented by computerized and semi-automatic ones. The common area between biometrics and forensics is the automated investigation of the captured evidence (trace) from a crime scene [51]. This section reports some impacts of using biometric systems in general, and the previously explained modalities in particular.

Of course, the deployment of biometric modalities in forensic applications has two types of impact. On one hand, it leverages the proven performance of an ABIS with respect to time efficiency, especially with large databases, and identification accuracy, especially with degraded biometric input samples. On the other hand, it imposes a risk of violating user privacy. Moreover, special techniques are needed for protecting the template database. Additional approaches are also needed to protect the overall system from criminals' hacking and attacking.

The strong point of biometric systems lies in their ability to operate in identification and verification modes [10]. The identification mode in forensics works analogously to score-based biometric systems. The trace of the suspicious person is compared/matched against a group of suspect models, (1:N) matching. A verification mode implies that the trace of the suspicious individual is compared to a

claimed identity, and according to a similarity threshold, it is decided whether or not the trace belongs the claimed identity [17], [52]. Coupling biometric systems with forensic applications reflects the strength of biometric systems in forensic applications with respect to accuracy and precision.

While biometric systems can be deployed in forensic applications, traditional score-based decision biometric systems have not been working well in forensics, where a Likelihood Ratio (LR) must be used [52]. Score to likelihood ratio methods, and calibration methods, such as Kernel Density Estimation (KDE) and Logistic Regression (Log Reg), must be used to calibrate the biometric system score prior to any forensic usage [53]. Individualization is a common notion in forensic applications as the criminologist is interested in knowing of the source of the biometric data instead of just the similarity score between two biometric samples [12].

Biometric identifiers are time immutable, (see Section 2). This property may be another strong point for the biometric systems as the presented biometric sample is time invariant. On the other side, time immutability involves some negatives for forensic science. If the biometric database is hacked, then the identity of all registered individuals is at stake. Moreover, there is a possibility for matching crime evidence with fraudulent biometric templates that badly affect system performance and accuracy. Cancellable biometric techniques have been proposed to protect biometric databases, and hence, protecting individual identity and privacy [3].

Generally speaking, using biometric systems in forensic science has plenty of positive results on the performance of forensic applications. These results include greater system accuracy and reliability, shorter processing time, and much more productivity. The down side of coupling biometric modalities and forensic science lies in stripping the user privacy.

## 5 Future Vision

Biometrics technology is considered as an emerging approach for personal identification and recognition in both civilian and forensic applications. Biometric technology is used in forensic applications not only for suspects individualization, but also for preventing most cyber crimes by denying access to sensitive data. Tackling the current challenges of biometrics technology deployment in forensic applications will be an interesting research direction. Furthermore, biometric systems calibration techniques will receive great attention for harmonizing biometric systems with forensic applications.

Nowadays, new biometric identifiers such as human gait and human ear are being investigated for forensic science deployment [7]. Due to their biometric properties, (see Section 2), unique identity is assigned to each individual; they are therefore good potential identifiers for suspect individualization. Additionally, they are considered as valuable evidences that can be traced from a crime scene.

Artificial intelligence techniques such as Artificial Neural Networks (ANN), and Genetic Algorithms (GA), and Support Vector Machines (SVM) [54], [55]

play an important role in presenting non-traditional solutions for the challenges of biometrics technology, especially reducing the processing time and increasing the system's accuracy. The common idea behind these techniques is to build a feature vector and train (learn) a machine how to process that vector according to particular rules. Thus, machine learning techniques can efficiently process complicated biometrics data, and hence reflect these enhancements in the performance of forensic applications [36].

In order to alleviate the problem of biometric system processing time, a result of processing extremely large databases, the effect of advanced computing approaches and parallel processing techniques [35] on biometric systems will need to be brought in focus for future investigation. Furthermore, we expect more deployments of computational intelligence techniques on biometric identifiers for enhancing a system's performance in forensic applications [36].

## 6 Conclusions

Developing a reliable and productive forensic application is imperative for identifying and convicting criminals. The deployment of biometrics technology, such as fingerprints and facial images, as well as more recent developments in identifying human gait and ears, achieves greater accuracy and reliability than current forensic applications. Biometric systems need to be adopted from score-based decisions into likelihood ratio decisions for reliable forensic system performance.

This chapter has presented a study of the impact of using generic biometric systems, specifically fingerprints, facial imagery, and iris patterns, in forensic science. The chapter started with a common review of biometric technologies, the requirements for selecting biometric identifiers, and the Automatic Biometric Identification System (ABIS) components and errors. Subsequently, some biometric modalities were explained from a forensic applications perspective.

The deployment of biometric systems provides an automatic and convenient way to investigate a crime scene. The accuracy and the reliability of a biometric system will be reflected in the forensic application. Moreover, the performance of the forensic system will be enhanced in terms of evidence tracking, processing time, and individualization accuracy. On the other hand, this deployment involves drawbacks in terms of protecting the database and preserving the privacy of the individual.

## References

1. Yan, Y., Osadciw, L.A.: Bridging biometrics and forensics. In: Proceedings of SPIE 6819, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, pp. 6819Q–6819Q–8 (February 2008)
2. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 14(1), 4–20 (2004)
3. Jain, A.K., Ross, A.A., Nandakumar, K.: *Introduction to Biometrics*, 1st edn. Springer (2011)

4. Giot, R., El-Abed, M., Rosenberger, C.: Fast computation of the performance evaluation of biometric systems: Application to multibiometrics. *Future Generation Computer Systems* 29(3), 788–799 (2013), Special Section: Recent Developments in High Performance Computing and Security
5. Odinaka, I., Lai, P.H., Kaplan, A.D., O’Sullivan, J.A., Sirevaag, E.J., Rohrbaugh, J.W.: ECG biometric recognition: A comparative analysis. *IEEE Transactions on Information Forensics and Security* 7(6), 1812–1824 (2012)
6. Schouten, B., Jacobs, B.: Biometrics and their use in e-passports. *Image and Vision Computing* 27(3), 305–312 (2009), Special Issue on Multimodal Biometrics
7. Nixon, M.S., Bouchrika, I., Arbab-Zavar, B., Carter, J.N.: On use of biometrics in forensics: Gait and ear. In: *European Signal Processing Conference* (August 2010)
8. Spaun, N.A.: Forensic biometrics from images and video at the federal bureau of investigation. In: *First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2007)*, pp. 1–3 (2007)
9. Goudelis, G., Tefas, A., Pitas, I.: Emerging biometric modalities: A survey. *Journal on Multimodal User Interfaces* 2(3-4), 217–235 (2008)
10. Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: *Handbook of Fingerprint Recognition*, 2nd edn. Springer (2009)
11. International Biometric Group: Biometrics market and industry report 2009-2014 (March 2008), <http://www.biometricgroup.com>
12. Meuwly, D.: Forensic individualisation from biometric data. *Science & Justice* 46(4), 205–213 (2006)
13. Egawa, S., Awad, A.I., Baba, K.: Evaluation of acceleration algorithm for biometric identification. In: Benlamri, R. (ed.) *NDT 2012, Part II. CCIS*, vol. 294, pp. 231–242. Springer, Heidelberg (2012)
14. Ratha, N.K., Connell, J.H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal* 40(3), 614–634 (2001)
15. Lee, Y., Filliben, J.J., Micheals, R.J., Phillips, P.J.: Sensitivity analysis for biometric systems: A methodology based on orthogonal experiment designs. *Computer Vision and Image Understanding* 117(5), 532–550 (2013)
16. Li, Y.: Biometric technology overview. *Nuclear Science and Techniques* 17(2), 97–105 (2006)
17. Jain, A.K., Bolle, R., Pankanti, S. (eds.): *Biometrics: Personal Identification in Networked Society*, 2nd edn. Springer (2005)
18. Luis-Garcia, R.D., Alberola-Lopez, C., Aghzout, O., Ruiz-Alzola, J.: Biometric identification systems. *Signal Processing* 83(12), 2539–2557 (2003)
19. Awad, A.I., Hassanien, A.E., Zawbaa, H.M.: A cattle identification approach using live captured muzzle print images. In: Awad, A.I., Hassanien, A.E., Baba, K. (eds.) *SecNet 2013. CCIS*, vol. 381, pp. 143–152. Springer, Heidelberg (2013)
20. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: A tool for information security. *IEEE Transactions on Information Forensics and Security* 1(2), 125–143 (2006)
21. Toledano, D., Fernandezpozo, R., Hernandeztrapote, A., Hernandezgomez, L.: Usability evaluation of multi-modal biometric verification systems. *Interacting with Computers* 18(5), 1101–1122 (2006)
22. Islam, S., Davies, R., Bennamoun, M., Owens, R., Mian, A.: Multibiometric human recognition using 3D ear and face features. *Pattern Recognition* 46(3), 613–627 (2013)
23. Tresadern, P., Cootes, T.F., Poh, N., Matejka, P., Hadid, A., Levy, C., McCool, C., Marcel, S.: Mobile biometrics: Combined face and voice verification for a mobile platform. *IEEE Pervasive Computing* 12(1), 79–87 (2013)

24. Yang, K., Du, E.Y., Zhou, Z.: Consent biometrics. *Neurocomputing* 100(0), 153–162 (2013)
25. Jain, A.K., Nandakumar, K.: Biometric authentication: System security and user privacy. *Computer* 45(11), 87–92 (2012)
26. Lee, H., Gaensslen, R.: *Advances in Fingerprint Technology*, 2nd edn. CRC Series in Forensic and Police Science. Taylor & Francis (2010)
27. Yager, N., Amin, A.: Fingerprint verification based on minutiae features: a review. *Pattern Analysis & Applications* 7(1), 94–113 (2004)
28. Maltoni, D., Cappelli, R.: Advances in fingerprint modeling. *Image and Vision Computing* 27(3), 258–268 (2009)
29. Yager, N., Amin, A.: Fingerprint classification: a review. *Pattern Analysis & Applications* 7(1), 77–93 (2004)
30. Awad, A.I., Baba, K.: Fingerprint singularity detection: A comparative study. In: Mohamad Zain, J., Wan Mohd, W.M.b., El-Qawasmeh, E. (eds.) ICSECS 2011, Part I. CCIS, vol. 179, pp. 122–132. Springer, Heidelberg (2011)
31. Bolle, R.M., Senior, A.W., Ratha, N.K., Pankanti, S.: Fingerprint minutiae: A constructive definition. In: Tistarelli, M., Bigun, J., Jain, A.K. (eds.) ECCV 2002. LNCS, vol. 2359, pp. 58–66. Springer, Heidelberg (2002)
32. Muñoz-Briseño, A., Alonso, A.G., Palancar, J.H.: Fingerprint indexing with bad quality areas. *Expert Systems with Applications* 40(5), 1839–1846 (2013)
33. Jain, A.K., Chen, Y., Demirkus, M.: Pores and ridges: High-Resolution fingerprint matching using level 3 features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(1), 15–27 (2007)
34. Wynters, E.: Parallel processing on NVIDIA graphics processing units using CUDA. *Journal of Computing Sciences in Colleges* 26(3), 58–66 (2011)
35. Awad, A.I.: Fingerprint local invariant feature extraction on GPU with CUDA. *Informatica (Slovenia)* 37(3), 279–284 (2013)
36. Awad, A.I.: Machine learning techniques for fingerprint identification: A short review. In: Hassanien, A.E., Salem, A.-B.M., Ramadan, R., Kim, T.-h. (eds.) AMLTA 2012. CCIS, vol. 322, pp. 524–531. Springer, Heidelberg (2012)
37. Peacock, C., Goode, A., Brett, A.: Automatic forensic face recognition from digital images. *Science & Justice* 44(1), 29–34 (2004)
38. Bereta, M., Karczmarek, P., Pedrycz, W., Reformat, M.: Local descriptors in application to the aging problem in face recognition. *Pattern Recognition* 46(10), 2634–2646 (2013)
39. Jain, A.K., Li, S.Z.: *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus (2005)
40. Park, U., Jain, A.K.: Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security* 5(3), 406–415 (2010)
41. Soukup, D., Bajla, I.: Robust object recognition under partial occlusions using NMF. In: *Computational Intelligence and Neuroscience*, vol. 2008, 14 pages. Hindawi Publishing Corporation, ID 857453 (2008)
42. Gabor, D.J.: Theory of communication. *IEE* 93(26), 429–457 (1946)
43. Daugman, J.: High confidence visual recognition of persons by a test of statistical independence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15(11), 1148–1161 (1993)
44. Chen, W.K., Lee, J.C., Han, W.Y., Shih, C.K., Chang, K.C.: Iris recognition based on bidimensional empirical mode decomposition and fractal dimension. *Information Sciences* 221, 439–451 (2013)

45. Belcher, C., Du, Y.: A selective feature information approach for iris image-quality measure. *IEEE Transactions on Information Forensics and Security* 3(3), 572–577 (2008)
46. He, X., Yan, J., Chen, G., Shi, P.: Contactless autofeedback iris capture design. *IEEE Transactions on Instrumentation and Measurement* 57(7), 1369–1375 (2008)
47. Rankin, D., Scotney, B., Morrow, P., Pierscioneck, B.: Iris recognition failure over time: The effects of texture. *Pattern Recognition* 45(1), 145–150 (2012)
48. Daugman, J., Downing, C.: No change over time is shown in Rankin et al. iris recognition failure over time: The effects of texture. *Pattern Recognition* 46(2), 609–610 (2013)
49. Rankin, D., Scotney, B., Morrow, P., Pierscioneck, B.: Iris recognition—the need to recognise the iris as a dynamic biological system: Response to Daugman and Downing. *Pattern Recognition* 46(2), 611–612 (2013)
50. Juhola, M., Zhang, Y., Rasku, J.: Biometric verification of a subject through eye movements. *Computers in Biology and Medicine* 43(1), 42–50 (2013)
51. Srihari, S.N., Huang, C., Srinivasan, H., Shah, V.: Biometric and forensic aspects of digital document processing. In: Chaudhuri, B.B. (ed.) *Digital Document Processing*. Springer (2005)
52. Gonzalez-Rodriguez, J., Fierrez-Aguilar, J., Ramos-Castro, D., Ortega-Garcia, J.: Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International* 155(2-3), 126–140 (2005)
53. Ali, T., Spreuwers, L., Veldhuis, R.: A review of calibration methods for biometric systems in forensic applications. In: 33rd WIC Symposium on Information Theory in the Benelux, pp. 126–133. WIC, The Netherlands (2012)
54. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer (2006)
55. Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D.: *Introduction to Pattern Recognition: A Matlab Approach*. Academic Press (2010)

# Incorporating Language Identification in Digital Forensics Investigation Framework

Nicholas Akosu and Ali Selamat

Software Engineering Department,  
Faculty of Computing,  
Universiti Teknologi Malaysia,  
81310 Skudai, Johor, Malaysia  
nickakosu@yahoo.com,  
aselamat@utm.my

**Abstract.** In current business practices, majority of organizations rely heavily on digital devices such as computers, generic media, cell phones, network systems, and the internet to operate and improve their business. Thus, a large amount of information is produced, accumulated, and distributed via electronic means. Consequently, government and company interests in cyberspace and private networks become vulnerable to cyberspace threats. The investigation of crimes involving the use of digital devices is classified under digital forensics which involves adoption of practical frameworks and methods to recover data for analysis which can serve as evidence in court. However, cybercrime has advanced to the stage where criminals try to cover their tracks through the use of anti-forensics strategies such as data overwriting and data hiding. Research into anti-forensics has given rise to the concept of ‘live’ forensics which comprises proactive forensics approaches capable of digitally investigating an incident as it occurs. However, information exchange using ICT facilities has reduced the world into a global village without eliminating the linguistic diversity on the planet. Moreover, existing digital forensics frameworks have assumed the language of stored information. If such assumption turns out to be wrong, semantic interpretation of extracted text would also be wrong leading to wrong conclusions. We propose incorporation of language identification (LID) in digital forensics investigation (DFI) models in order to help law enforcement to be a step ahead of criminals. In this chapter, we outline issues of language identification in DFI frameworks and propose a new framework with language identification component. The LID component is to carry out digital surveillance by scrutinizing emails, SMS, and text file transfers, in and out of the system of interest. The collected text is then subjected to language identification. Determining the language of the text would help to decide if the communication is regular and safe or suspicious and should be subjected to further forensic analysis. Finally we discuss results from a simple language identification scheme that can be easily and quickly integrated to a DFI model yielding very high accuracy without compromising speed performance.

**Keywords:** Digital forensic framework, Anti-forensics, Language identification, Under-resourced languages, Spelling checker.

# 1 Introduction

In today's world a lot of business/private information is stored on computers and other electronic devices on a regular basis. Not only is the volume of such information enormous, it is also movable from one place to another and in diverse ways. The possibilities of storage, retrieval, and sharing of information are so great and efficient that some business transactions are only possible by such means. For example, it is now possible to complete a business transaction by sending an email, speaking on a GSM phone, sending a text message, all in a bid to clarify the business transaction and then finalizing it by means of a bank transfer or payment by means of online payment using credit/debit cards. This scenario affords such a good mix of business data, social data, private data, economic data, etc. making it hardly possible to determine the kind of information that is flowing in/out of a computerized system at any given time. This situation poses a great security risk in terms of the kind of information that may be passed to an outsider for ulterior motives or the kind of database that may be broken into using technical expertise. The array of information that gets stored or transferred through computer systems in the ordinary course of business is as follows:

- Company/Government database (financial data, stocks, staff information, contact information, etc.).
- Telephone communication.
- Mobile Short messages service (SMS).
- Email with or without attachments e.g. reports.
- Data/file sharing through Local/Wide Area networks.
- Internet online business (orders, parcel monitoring, payments, etc.).

Given such a complicated and diverse setup it is clear that trying to find out 'what went wrong' especially long after such an event has taken place can be an uphill task. This establishes the need for DFI. The term digital forensics has been defined in many ways by different researchers and practitioners. Nikkel [1] defined digital forensic as the utilization of scientific methods for the identification, collection, validation, preservation, analysis, interpretation, documentation and presentation of digital evidence for the purpose of realizing the reconstruction of criminal events or facilitating the detection of unauthorized actions capable of disrupting planned operations. Because of current advances in Computer technology and the ever increasing dependence of the business community on ICT this definition must include evidence from computer-related media, like cell phones, memory sticks, PDA's, and evidence from network traffic. However, it is not to be assumed that DFI work is usually carried out in a system that must be running normally. Sometimes the criminals may even try to cover their tracks through making the system un-operational, hiding data or overwriting data, etc. In some cases DFI experts need to take out hard disks and other media to extract information. This would most likely be at the suspect's office or at the crime scene. In some cases, hand-held devices like GSM may be used and the DFI expert may need to examine the content of such devices.



However, as earlier stated criminals may make deliberate efforts to block DFI, the practice that has come to be known as anti-forensics. The implication of anti-forensics is that it may delay investigation, lead to wrong information, result in no information being found or its relevance being hard to connect to the rest of the investigation. As a reaction to anti-forensics, the concept of live forensics or proactive forensics has emerged. Its major aim is to put in place digital infrastructure that is capable of monitoring any digital system of value in order to ensure that unauthorized access/conduct is prevented or arrested at an early stage.

In this chapter, we present a component based digital forensic framework with a language identification component as an important surveillance system – a unit capable of tracking all text transfers in and out of the system of interest aimed at enhancing the proactive digital forensic strategy. The functions of the language identification component include running concurrently with the system it is supposed to protect, regularly extracting text and identifying the natural language of the text, and feeding its output into the semantic unit that determines the meaning of the text. In this way, the system is assured of the nature/purpose of almost every interaction. Thus, any communication or interaction outside the ordinary is alerted for further DFI. As can be gathered from the literature, digital forensics (DF) is concerned with using scientific techniques to reconstruct an event after it has happened. It involves data collection/extraction, identification, validation, analysis, interpretation, preservation and documentation of data, for use as evidence in prosecution of a crime, or discovery of plans to disrupt normal operations. In this process the language of communication is vital and must be determined before any further meaningful processing can take place. So far most DFI frame works have assumed the language of the data in question, an assumption that could jeopardize the entire effort due to the linguistic diversity in the World.

## **2 Linguistic Diversity and Digital Forensics**

According to [2] there are over 7000 natural languages in the World. This fact is becoming more important now that criminals are embarking on anti-forensics as a strategy for evading Justice. While steganography is a much more sophisticated means of information hiding in pictures, etc., minority languages could easily serve a similar purpose if criminals choose to use one for communication. This raises the question of how information hiding using linguistic approaches can be stopped or discovered either in the course of live forensic investigation or in the ordinary course of a reactive DFI.

Naturally, the first step in processing information in natural language is to identify the language. It is only after the language has been identified that steps to be taken in further processing can be decided. If the language is one that the DF investigator can understand, the matter is straight forward. If she/he cannot understand the language, knowledge of the language of the text helps him/her in the search for a translator, either human or automatic. After translation the decision as to whether the text is of importance to the investigator can be easily made. Hence language identification is a key technology in tackling linguistic diversity in digital forensics.

### **3 Issues of Language Identification in Digital Forensics Investigation**

It would appear that identifying the language of a text in order to proceed with a DFI is a straight forward problem. However, the reality is quite different because of the volume of data that needs to be analyzed in a relatively short time. Moreover, we have already noted that there are over 7000 living languages available on Earth. The sheer number of languages alone poses a big challenge for identification coupled with the fact that these languages are at varying levels of development. For example, only few of the languages have been studied with respect to automatic language identification. In such cases some statistical and other methods have been developed for distinguishing them given varying amounts of test text. It has been established that some methods fail if the text available (for training or identification) is too small. However, the languages currently identifiable by computational techniques, are rather few and even then a list of such languages is not available anywhere. From the literature, it can be estimated that the number of languages so far studied may be about 200, even though one recent study by Brown [3] claimed the capability to identify more than 900 languages without listing them, of course.

Apart from the languages discussed above, there is yet a second category of languages that cannot be identified using statistical methods because there are no digital resources to enable such a study in the first place. These are what are generally referred to as minority languages or under-resourced languages. Another issue of immense importance is that of multilingual identification. Statistical methods are best suited for identifying a text that is written in a single language. For example, [4] noted that statistical (commercially available) methods can identify two individual strings as belonging to two different languages; but when the separate strings are concatenated, the methods frequently fail to identify the resulting multilingual string. In some cases the methods identify the text as belonging to a different language entirely, not even one of the original languages of the constituent parts. This is extremely worrying because in DFI there is already the likely hood that the criminals would try many ways to hide information, one of which could be the use of multiple languages.

The third and closely related issue has to do with the volume of text to be identified. If the amount of text is too small some techniques fail to identify the language correctly. The fourth issue worth mentioning here is the very important question of how much time these language identification methods take to determine the language of a text. In this regard, we find that most proposed techniques are not adequately investigated with respect to time complexity issues. However, some techniques are actually reported to be computationally expensive. DFI is already known to be a time consuming task. Therefore, it is important to seek ways that will improve performance, not aggravate it. We submit that there is need to intensify research on language identification particularly aimed at solving the above stated issues and to try to chart technical paths for making integration of language identification into proactive (live) DFI a viable endeavor.

## 4 Background on Digital Forensic Investigation Framework

Digital forensic investigation is a process that employs science and technology to develop and test theories, which can be tendered in a court of law, to answer questions concerning events under investigation. The central point in any DFI is the evidence. However, digital evidence has peculiar characteristics. It may be viewed as data of investigative value that is stored on or transmitted by some digital device. Thus, digital evidence is essentially, hidden evidence similar to Deoxyribonucleic Acid (DNA) and fingerprint evidence which are also hidden. According to [5] digital evidence is fragile and can be changed or damaged by improper handling and inappropriate methods of examination thereby altering its original state; extra care must be taken in collecting, documenting, preserving, and analyzing digital evidence.

Digital forensics analysis involves implementation of many functions involving the use of tools, such as cryptographic hashing of files, thumbnail generation, and extraction of ASCII sequences; also used are, keyword indexing, and the examination of images for evidence of steganography. Three distinct types of digital forensic analysis were identified at the Digital Forensic Research Workshop (DFRWS) namely, Code analysis, Media analysis and Network Analysis. Recent developments in the field suggest that digital forensics can be divided into four categories: network forensics, database forensics, mobile forensics and small device forensics. Moreover, cloud forensics is viewed as a subset of network forensics (DFRWS, 2001) [6] because network forensics is concerned with forensic investigations in public and private networks. Because cloud computing is based on broad network access, it can follow the main phases of network forensics adapted to the cloud computing environment in each phase. Special challenges attend the area of DF investigation in cloud computing. Here the great variety in processing devices poses a challenge for data discovery and evidence collection. The effect of crime and the workload involved in the investigation of crimes in cloud computing is aggravated by the large number of resources connected to the Cloud[7]. In addition, the use of audit logs can be compromised by time synchronization in the process of sourcing evidence. It is established that accurate time synchronization is an issue in network forensics, and the challenge is even greater in a cloud environment requiring timestamps to be synchronized across many physical machines spread across geographical regions involving several cloud infrastructure and remote web clients with numerous end points. Furthermore, DFI in the cloud can be hampered by incompatibility of log formats among two or more parties carrying out joint investigations[8].

The need for a standard framework has been in the focus of researchers as far back as 2001 when the first DFRWS was held [6]. Even then researchers are not yet in agreement on a particular framework. However, being a framework that must produce results in a highly dynamic environment, a DFI framework needs to be flexible in order to accommodate the development of requirements for the needed tools and procedures for testing each phase. In 2001, DFRWS accepted a framework which involves identification, collection, preservation, examination, analysis, presentation and decision. This framework has been the basis for all the models that have been proposed since then. In 2006, [9] proposed a model that merged the most essential features of all the existing models. More recently, [10] proposed an eleven-phase systematic DFI model and spelt out techniques for handling volatile and non-volatile

evidence collection. The researchers noted that using a combination of tools in the data collection phase can improve results.

In the past few years, DFI has been faced with several challenges namely increasing volume of data, the time required to process such data, types and variety of data sources, etc. In addition the emergence of anti-forensics as a means of disrupting DFI (often used by criminals) has further complicated the procedures for DFI. In reaction to the above developments researchers have proposed a number of ways in which to tackle these challenges. In their research, [11] have proposed distributed digital forensics (DDF) as a means of tackling the volume of data and processing it in a timely manner. They outlined system requirements for the distributed digital forensic system and suggested a light weight frame work for DDF. According to [12] anti-forensics refers to methods and activities that prevent forensic tools and investigators from achieving their objectives. Examples of anti-forensics techniques are data overwriting and data hiding. Anti-forensics can prevent evidence collection, increase the investigation time, advance misleading evidence aimed at jeopardizing the investigation, and prevent identification of digital crime. The author further stated that a future DFI process will be facilitated by development of future tools and techniques to cope with collection and preservation of proactive evidence. In an earlier research, [13] emphasized the need for organizations to decide on the data to collect in order to investigate anti-forensics effectively. The need to develop standard procedures for live DFI cannot be over emphasized.

Another approach for dealing with the ever-increasing complexity of DFI was proposed by [12]. The researchers considered the development of a component based framework consisting of three components: Proactive digital forensics (ProDF), Reactive digital forensics (ReDF) and Active digital forensics (ActDF). According to the researchers, the goals of ReDF include determining the root-cause of the incident, linking the perpetrator to the incident, minimizing the level of damage of the incident and making it possible to successfully investigate the incident. The researchers further define the goals for ActDF as aimed at collecting relevant live criminal digital evidence (CDE), including volatile evidence, on a live system, minimizing the impact of an ongoing attack and helping to provide a meaningful starting point for a reactive investigation. Other researchers [14] have proposed new forensic techniques and tools for the investigation of anti-forensics methods, and have suggested automation of live forensic investigation. These approaches aim at dealing with digitally investigating an incident while it occurs.

According to [15] digital crime scene investigation has 3 main phases:

1. System preservation & documentation
2. Evidence searching & documentation
3. Event reconstruction & documentation

In the framework proposed by [16], the phases under the proactive component include:

- i) Proactive Collection (concerns automated live collection of a pre-defined data in the order of volatility and priority, specific to the requirement of an organization),

- ii) Event Triggering Function (targets suspicious events that can be triggered from the collected data),
- iii) Proactive Preservation (involves automated preservation of the evidence related to the suspicious event),
- iv) Proactive Analysis (deals with automated live analysis of the evidence, using forensics techniques such as data mining),
- v) Preliminary Report (refers to automated report for the proactive component).

The goals to be achieved by the addition of the proactive component include development of new proactive tools and techniques to investigate anti-forensics methods, capturing more accurate and reliable evidence in real time, promoting automation in subsequent phases in the proactive component of the DFI, providing reliable leads for the reactive component to take place and, finally saving time and money by reducing the resources required to carry out an investigation[16].

In this chapter we propose the incorporation of a language identification component in the DFI framework for effective live forensic investigation aimed at sound management of the volume of forensic data to be analyzed and reducing the time spent on processing forensic data.

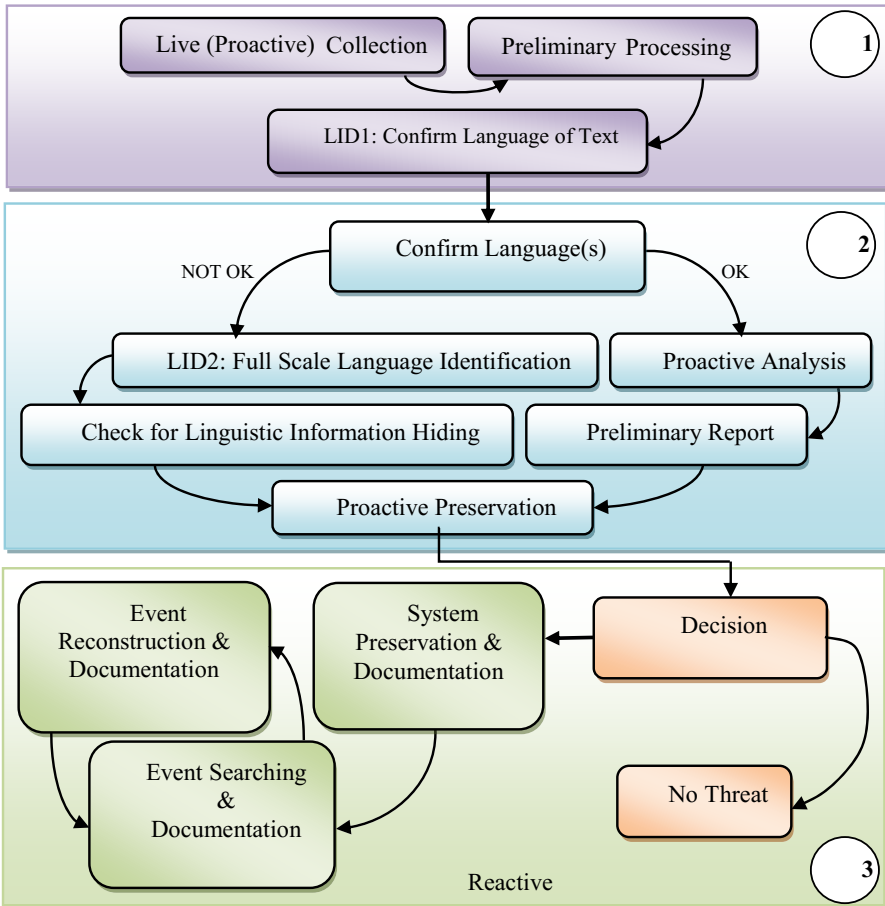
## 5 The Proposed Framework

Our framework is intended to be implemented as part of firewall forensics in which language identification is performed at 2 levels. It has three stages. In phase 1, the first level of language identification (LID1), is used for confirmation purposes i.e. it is understood that the system to be protected has boundaries and at these boundaries, there is need to perform a check to confirm that whatever text is allowed into the system is in the language(s) the system was built to work with. Therefore at the first step, language confirmation is performed. If this is passed, the level of alertness with respect to this text is reduced but the input is still passed to proactive analysis and ultimately to proactive preservation before a decision is made as to whether this is part of normal processing or not. The case would naturally be different when LID1 fails to recognize the language of a particular input. In which case the system is at heightened alert (phase 2), and full-scale language identification must be performed to determine the language of the input using LID2. Then there is a check for possible linguistic information hiding before proactive preservation. At this point a decision can be made as to whether or not such input should be subjected to full reactive DFI (phase 3). The proactive framework with language identification component is shown in Figure 1. The requirements to go in phase 1 include:

- a) Availability of all the language models for the intended languages in the particular system
- b) The readiness of the language identification technique which must already be installed and running in the background.

The requirements to enter in phase 2 of the framework include:

- i) The readiness to install as many language models as necessary to identify the language of the input text.
- ii) Availability of multi-processing facilities using GPU programming for efficient implementation of the language identification component.
- iii) Availability of automatic translators to enable further semantic processing of input after identification of the language.



**Fig. 1.** Pro-Active Digital Forensic Investigation Framework with Language Identification component

This figure shows the proposed framework made up of 3 phases. Inside the figure, Phase 1 specifies a need to carry out language confirmation. Phase 2 specifies full scale language identification, which will be performed when phase 1 reveals that the input text is not in any of the languages anticipated by the system. Phase 3 is entered when the need for a reactive forensic investigation is established. The usual procedures are then followed.

The requirements to go in phase 3 include:

- a) A strategic decision support system including possible threats in the given circumstances of the specific system.
- b) Adequate storage facilities and algorithms for event searching and event reconstruction

## **6 Proposed LID Technique Suitable for Digital Forensic Investigation**

There are two main objectives for incorporating language identification in DFI framework, namely monitoring the system of interest at the boundaries to prevent any criminal activities before they happen and reducing the volume of data to be processed, as well as the time required for analysis, if/when a reactive forensic investigation becomes necessary. Both objectives cannot be realized if the language identification technique is not efficient or is too cumbersome to operate. We propose the Spelling Checker approach for forensic language identification due to the following reasons:

- i) The technique is easy to launch; can be quickly launched for any language as long as such a language is written using an orthographic form that permits tokenization.
- ii) Uses whole words as features to carryout identification, i.e. uses language features that are basic and closest to the most natural form of communication by humans.
- iii) Fast in operation; speed can be improved by performance tuning.
- iv) This technique does not need large volumes of text to implement.
- v) It can correctly identify a single sentence.
- vi) The technique is capable of multi-lingual identification.

### **6.1 Theoretical Background**

The Spelling Checker technique is a simple method which identifies the language of a target text by comparing the words in the document with the list of words that exist in the vocabulary of any set of available languages. If a particular language emerges as having, in its lexicon, the largest number of words in the target text the system concludes that the target document must have been written in that language. The process starts with construction of the language models (vocabulary or lexicon for

each language) which are generated by tokenizing some reasonable amount of text in the various languages and elimination of duplicate words after pre-processing. The resulting language models are word lists comprising unique occurrences of words in each language. These models are taken as functional definitions of each language, which may grow with usage. The system determines the language of any input document by computing a binary matrix of the text by searching for each word (in the document) across all the language models.

The goal is to determine the status of each word in a document with respect to the vocabulary of a particular language. Any word,  $w$ , can only be a member of the vocabulary of a language if it is a valid word in the language. We have two conditions that need to be fulfilled,

1. Word,  $w$  is in document  $D$
2. Word,  $w$  is valid in vocabulary,  $V$ .

This can be expressed using statistical notation as follows:

$$\{w \mid w \in V \ \& \ D(w)\} \quad (1)$$

This captures “the set of all words,  $w$  such that  $w$  is an element of  $V$  (the vocabulary of the language) and  $w$  has property  $D$ . Equation (1) is used to build the binary matrix which is then analyzed to identify the language of the document,  $D$ .

The proposed technique considers language identification as a problem of computing the distribution over some set  $X$  of variables  $X_1 \dots X_n$ , (i.e. words) each of which takes values in the domain  $Val(X_i)$ , the vocabulary of the language. Thus, given a document  $D$ , the data set,  $D = \{x(1) \dots x(m)\}$ , where each  $x(m)$  is a complete assignment to the variables  $X_1 \dots X_n$  in  $Val(X_1 \dots X_n)$ . In order to compute an ‘ $N \times K$ ’ binary matrix which can be used to predict the language of  $D$ , we define a scoring function,  $Score(L: D)$  which generates the ‘ $N \times K$ ’ matrix relative to the data set,  $D$ .

The score is reduced to summary statistics with respect to the individual language models, using the generated binary matrix,  $M$ , defined as  $M[x_i, u]$  for each  $x_i \in Val(X_i)$ ;  $u \in \{L\}$ , set of language models. Therefore,

$$Score = \begin{cases} 1 & \text{if } (w \in V \text{ and } D(w)) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The sum of scores associated with each language model (wordlist spelling checker) is computed using the function:

$$Score(L: D) = \sum_{i=1}^n score(X_i) \quad (3)$$

The language is then determined based on some threshold value set by the user. For example, if the user decides on a threshold value of 70% this means that at least 70% of the words in the document must be confirmed as valid in a particular language in order to determine that the document is in the stated language. Research done by the



authors shows that a benchmark of 50% often gives 100% accuracy. This can serve as guide for users. The percentage (%) score is computed using equation (4),

$$Score (\%) = 100 * \sum_{i=1}^n score(xi)/n, \quad (4)$$

where n is the number of words in dataset, D.

The language of the document is determined by comparing the score (%) to the threshold value. The document language is unknown if score (%) is less than the required threshold value.

## 7 Spellchecker Model and Its Experimental Validation

In this Section we present the wordlist based spellchecker model and demonstrate its viability for language identification. To do this we explain the source documents, the tokenization process and the resulting language models and experimental results for some selected languages. The modeling of the spellcheckers was done using the tokenization method in which the input documents were stripped of special characters and numbers. The tokenization process then split the raw data into the individual words to form the wordlists of the spellcheckers for each language. The Universal Declaration of Human Rights (UDHR) act translations for the chosen languages provided the needed source documents for experimentation. The UDHR is translated into over 300 languages [17]. Spellcheckers for 4 languages were generated using the wordlist-based approach. The process of tokenization was preceded by the preprocessing step involving removal of special characters and numeric characters. Thereafter all duplicate words were removed and all characters were converted to lower case. In order to demonstrate the performance of the spellcheckers, the spellchecker lexica were divided into five layers as follows:

- i) We divided the UDHR corpus translations for the 4 languages into 2 parts, one part made up of 90% of the corpus was used for training, (the lexica of the spellcheckers), while the remaining 10% of the corpus served as the testing set.
- ii) The training set was then split it into 5 layers using word frequency; the first layer consisted of words that occur '10 or more' times, the second part consisted of words that appeared between '5 and 9' times, the third layer was formed by words that occur '3 or 4' times, the fourth layer consisted of words that appeared twice while the fifth layer consisted of words that occur once. A similar stratification was used by [18] and [19].
- iii) We tested the spellcheckers using the testing set by determining the number of words that were recognized by the first layer of the spellchecker, then the number of words that were recognizable using the first layer plus the second layer and by cumulatively adding the subsequent layers, the performance of the spellchecker was recorded as it changed from cumulating layer 2 to layer 5 of the lexicon.

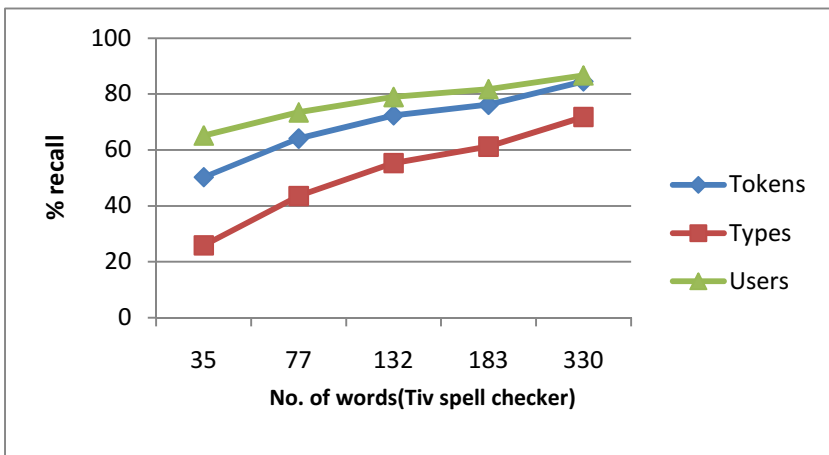
- iv) As an additional further validation step, we downloaded a larger English language corpus from project Gutenberg, consisting 192,427 tokens which were used to carry out 2 validation tests. In the first test, all the 192,427 tokens were used as training set while the UDHR translation (in English) was used as testing set. For the second validation test, we used 90% of the text from project Gutenberg for training and 10% for testing. Both experiments were done using the layering system outlined in sub-paragraphs (i) to (iii) above.

The process of determining ‘not recognized’ words was carried out using a computer program written in Python programming language.

## 7.1 Results

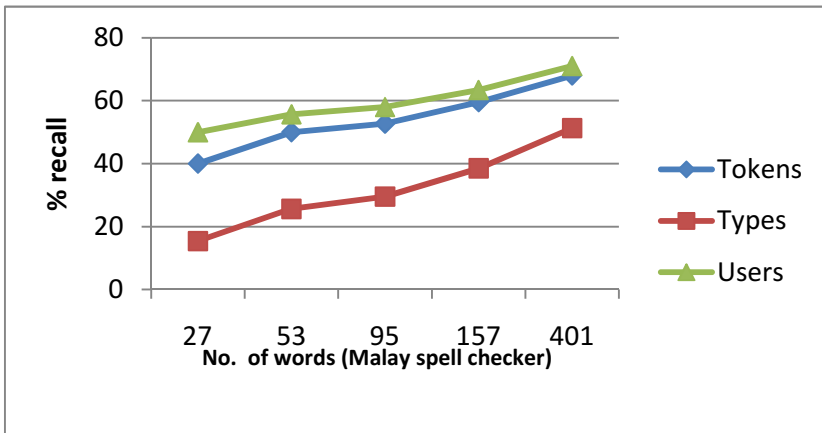
The spellcheckers for Tiv, Malay, and ‘English1’ were built using 90% of the UDHR translation in the respective languages. At the first level, i.e., using only words that have frequency of ‘10 or more’ we obtained a token recall of 50% (Tiv), 40% (Malay), and 51.3% (English1). Notice that we used ‘English1’ to distinguish the data set from UDHR as against English2 and English3 representing data set from the larger English corpus. Results for all the validation experiments are shown in Figures 2-4.

Also at the first layer, the type recall was 25.9% (Tiv), 15.4% (Malay), and 20.2% (English1). To the user, what matters is the type/token ratio in recall since the user only needs to include a word once for it to be recognized several times by the spellchecker. Therefore, the user’s recall was 65.2% (Tiv), 50% (Malay), and 55.1% (English1) at the first level.



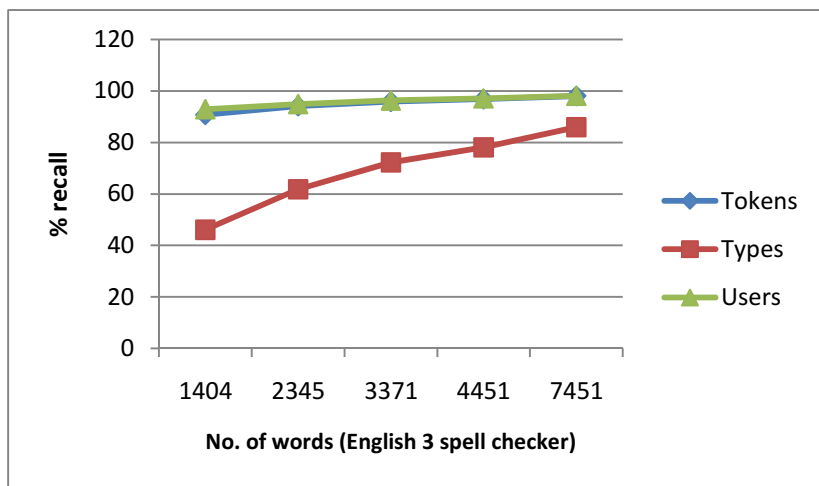
**Fig. 2.** Validation results for wordlist based spellcheckers for Tiv language

This figure shows graph of performance of the spellchecker for Tiv language. Inside the figure, it can be seen that the more words in the lexicon, the better the performance.



**Fig. 3.** Validation results for wordlist based spellcheckers Malay language

This figure shows graph of performance of the spellchecker for Malay language. Inside the figure, it can be seen that the more words in the lexicon, the better the performance.



**Fig. 4.** Validation results for wordlist based spellcheckers for English language

This figure shows graph of performance of the spellchecker for English language using a larger corpus. Inside the figure, it can be seen that the more words in the lexicon, the better the performance.

At the second level, the token recall increased to 64.1% (Tiv), 50% (Malay), and 59.5% (English1) while the type recall moved to 43.5% (Tiv), 25.6% (Malay), and 31.5% (English1) and the user’s recall climbed to 73.5% (Tiv), 55.7% (Malay), and 61.4% (English1). The results continued to improve until at the level of adding

the hapaxes we found that the token recall went to 84.5% (Tiv), 68% (Malay), and 78.5%(English1), while the user's recall finally stood at 86.7% (Tiv), 71.0% (Malay), and 79.1% (English1). The result for 15 languages is shown in Table 1 including the token, type and users' recall for English2 and English3. From these results it can be observed that the recall values did not show such a wide margin irrespective of the size of corpus used. Table 1 shows the accuracy of spellcheckers along with the size of corpus used to build the language models.

**Table 1.** Accuracy / Size of corpus

Language	No of words used to build spellchecker	No of types	Token/type ratio	Test set		Users' Recall (%)
				Tokens	Types	
Hausa	1643	411	4.0	183	115	86.9
Tiv	1622	330	4.9	181	85	86.7
Igbo	1728	363	4.8	193	81	91.7
Yoruba	1422	122	11.7	158	44	96.7
Malay	1175	401	2.9	131	78	71.0
Zulu	905	603	1.5	101	87	42.6
Swahili	1504	343	4.4	168	89	79.8
Ndebele	865	520	1.7	97	81	50.5
Indonesian	1213	417	2.9	135	91	65.2
Croatian	1235	637	1.9	138	106	62.3
Serbian	1288	639	2.0	144	114	61.1
Slovak	1199	640	1.9	134	104	59.0
Asante	1732	348	5.0	193	107	73.9
Akuapem	1784	263	6.8	199	83	90.0
English-1	1418	450	3.2	131	78	79.1
<b>English-2</b>	<b>192,427</b>	<b>7,811</b>	<b>24.6</b>	<b>1,781</b>	<b>533</b>	<b>89.1</b>
<b>English-3</b>	<b>173,184</b>	<b>7,450</b>	<b>23.3</b>	<b>19,243</b>	<b>2,550</b>	<b>98.1</b>

The above results and accompanying observations confirm that the first generation spellcheckers derived from the UDHR translations are of adequate accuracy and reliability and are thus suitable for application to other Natural Language Processing tasks such as Language Identification. This could be a significant contribution to the development of a digital resource base for under-resourced languages.

## 8 Discussion

The spellchecker model has very interesting properties, one of which is the fact that its performance is capable of improving with usage. We have seen this in the performance of the spellcheckers in Section 7. All the spellcheckers continued to improve in performance as more and more layers of words were added to the spellchecker lexica. This is particularly important because natural languages are living languages and they tend to evolve by inclusion of additional words. Secondly, all the

spellcheckers showed high recall values at the high frequency layers, i.e. layers 1 and 2. This confirms that there is a class of words that are consistently used more often than other words for every language. It is also interesting to mention that like in ordinary usage of spellcheckers for word processing, where the spellchecker lexicon is allowed to grow with usage, using spellcheckers for language identification can also benefit from the same feature – including newly discovered words in the lexica would enhance the capability of the spellcheckers. Finally we note that the ease with which a spellchecker model can be built is another advantage in that it is relatively easy to expand a given system by adding more languages to those it is capable of identifying. We have already noted that it is not necessary to have a large corpus in order to build a viable spellchecker model for language identification. From Table 1 we observe that in the case of English, the last 3 rows gave a comparable percentage recall irrespective of the size of text used to build the spellchecker models.

## 9 Summary and Conclusion

In this chapter we proposed the integration of a language identification component in DFI framework as a strategy for tackling anti-forensics. Using this approach promises some interesting advantages including, reducing the volume of data to be analyzed if/when a reactive DFI becomes necessary, reducing the time required for DFI, keeping investigators one step ahead of criminals and reducing the likelihood of using natural language for linguistic information hiding. Adopting this approach in the logical crime scene procedures will increase the DFI foot print, thereby further empowering law enforcement. Finally we propose the use of multi-processing by means of GPU programming for the language identification component. This will further reduce the time needed for detecting the language of online communication for effective surveillance of important data bases.

## References

1. Nikkel, B.J.: The Role of Digital Forensics within a Corporate Organization. In: Proceedings of the IBSA Conference, Vienna (2006)
2. Gordon, R.G.: Ethnologue: Languages of the world. SIL International, Dallas (2005)
3. Brown, R.D.: Finding and Identifying Text in 900+ Languages. Digital Investigation (2012)
4. Hammarstr-om, H.: A Fine-Grained Model for Language Identification. In: Workshop of Improving Non English Web Searching. Proceedings of iNEWS 2007 Workshop at SIGIR, pp. 14–20 (2007)
5. Carrier, B., Spafford, E.: Getting physical with the digital investigation process. International Journal of Digital Evidence 2(2) (2003)
6. Palmer, G.: A Roadmap for Digital Forensic Research. DFRWS Technical Report (2001), <http://www.dfrws.org/2001/dfrwsrmfinal.pdf>
7. Roussev, V., Wang, L., Richard, G., Marziale, L.: A Cloud Computing Platform for Large-Scale Forensic Computing. In: Peterson, G., Sheno, S. (eds.) Advances in Digital Forensics V. IFIP AICT, vol. 306, pp. 201–214. Springer, Heidelberg (2009)

8. Ruan, K., Baggili, I., Carthy, J., Kechadi, T.: Cloud Forensics: An Overview (2012), [http://www.loudforensicsresearch.org/publication/Cloud\\_Forensics\\_](http://www.loudforensicsresearch.org/publication/Cloud_Forensics_)
9. Kohn, M., Eloff, J., Olivier, M.: Framework for a Digital Forensic Investigation. In: Proceedings of the Information Security South Africa (ISSA), from Insight to Foresight Conference, Sandton, pp. 1–7 (2006)
10. Agarwal, M.A., Gupta, M.M., Gupta, M.S., Gupta, S.C.: Systematic Digital Forensic Investigation Model. *International Journal of Computer Science and Security (IJCSS)* 5(1) (2011)
11. Roussev, V., Richard, III., G.G.: Breaking the Performance Wall: The Case for Distributed Digital Forensics. In: Proceedings of the 2004 Digital Forensics Research Workshop, Baltimore, MD (2004)
12. Garfinkel, S.L.: Digital Forensics Research: The Next 10 Years. *Digital Investigation* (2010)
13. Garfinkel, S.L.: Anti-Forensics: Techniques, Detection and Counter Measures. In: Proceedings of the 2nd International Conference on i-Warfare and Security, p. 77 (2007)
14. Alharbi, S., Weber-Jahnke, J., Traore, I.: The Proactive and Reactive Digital Forensics Investigation Process: A Systematic Literature Review. *International Journal of Security and Its Applications* 5(4) (2011)
15. Carrier, B., Spafford, E.: An Event-Based Digital Forensic Investigation Framework. In: Proceedings of the Fourth Annual Digital Forensic Research Workshop, Baltimore, MD (2004)
16. Grobler, C.P., Louwrens, C.P., Solms, S.H.: A Multi-component View of Digital Forensics. In: ARES 2010 International Conference on Availability, Reliability, and Security, pp. 647–652 (2010)
17. UNESCO, Office of the High commissioner for Human Rights 1948, Universal Declaration of Human Rights., <http://193.194.134.190/udhr/index.htm> (accessed on August 19, 2011)
18. Prinsloo, D., Schryver, G.M.: Non-Word Error Detection in Current South African Spellcheckers. *Southern African Linguistic and Applied Language Studies* 21(4) (2003)
19. Veken, A.V., Schryver, G.M.: Les langues africaines sur la Toile. Etude des cas Haoussa, Somali, Lingala et isi-xhosa (Title set into English: Non-Word Error Detection in Current South African Spellcheckers). In: *Cahiers du Rifaal* 23 (Theme: Le traitement informatique des langues Africaines), pp. 33–45 (2003)

# Cost Optimized Random Sampling in Cellular Automata for Digital Forensic Investigations

Arnab Mitra<sup>1,3</sup> and Anirban Kundu<sup>2,3</sup>

<sup>1</sup> Adamas Institute of Technology, West Bengal-700126, India  
mitra.arnab@gmail.com

<sup>2</sup> Kuang-Chi Institute of Advanced Technology, Shenzhen-518057, P.R. China  
anirban.kundu@kuang-chi.org

<sup>3</sup> Innovation Research Lab (IRL), West Bengal-711103, India  
anik76in@gmail.com

**Abstract.** In today's world, advancement of Information Technology has been simultaneously followed by cyber crimes resulting in offensive and distressful digital contents. Threat to the digital content has initiated the need for application of forensic activities in digital field seeking evidence against any type of cyber crimes for the sake of reinforcement of the law and order. Digital Forensics is an interdisciplinary branch of computer science and forensic sciences, rapidly utilizing the recovery and/or investigation works on digital data explored in electronic memory based devices with reference to any cyber based unethical, illegal, and unauthorized activities. A typical digital forensic investigation work follows three steps to collect evidence(s): content acquisition, content analysis and report generation. In digital content analysis higher amount of data volumes and human resource(s) exposure to distressing and offensive materials are of major concerns. Lack of technological support for processing large amount of offensive data makes the analytical procedure quite time consuming and expensive. Thus, it results in a degradation of mental health of concerned investigators. Backlog in processing time by law enforcement department and financial limitations initiate huge demand for digital forensic investigators turning out trustworthy results within reasonable time. Forensic analysis is performed on randomly populated sample, instead of entire population size, for faster and reliable analysis procedure of digital contents. Present work reports about an efficient design methodology to facilitate random sampling procedure to be used in digital forensic investigations. Cellular Automata (CA) based approach has been used in our random sampling method. Equal Length Cellular Automata (ELCA) based pseudo-random pattern generator (PRPG) has been proposed in a cost effective manner utilizing the concept of random pattern generator. Exhibition of high degree randomness has been demonstrated in the field of randomness quality testing. Concerned cost effectiveness refers to time complexity, space complexity, design complexity and searching complexity. This research includes the comparative study for some well known random number generators, e.g., recursive pseudo-random number generator (RPRNG), atmospheric noise based true-random number generator (TRNG), Monte-Carlo (M-C) pseudo-random number generator, Maximum Length Cellular Automata

(MaxCA) random number generator and proposed Equal Length Cellular Automata (ELCA) random number generator. Resulting sequences for all those above mentioned pattern generators have significant improvement in terms of randomness quality. Associated fault coverage is being improved using iterative methods. Emphasis on cost effectiveness has been initiated for proposed random sampling in forensic analysis.

**Keywords:** Digital Forensic Investigation, Random sampling, Recursive pseudo-random number generator (RPRNG), True-random number generator (TRNG), Monte-Carlo (M-C) pseudo-random number generator, Cellular Automata (CA), Maximum Length Cellular Automata (MaxCA), Equal Length Cellular Automata (ELCA).

## 1 Introduction

Forensic researcher E. Casey has described forensic procedure [1] as a collection of work modules to be executed for serving crime alert. Digital forensics is a recognized scientific and forensic procedure used in digital forensics investigation and has been abruptly used to collect evidence against cyber/computer based unethical, illegal and unauthorized activities. Content acquisition, content analysis and report generation are followed in a typical digital forensic investigation. Digital devices detained for exploration purpose are referred as ‘exhibits’ in legal terminology. Scientific techniques have been utilized by the investigators to recovery the digital evidences validating assumptions for law and/or civil records [2].

Major problem is faced in analyzing the contents of a digital device. The volume of data required to be inspected in forensic investigation is rapidly increasing with time. In recent days, digital investigation is established as an expensive and time critical task for instability of existing forensic software dealing with large data and for the presence of existing backlogs in processing time by law enforcements. Exposures to distressed and offensive materials are often responsible for lowering physical and mental conditions of concerned investigators. The requirement of cost effective solutions in case of automatic digital forensic investigations has initiated the usage of calculated samples instead of investigating total population of data. The samples are fetched in a guided random way from entire data collected from ‘exhibits’ [3].

Random numbers are defined as homogeneously distributed values over a well specified interval. Prediction for the next values is unfeasible for a random sequence [4].

The characteristics of random number [4-5] have described the fundamental distribution in a random sequence. No correlations between the successive numbers are found. Random numbers over a specified boundary are essentially normalized with some distributions such that each differential area is equally populated. Power-law distribution for random number has been described in Equation 1 [4].

$$P(x) = CX^n \text{ for } X \in [x_0, x_1] \quad (1)$$

where ‘P(x)’ is power-law distribution and ‘C’ is a constant;



Random numbers (patterns) [4-5] obtained by the execution of a recursive computer program are referred to as recursive pseudo-random number generator. ‘Pseudo’ is considered as the implicit deterministic way to form ‘randomness’.

An alternative way for generation of pseudo-random numbers is the usage of Monte-Carlo (M-C) RNG [6-7]. M-C is described as a stochastic method [6-7]. The term ‘Monte-Carlo’ has been introduced by Von Neumann and Ulam during World War II. M-C method has been applied to problems related to the atomic bomb. The mean values of stochastic variables are expressed as integral of variables in M-C method [8-9] as illustrated in Equation 2.

$$I = \int_D h(x)f(x)dx \tag{2}$$

where ‘D’ is high dimensional domain with coordinates ‘x’ and ‘f(x)’ is a non-negative function;

Equation 3 is further satisfied by ‘f(x)’.

$$\int_D f(x)dx = 1 \tag{3}$$

Non-deterministic method is primarily required in ‘seed’ selection for generation of true-random numbers (TRNs) [10]. ‘Seed’ is fetched from physical procedures such as radioactive decay, photon emissions or atmospheric noise.

Alternative method for generating pseudo-random number has been established with the usage of Cellular Automata (CA). CA [11] has been described as a dynamic mathematical model to represent the dynamic behavior of system. Application area of CA ranges from the field of computability theory to complexity science or from theoretical biology to micro-structure modeling. Regular frameworks of cells are found in CA structure. Each of the cells is either in ‘On’ or ‘Off’ state. The frameworks of CA cells could have different dimensions. Each cell surrounded by neighborhood cells is defined with respect to the particular cells. A typical structure of an n-cell Null Boundary CA is represented in Fig. 1 [12].

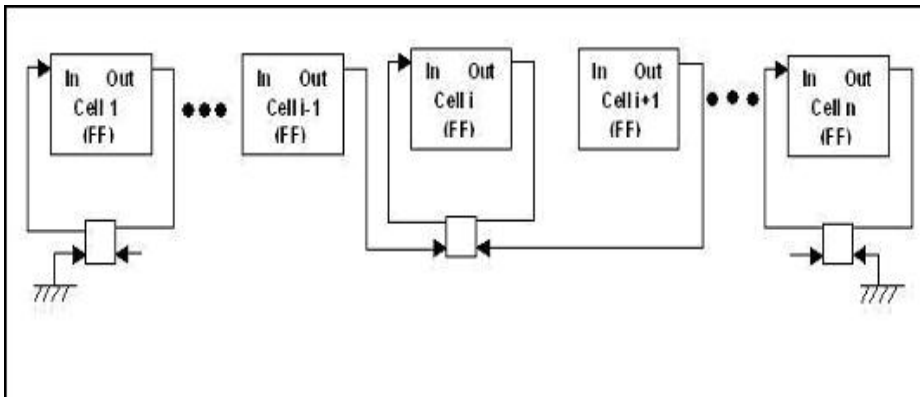


Fig. 1. Typical structure of n-cell Null Boundary CA [12]

Earlier researches on maximum length Cellular Automata (MaxCA) have established maximum randomness using its cycle having a length of  $2^{n-1}$  or  $2^n$  [13-18]. It is required to exclude prohibited patterns from generated random numbers using the same cycle [19-21].

Quality of randomness generated by random number generator is required to be verified. Diehard tests are a battery of statistical tests for measuring the quality of a random number generator [22].

Rest of the chapter is organized as follows: related works are discussed in Section 2; background has been reported in Section 3; proposed work is explained in Section 4; experimental observations and result analysis are shown in Section 5 and conclusion is drawn in Section 6.

## 2 Related Works

Previous researches [3, 23-24] have reported cost efficient randomized samples for digital forensics investigations. B. Jones et al. have mentioned the reduction in parallel clients' waiting time from three months to twenty four hours in case of digital forensics investigations using random sampling [3]. R. Mora et al. have reported reduction in processing time, higher accuracy and reduced chances of human errors in digital investigations [23]. O.D. Vel et al. have reported efficient usage of hidden markov model for tracking and prediction for degree of criminal activity over time using randomly sampled forensics scenarios [24]. Lack of focus on quality of randomness in randomized sample has been found in past.

Major efforts have been established to produce quality random numbers [8, 10, 14-22]. "http://www.random.org" [10] has pursued for generation of TRN using physical phenomenon like atmospheric noise as 'seed'. Important efforts have been reported in [14-22] for generation of pseudo-random numbers using CA. High degree of randomness has been recognized in MaxCA PRNG.

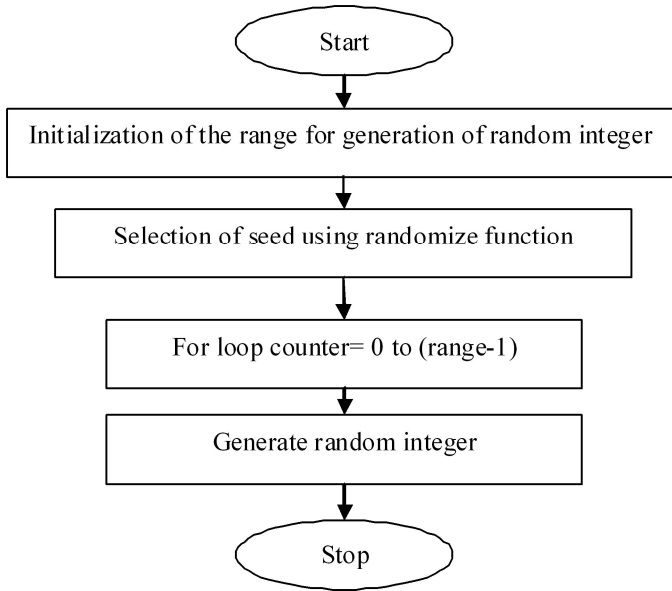
Pseudo-random numbers are achieved using a combination of 'randomize' and 'rand' functions. Equation 4 is utilized to achieve pseudo-random numbers [5].

$$X_{n+1} = P_1 X_n + P_2 (\text{mod} N) \quad (4)$$

where 'P<sub>1</sub>', 'P<sub>2</sub>' are prime numbers; 'N' is the range for random numbers; 'X<sub>n</sub>' is calculated recursively using the base value 'X<sub>0</sub>'; 'X<sub>0</sub>' is a prime number and referred as 'seed';

Pseudo-random pattern is achieved for a condition where 'X<sub>0</sub>' (seed) is fixed or it is selected in a deterministic procedure [4].

Recursive algorithm based computer program which is most frequently used as a source of random sequence has been considered in this chapter. Generation of pseudo-random numbers by recursive algorithm has been described in Fig. 2.



**Fig. 2.** Flowchart for pseudo-random number generation by recursion

Algorithm 1 has been used to prepare pseudo-random number as follows:

### **Algorithm 1. Recursive\_Pseudo-Random\_Pattern\_Generation**

Input: Upper limit for random numbers to be generated ( $n$ )

Output: Random pattern of integers

*Step 1: Start*

*Step 2: Initialize the range for which random integers to be generated*

*Step 3: Setting up of randomize seed*

*Step 4: Repeat Step 5 until required number of iteration has been achieved*

*Step 5: Generate random number using the random seed*

*Step 6: Stop*

M-C PRNG is optionally used to produce pseudo-random numbers within a specific boundary [6-9]. It is found with Monte-Carlo Simulator. M-C PRNG has been considered in this chapter.

TRNG is described by random.org in [10]. The atmospheric noise is fetched as 'seed' by random.org producing true random numbers.

Different degrees of randomness have been found using generated patterns by concerned RNGs. Recursive algorithm based RNG is deterministic in the procedure of 'seed' selection. The characteristics of TRNG are different from PRNG. TRNGs are ineffective compared to PRNGs over a time period towards generation of random numbers [10]. Particular sequence of numbers couldn't be reproduced in case of TRNG having non-periodic nature. There are chances for repetitions for the same sequence.

It has been observed that MaxCA PRNG is not a cost effective PRNG [25-26]. It is mandatory to keep track of PPS in maximum length cycle for excluding the prohibited patterns. Time, design and searching costs associated with random number generation have higher values. Careful consideration is required for selection of cost effective PRNG in Digital Forensics Applications.

Proposed work is emphasized on a cost optimized ELCA based PRNG for random sampling in digital forensic investigations. The detailed discussion of the proposed methodology is further illustrated in Section 4.

### 3 Background

An approach achieving high degree of randomization having better cost optimization with respect to all concerned complexities has been reported in [25-26]. The overall procedure is suitable for hardware implementation and its mathematical calculations have been reported as follows:

Consider, CA size of 'n';

Then,  $2^n = 2^{n-1} + 2^{n-1}$

$= 2^1 * (2^{n-1})$  (i.e. two number of equal length cycles)

$= 2^2 * (2^{n-2})$  (i.e. four number of equal length cycles)

$= 2^m * (2^{n-m})$  (i.e.  $2^m$  number of equal length cycles) for  $n \geq 1$  and  $m=1, 2, 3$

..... (n-1).

So we have,

$$2^n = 2^m * 2^{(n-m)} \quad (5)$$

Thus 'm' is always less than 'n' [25-26].

Example 1:

Consider, CA size 'n' is equal to 4.

So,  $2^4 = 2^1 * 2^{(4-1)}$  (i.e., total two numbers of equal length cycles of size eight),

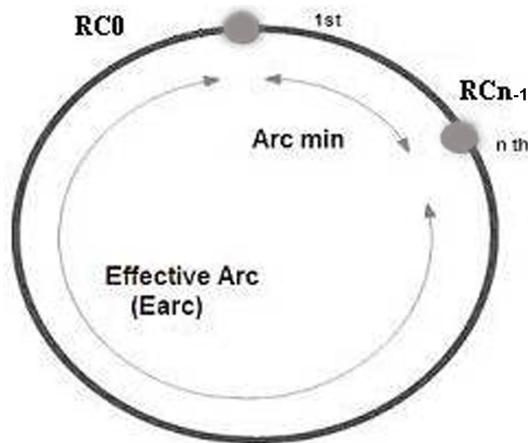
$= 2^2 * 2^{(4-2)}$  (i.e., total four numbers of equal length cycles of size four),

$= 2^3 * 2^{(4-3)}$  (i.e., total eight numbers of equal length cycles of size two).

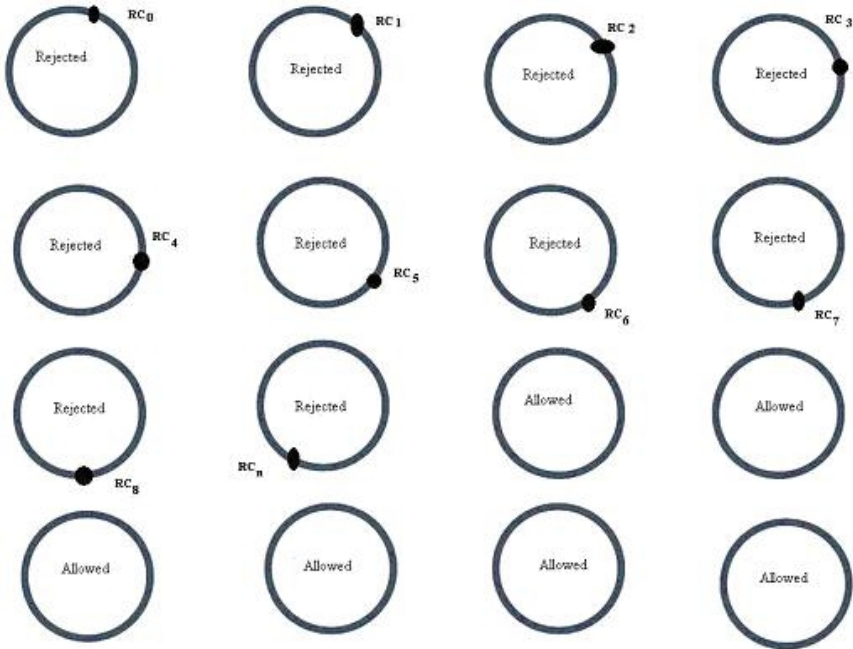
Generated ELCA are capable of producing random numbers as equivalent to the randomness quality achieved from MaxCA [25-26].

Proposed methodology is efficient to deal with prohibited pattern set (PPS). Prohibited pattern is referred to as a bit configuration found in a digital circuit for non-computability [25-26]. The occurrence of each prohibited patterns should be enclosed in any of the smaller sub-cycles such that the remaining cycles should be allowed to generate random numbers showing better cost effectiveness [25-26]. Hence, simplified design complexity and empowered searching complexity are found in our proposed approach considering zero overheads for keeping track for PPS in random number generation. More number of smaller equal length cycles is used in case of proposed n-cell ELCA as compared to n-cell MaxCA.

Each prohibited pattern is excluded from the cycle as per procedure of generating random numbers in MaxCA (refer Fig. 3(a)). In this scenario, PPS is excluded from the cycle of MaxCA. Assume, n-number of prohibited patterns in case of a MaxCA. Let, PPS is  $\{RC_0, RC_1, \dots, RC_{n-1}\}$ . Minimum length of arc ( $Arc_{min}$ ) between the prohibited patterns  $RC_0$  and  $RC_n$  should be measured for utilizing remaining arc (i.e. effective arc ( $E_{arc}$ )) for random number generations (refer Fig. 3(a)). Cycles containing prohibited patterns have been excluded from the procedure of generating pseudo random numbers for ELCA based PRNG [25-26] (refer Fig. 3(b)).



**Fig. 3.** (a) Typical cycle structure to deal with the problem of PPS in MaxCA



**Fig. 3. (b)** Typical cycle structure to deal with the problem of PPS in ELCA

Fig. 3 is generated having ‘n’ number of restricted configurations with following set:  $PPS = \{RC_0, \dots, RC_{n-1}\}$ .

It is mandatory to measure  $Arc_{min}$  and  $E_{arc}$  in case of n-cell MaxCA for complete fault coverage in subsequent random pattern generation.

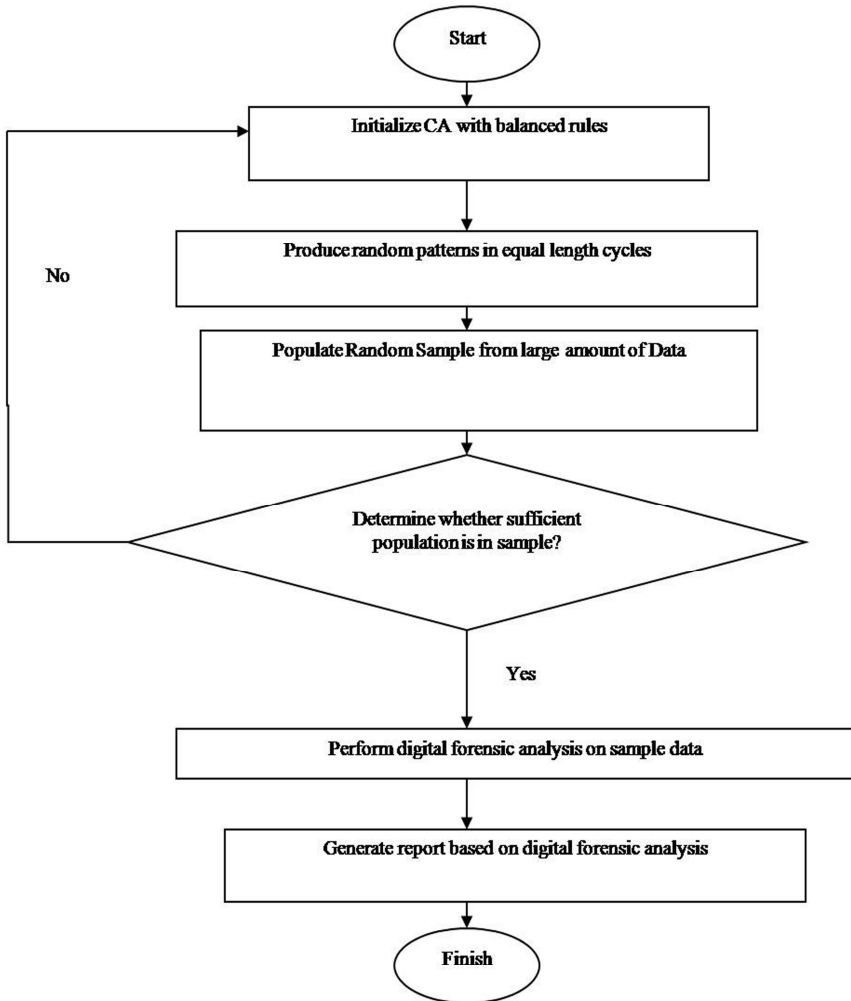
**Definition 1:** *Minimum length of Arc ( $Arc_{min}$ )* -  $Arc_{min}$  is the minimum distance between first and last prohibited patterns within the cycle of an n-cell MaxCA.

**Definition 2:** *Effective Arc ( $E_{arc}$ )* -  $E_{arc}$  is the remaining arc length of the cycle of an n-cell MaxCA excluding  $Arc_{min}$ . It is responsible for generating pseudo random patterns.

### 4 Proposed Work

Efficient random sample files have advantages for completion of digital forensics investigations to process within specified time [3, 23-24]. Cost effectiveness and simple design methodology are beneficial for generation of effective random sample files. The usage of ELCA based PRNG for random sampling in digital forensics applications has been emphasized.

We have emphasized the usage of cost effective ELCA based PRNG instead of opting for any other PRNG as CA based systems are easy to implement using D-type flip-flops. We have proposed a method to decompose an n-cell CA having large cycles into equal length cycles of smaller length such that concerned complexity costs should be reduced and the fault coverage should be flexible. The proposed digital forensic analysis system has been described using the flowchart in Fig.4.



**Fig. 4.** Flowchart of proposed Digital Forensic Analysis System

The preparation of effective random sample files using ELCA PRNG has been focused on particular forensics investigations as required (refer Fig. 4). Algorithm 2 has been used for ELCA generation.

### Algorithm 2. ELCA\_Generation

Input: CA size (n)

Output: m-length ELCA

*Step 1: Start*

*Step 2: Initialize the number of n-cell CA to generate random numbers using n-cell CA*

*Step 3: Initialize balanced CA rule to all the cells for generation of ELCA*

*Step 4: Decompose the cell number (n) into equal numbers (m) such that  $2^m * (2^{n-m})$*

*(i.e. 'm' number of ELCA) for  $n \cdot 1$  and  $m=1, 2, 3 \dots (n-1)$*

*Step 5: Generate random pattern*

*Step 6: Stop*

Proposed methodology is allowed only to generate random numbers from smaller cycles that do not contain PPS. The PPS exclusion feature from the main cycle has improved the design complexity. The logic behind this simplicity is that the proposed methodology has simply discarded the equal length cycles containing prohibited patterns. So there is no need to keep track of  $Arc_{min}$  length in the cycle. Concepts of  $Arc_{min}$  and  $E_{arc}$  are only applicable for MaxCA based design. Let, the time taken for calculating  $Arc_{min}$  and  $E_{arc}$  are  $T_{arc}$  and  $T_E$  respectively. So, the pattern generation time is  $T (T_{arc} + T_E)$ . There is no concept of calculating  $T_{arc}$  and  $T_E$  in our proposed ELCA based design. All the smaller cycles having PPS are discarded from pattern generation procedure. Therefore, effective time taken for pattern generations  $T_{ELCA}$  is equal to the execution time for remaining smaller length cycles having no PPS. Thus,  $T_{ELCA}$  is free from the overhead of calculation of  $T_{arc}$  and  $T_E$ . Hence, its design complexity is simpler compared to MaxCA.

### Example 2

A 4-cell CA is decomposed into some equal length smaller cycles instead of one maximum length cycle. There are several options to make the decomposition as per the real-time requirement. It can be decomposed into 4 smaller cycles of length 8; or, it can also be decomposed into 8 smaller cycles of length 4, and so on. Overall pattern generation in this type of scenario is followed as per Fig. 5. One maximum length cycle has been shown in Fig. 5(a). 4 equal length smaller cycles are shown in Fig. 5(b). Further, 8 equal length smaller cycles are shown in Fig. 5(c). Fig. 5(a) is based on Null Boundary 4-cell CA having rules in specified sequence <90, 150, 90, 150>. The synthesis of this example to generate ELCA is achieved by the 'ruleset' <195, 195, 195, 195> (refer Fig. 5(b)) and <51, 51, 51, 51> (refer Fig. 5(c)).



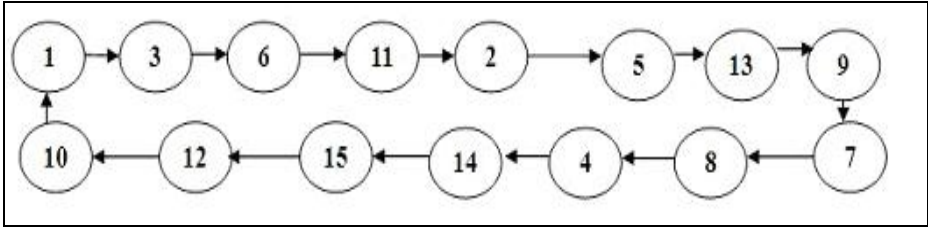


Fig. 5. (a). MaxCA Cycle for n=4 for <90, 150, 90,150>

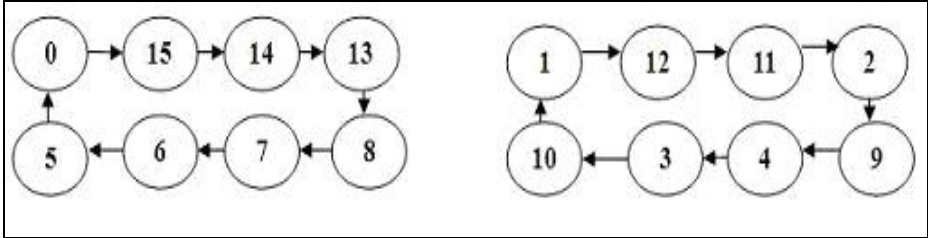


Fig. 5. (b). Proposed 2 ELCA of cycle size 8 for <153, 153, 153, 153>

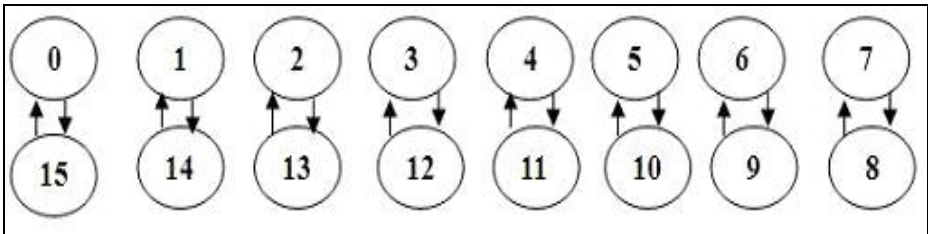


Fig. 5. (c). Proposed 8 ELCA of cycle size 2 for <51, 51, 51, 51 >

## 5 Experimental Observations and Result Analysis

Data sets generated by different RNGs have been reported in Fig. 6 visualizing randomness quality of corresponding RNGs.

It is observed in Fig. 6 that “Recursive PRNG” has least degree of randomness. “MaxCA PRNG” and “ELCA PRNG” share almost equal degree of randomness which is maximal compared to all other RNGs.

Diehard tests are being performed on data sets generated by PRNGs. ‘P-value’ has been calculated based on each of the tests. The ‘p-value’ generated for each sample data set by Diehard battery series test is convenient for deciding pass/fail for the test data set. The ‘p-value’ is uniform over [0, 1) for an input file in which truly



**Table 1.** (continued)

6	Monkey Tests OPSO, OQSO, DNA	Fail	Fail	Pass	Fail	Pass	Fail	Pass
7	Count the 1's in a Stream of Bytes	Fail	Fail	Pass	Pass	Pass	Pass	Pass
8	Count the 1's in Specific Bytes	Fail	Fail	Pass	Fail	Pass	Fail	Pass
9	Parking Lot Test	Fail	Fail	Pass	Pass	Pass	Pass	Pass
10	Minimum Distance Test	Fail	Fail	Pass	Pass	Pass	Pass	Pass
11	The 3DSpheres Test	Fail	Fail	Pass	Pass	Pass	Pass	Pass
12	The Squeeze Test	Fail	Fail	Fail	Fail	Pass	Fail	Pass
13	Overlapping Sums Test	Fail	Fail	Pass	Fail	Pass	Fail	Pass
14	Runs Test	Fail	Fail	Pass	Pass	Pass	Pass	Pass
15	The Craps Test	Fail	Fail	Pass	Pass	Pass	Pass	Pass
	<b>Total No. of Diehard Test Passes =</b>	0	0	11	10	14	10	14

Results obtained for different RNGs are further illustrated graphically in Fig. 7.

The results obtained in Table 1 and Fig. 7 have ensured maximum degree of randomization in proposed ELCA based design similar to the results achieved for Max-CA based RNG. Hardware, time, design, and searching complexities of different PRNGs have been enlisted in Table 2 for cost analysis in generation procedure of pseudo-random patterns. Required number of flip-flops for physical implementation of concerned PRNG system has been referred to as hardware complexity. Time required for generation of a pseudo-random pattern by RNG has been referred to as time complexity. The inherent design methodology dealing with problems of PPS has been considered as design complexity. The complexity associated with searching PPS free pseudo-random patterns is referred to as searching complexity.

## Randomness Quality Through Diehard Tests

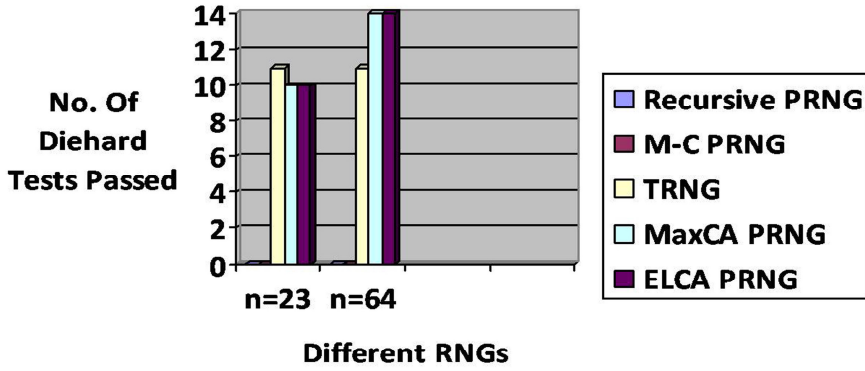


Fig. 7. Diehard Test Performance Graph

Table 2. Complexity comparison among different pattern generators

Name of the Complexity	Recursive	M-C	TRNG	MaxCA	ELCA	Remarks
Hardware	Not available	Not available	Not available	O(n)	O(n)	CA PRNGs are feasible for implementation using flip flops
Time	O (n) Here 'n' denotes number of iteration required in the concerned program	O (n) Here 'n' denotes number of iteration required in the concerned program	O (n) Here 'n' denotes number of iteration required in the concerned method	O(n) Here 'n' denotes length of cycle	$\sum O (m_i)$ Here 'm' denotes length of cycle and 'i' denotes number of ELCA	Single cycle in ELCA has less time complexity.

**Table 2.** (continued)

Design	Require randomize () for random seed selection and no such particular method to deal with PPS	Require specific mechanism for random seed selection and no such particular method to deal with PPS	Require natural source for random seed selection and no such particular method to deal with PPS	Require requirements for Calculation of $Arc_{min}$ to deal with PPS	Does not require to calculate any $Arc_{min}$ to deal with PPS	ELCA is simpler design to deal with PPS
Searching	No such particular method to deal with PPS	No such particular method to deal with PPS	No such particular method to deal with PPS	Require requirements for calculation of $E_{arc}$	Does not require to calculate $E_{arc}$	ELCA has simpler searching to deal with PPS

Advantages and superiority of ELCA based PRNG over other RNGs are reported in Table 2.

## 6 Conclusion

Degree of randomness achieved from random data sets for different RNGs have ensured ELCA along with MaxCA PRNG as the best options. ELCA PRNG is a cost effective solution for physical implementation compared to other RNGs as obtained by further analysis. ELCA has been utilized for preparation of small sets of random samples using generated smaller cycle lengths in an efficient way. Thus, proposed methodology is suitable for generating random sequences as samples in case of digital forensic investigations.

## References

1. Casey, E.: Digital Evidence and Computer Crime, 2nd edn. Elsevier (2004)
2. Digital Forensic Procedure, [http://en.wikipedia.org/wiki/Digital\\_forensic\\_process](http://en.wikipedia.org/wiki/Digital_forensic_process)
3. Jones, B., Pleno, S., Wilkinson, M.: The use of random sampling in investigations involving child abuse material. Digital Investigation 9 (2012), <http://www.dfrws.org/2012/proceedings/DFRWS2012-11.pdf>

4. Wolfram, S.: Wolfram Mathematica Tutorial Collection: Random Number Generation, <http://mathworld.wolfram.com/RandomNumber.html>
5. Eddelbuettel, D.: Random: An R package for true random numbers, <http://dirk.eddelbuettel.com/bio/papers.html>
6. [http://en.wikipedia.org/wiki/Monte\\_Carlo\\_method](http://en.wikipedia.org/wiki/Monte_Carlo_method)
7. Zio, E., Podofillini, L., Zille, V.: A combination of Monte Carlo simulation and cellular automata for computing the availability of complex network systems. Reliability Engineering System Staff (2006)
8. Gurov, T., Ivanovska, S., Karaivanova, A., Manev, N.: Monte Carlo Methods Using New Class of Congruential Generators. In: Kocarev, L. (ed.) ICT Innovations 2011. AISC, vol. 150, pp. 257–267. Springer, Heidelberg (2012)
9. <http://www.projectsmart.co.uk/docs/monte-carlo-simulation.pdf>
10. True Random Numbers, <http://www.random.org/>
11. Wolfram, S.: Theory and Application of Cellular Automata. World Scientific (1986)
12. Chaudhuri, P.P., Chowdhury, D.R., Nandi, S., Chattopadhyay, S.: Additive Cellular Automata Theory and Applications, vol. 1. IEEE Computer Society Press (1997)
13. Das, S., Kundu, A., Sikdar, B.K.: Nonlinear CA Based Design of Test Set Generator Targeting Pseudo-Random Pattern Resistant Faults. In: Asian Test Symposium, Taiwan (2004)
14. Das, S., Sikdar, B.K., Chaudhuri, P.P.: Nonlinear CA Based Scalable Design of On-Chip TPG for Multiple Cores. In: Asian Test Symposium, Taiwan (2004)
15. Martinez, D.G., Doinguez, A.P.: Pseudorandom number generation based on Nongroup Cellular Automata. In: IEEE 33rd Annual International Carnahan Conference on Security Technology (1999)
16. Das, S., Rahaman, H., Sikdar, B.K.: Cost Optimal Design of Nonlinear CA Based PRPG for Test Applications. In: IEEE 14th Asian Test Symposium, India (2005)
17. Hortensius, P.D., Pries, W., Card, H.C.: Cellular Automata based Pseudorandom Number generators for Built-In Self-Test. IEEE Transactions on Computer-Aided Design 8(8) (1989)
18. Bardell, P.H.: Analysis of Cellular Automata Used as Pseudorandom pattern Generators. In: International Test Conference (1990)
19. Das, S., Kundu, A., Sikdar, B.K., Chaudhuri, P.P.: Design of Nonlinear CA Based TPG Without Prohibited Pattern Set In Linear Time. Journal of Electrical Testing Theory and Applications (2005)
20. Das, S., Kundu, A., Sen, S., Sikdar, B.K., Chaudhuri, P.P.: Non-Linear Cellular Automata Based PRPG Design (Without Prohibited Pattern Set) in Linear Time Complexity. In: Asian Test Symposium, China (2003)
21. Ganguly, N., Nandi, A., Das, S., Sikdar, B.K., Chaudhuri, P.P.: An Evolutionary Strategy To Design An On-Chip Test Pattern Generator Without Prohibited Pattern Set (PPS). In: Asian Test Symposium, Guam (2002)
22. Brown, R.G.: Dieharder: A Random Number Test Suite, C program archive dieharder, version 1.4.24 (2006a), <http://www.phy.duke.edu/~rgb/General/dieharder.php>
23. Mora, R., Kloet, B.: The Application of statistical sampling in Digital forensics. Hoffmann Investigations, Almere The Netherlands (2010), <https://blogs.sans.org/computer-forensics/files/2010/03/statisticalforensictriage.pdf>

24. Vel, O.D., Liu, N., Caelli, T., Caetano, T.S.: An Embedded Bayesian Network Hidden Markov Model for Digital Forensics. In: 4th IEEE International Conference on Intelligence and Security Informatics, USA (2006), doi: 10.1007/11760146\_41
25. Mitra, A., Kundu, A.: Cost Optimized Approach to Random Numbers in Cellular Automata. In: Wyld, D.C., Zizka, J., Nagamalai, D. (eds.) *Advances in Computer Science, Engineering & Applications*. AISC, vol. 166, pp. 609–618. Springer, Heidelberg (2012)
26. Mitra, A., Kundu, A.: Cellular Automata based Cost Optimized PRNG for Monte-Carlo Simulation in Distributed Computing. In: *CUBE International Information Technology Conference & Exhibition 2012, India* (2012)

# Building Multi-modal Crime Profiles with Growing Self Organising Maps

Yee Ling Boo and Dammina Alahakoon

School of Information and Business Analytics,  
Faculty of Business and Law, Deakin University, Victoria, Australia  
{yee.boo, d.alahakoon}@deakin.edu.au

**Abstract.** Profiling is important in law enforcement, especially in understanding the behaviours of criminals as well as the characteristics and similarities in crimes. It could provide insights to law enforcement officers when solving similar crimes and more importantly for pre-crime action, which is to act before crimes happen. Usually a single case captures data from the crime scene, offenders, etc. and therefore could be termed as multi-modality in data sources and subsequently has resulted a complex data fusion problem. Traditional criminal profiling requires experienced and skilful crime analysts or psychologists to laboriously associate and fuse multi-modal crime data. With the ubiquitous usage of digital data in crime and forensic records, law enforcement has also encountered the issue of big data. In addition, law enforcement professionals are always competing against time in solving crimes and facing constant pressures. Therefore, it is necessary to have a computational approach that could assist in reducing the time and efforts spent for the laborious fusion process in profiling multi-modal crime data. Besides obtaining the demographics, physical characteristics and the behaviours of criminals, a crime profile should also comprise of crime statistics and trends. In fact, crime and criminal profiles are highly interrelated and both are required in order to provide a holistic analysis. In this chapter, our approach proposes the fusion of multiple sources of crime data to populate a holistic crime profile through the use of Growing Self Organising Maps (GSOM).

**Keywords:** crime profiling, multi-modal, data mining, data fusion, artificial neural networks, growing self organising maps.

## 1 Introduction

Since a decade ago, many investigative computer systems, with and without the involvement of artificial intelligence, has been designed for law enforcement with the purpose of recording and combating crimes. Many special purpose software have been developed to aid forensic experts in tasks such as the identification and comparisons of forensic evidence, information retrieval, etc [4,14,22,34]. After the incident of 9/11, the necessity in analysing and understanding crimes and criminals from voluminous crime databases has encouraged the proliferation of research in intelligent computer



systems for combating crime and countering terrorism. Thus, the focus of crime investigation has been expanded from the ability to catch the criminals towards the ability to act before a crime happens or before an offender commits a crime, so called pre-crime. Therefore in crime data mining community, many methodologies have been discussed for the tasks such as data associations among crimes, criminal network analysis, criminal career analysis, repeat victimization predictions, detection of criminal identity deceptions and many more [13,33,49,16,46]. In addition, the rise of text mining techniques have made the analysis of unstructured data such as crime reports, social media, etc. possible and effective for digital forensic. Specifically, such studies have been reported by [18,23]. Nevertheless, these methodologies or algorithms mostly involved the application of artificial intelligence and data mining techniques as well as statistical and mathematical models.

Criminal profiling or psychological profiling, which also plays an important role in aiding law enforcement for crime fighting, has limited computational and empirical discussions in the field of data mining or artificial intelligence. It is very common in business, medical or web domain to profile customers, patients as well as web users by using techniques such as data mining. Therefore, it is also reasonable and important to profile criminals, victims and crimes in order to gain further insights holistically. Subsequently, the profiles help to identify persons who are at risk and answer the questions such as why and where. These would definitely help the law enforcement in developing proper strategies and allocating human resources on the targeted crimes and criminals. To date, computational discussions on profiling include the use of Bayesian network modelling by [9] and the application of artificial neural networks by [40]. Particularly, the Bayesian network model has shown the possibility of building offender profiles computationally and could be used as a decision tool to speed up the investigative process through reduction of the list of suspects.

Forensic scientist and criminal profiler, Brent Turvey defined that the process of inferring the characteristics of criminals is commonly referred as criminal profiling. Such process also bears other terms such as behavioural profiling, crime scene profiling, offender profiling, psychological profiling, etc [43]. On the other hand, profiling in crime could also be viewed as the process of combining or fusing the different variety of data sources such as reports from autopsy, forensics reports, law enforcement reports, crime scene analysis, victimology and so on in order to successfully build a profile. In particular, such type of profiling is deductive by nature because forensic evidence is used to associate crime scene with victim in order to deduce the characteristics of a criminal [40]. The result of profiling is a criminal profile that normally consists of components such as the probable age, sex, race, residence, occupation, crime behaviour factors and so on [32,26]. Thus, a profile could be important in law enforcement, especially in understanding the behaviours of criminals and identifying characteristics of similar crimes. In addition, profiling could provide insights to law enforcement professionals when solving similar crimes. More importantly, they are equipped with insights that could help them to act before a crime occurs.

Traditional criminal profiling requires experienced and skilful crime analysts or psychologists to laboriously combine multiple sources of data. The act of profiling becomes personal as it is a process that highly depends on the experience, skills and tacit

knowledge from a profiler and also there are differences between individual profilers who embrace different profiling frameworks [10]. Consequently, there are numerous critics and debates about the process of traditional profiling in terms of the systematic approaches used, the accuracy and reliability of the results in scientific and empirical point of views [40,35]. Hence, we advocate that data mining and data fusion could contribute in building crime profile systematically and scientifically. We can imagine that a single crime case requires a variety of data sources from the crime scene, victims, physical evidence, offenders, witnesses, and so on. This could be likened to a state of multi-modality in which different modes of data could be referred to multiple sources of crime data and thus resulting in big data and complex fusion problems. Since law enforcement professionals are always competing against time in solving crimes and facing constant pressures, a computational approach is necessary as it could provide a way out in reducing the time spent on the laborious fusion process and increase the efficiency in crime solving.

In most of the profiling literature, the final goal of crime profiling is to obtain the demographics and physical characteristics as well as the behaviours of a criminal. A crime profile could also comprise of crime statistics or crime trend for various types of crime over a period of time in comparison to the socio-demographic details [3]. Therefore, we suggest that understanding criminal behaviours alone is not sufficient to obtain a complete profile of a crime and that crime and criminal profiles are highly interrelated and both are required in order to provide a holistic analysis. In addition, [32] states that the hybrid of machine learning and human reasoning, domain experience and expertise will be the ideal method in profiling. Hereafter, this will create the paradigm shift in profiling crimes where traditional methodologies are enhanced with artificial intelligence techniques. In parallel with this, we are motivated by the idea of complementing two processes - data mining and data fusion, for holistically building multi-modal crime profiles.

In this chapter we propose the use of a soft computing based technique in developing multi-modal crime profiles. Section 2 defines the concept of multi-modality and describes the characteristics of crime data. Section 3 briefly discusses the two complementary processes, data mining and data fusion. Section 4 describes the soft computing techniques - Self Organising Maps (SOM) and Growing Self Organising Maps (GSOM) and highlights the previous research in crime domain. Section 5 proposes our conceptual framework and follows with detail explanations for each phase in the framework. Section 6 demonstrates the conceptual framework through some experiment results. Finally, Section 7 provides the conclusion and future research.

## 2 The Multi-modality in Crime Data

Once a crime is committed by a criminal, a large amount of crime data could be generated from multiple sources. For instance, the multiple sources of crime data could come from the interviews with victim, analysis of forensic evidence, phone records of offenders, and so on. Therefore, we propose that the multi-modality issue in this domain could be perceived in five characteristics as depicted in Table 1. Each characteristic is correlated to each other and it could be further divided into greater details and thus the

**Table 1.** The multi-modality in crime data

<b>Crime Type</b>	<b>Data Source</b>	<b>Data Format</b>	<b>Data Structure</b>	<b>Data Type</b>
Property crime	Narrative report	Image	Temporal	Numerical
Violent crime	Victim data	Audio	Spatial	Categorical
Enterprise crime	Forensic reports	Video	Structured	Ordinal
	Offender data	Text	Semi-structured	Ratio
			Unstructured	Nominal

different level of the modality allows us to examine crime data in multi-dimensional perspectives.

According to criminologists, crime typologies can be divided into property crime, violent crime, white collar crime, public order crime and cyber crime [39]. On the other hand, FBI's Uniform Crime Reporting Systems reports crime statistics according to the different crime categories in greater details [1]. In crime data mining literature, [16,15] have tabularized a list of crime types into two major categories, namely local law enforcement level and national security level. In addition, the two categories are ordered according to the increasing public influences. Similarly, [13] discussed about criminal career analysis by taking different crime types into consideration in order to build a typical criminal profile. In brief, crime types could range from minor crime such as illegal parking to violent crime such as homicide or volume crime such as burglary. A single crime type is possible to be further categorized into specific criminal act. For instance, violent crime could be categorized into rape, homicide, hate crime, armed robbery and terrorism. Therefore, it is reasonable to consider crime types as one of the components in describing the multi-modality in crime domain.

Depending on the crime type, different types of contextual data source for crime investigation and profiling could be produced. However, there are some standard data sources such as crime scene reports and modus operandi which are applied to all of the crime types. Relatively, certain data sources such as forensic reports, autopsy reports or victims interviews are collected only if volume crime or violent crime are involved. Thus, it is also possible to define the modality in terms of the different variety of contextual data sources such as victim data, physical evidences, modus operandi, and so on. In addition, the multiple sources of crime data could further be expanded when other data sources such as demographic data, phone call records, credit card statements, etc. are required to aid in crime investigation. For instance, [28] included census attributes such as incomes, race, population age and etc. for crime association. Hence, the multi-modality of crime data source could easily trigger the issue of big data explosion. Specifically, such issue refers to the rapid and exponential increase in the amount digital crime data which are captured in a variety sources. The impacts and effects of such abundance are paramount and have transformed the operations in law enforcement.

The advancement of technology provides the convenient for law enforcement in recording and collecting digital crime data, thus large amount of crime data sources are stored and recorded in many data formats, as depicted in 1. Multimedia crime data such as crime scene photos, surveillance videos, telephone tapping records, narrative reports,

etc. are examples of different data formats. It is important to identify the multi-modality in terms of the data formats because different processing and analysis techniques are specially developed to handle and analyse them. For instance, police narrative reports that are usually appeared in free text are processed with neural networks by [15] to extract entities such as name from the reports and [22] uses neural networks as well for forensic shoeprints image classification. Thus, it is very important to acknowledge the necessity of having different methods in handling, processing and analysis of the multi-modality in data formats, which could greatly assist the crime investigators and profilers in their daily operations.

The multi-modality of crime data could also be viewed in the facet of data structure and some examples of crime data are listed in Table 2. Such disparate crime data could be analysed separately or collectively based on the different data structures and the necessity during crime investigations. For instance, [50] used spatial data such as the crime scene location for modelling the possible spatial site selection by criminal and [23] promoted the use of text mining technique for analysing emails, interviews and phone calls of unsolved homicide cases.

**Table 2.** The structure of crime data

<b>Temporal</b>	<b>Spatial</b>	<b>Structured</b>	<b>Semi-Structured</b>	<b>Unstructured</b>
Time	Victim address	Demographic data	Autopsy reports	Forensic images
Date	Crime scene location	Transactional data	Weblogs	Narrative reports

From the statistical and mathematical viewpoint, data is measured in different data types as shown in Table 1. Thus, the multi-modality of crime data could also be defined in the perspective of data types. We observe that crime data are represented in the combinations of data types such as categorical, numerical, ordinal, etc. For instance, demographic details contain the continuous numerical values such as age, weight and height of criminals and nominal value such as race of the criminals. Therefore, it is necessary to distinguish crime data in different data types especially for purposes such as data pre-processing and data transformation.

### **3 The Complementary Processes – Data Mining and Data Fusion**

We would emphasize that the different multi-modality of crime data are correlated especially the data format, data structure and data type. In addition, the distinctive classification of the multi-modality in crime data has provided an overview of the big data issues in which crime data could exist in high volume, high variety, high velocity and also high veracity. Therefore, it is imperative to have a new form of multi-dimensional processing and analysis for building multi-modal crime profiles. By reviewing the capabilities and potentials of the two processes - data mining and data fusion, a novel approach for building multi-modal crime profiles could be developed. Specifically, crime data could be explored and analysed by complementing both data mining and data fusion processes. This section briefly discusses data mining and data fusion separately and also addresses the significance of the two complementary processes.

### 3.1 Data Mining

The need to understand large, complex and information rich data sources is increasingly important and common in many fields such as business, science, engineering and medical [20]. Thus, there exist many computational techniques to extract hidden patterns from these data sets to help the domain experts in gaining insights of their data sets and also to improve their daily decision making processes. In fact, data mining is a process of extracting nontrivial and implicit information from large volume of data and the extracted patterns, namely knowledge, are usually unknown and could be potentially useful [17]. We can as well perceive data mining as one of the components in the Knowledge Discovery from Database (KDD) process because the overall KDD process also includes other components such as data selection, pre-processing, transformation and interpretation [44]. It is important to note that KDD is a recursive process and each of the processes are related to each others and thus they play equally important roles in discovering useful knowledge.

In the domain of crime, various crimes are reported daily and the ability to analyse high volume and different variety of crime data for the purpose of discovering crime patterns is far behind the ability to gathering and recording of crime data. Therefore, it is very normal for law enforcement to encounter the issue of big data. Besides, [49] addressed some characteristics in crime data where incompleteness, incorrectness and inconsistency are the common data problems in law enforcement. The multi-modality in crime data has increased the complexity of processing and analysing crime data in such domain. Hence, [30] states that it is time for law enforcement to march towards the era of using data mining for better understanding and analysing of huge amount crime data. The necessity to tackle the different characteristics of crime data encourages the application of artificial intelligence and machine learning approaches in data mining process. Such approaches provide the mean to perform thorough and speedy explorations in crime domain and allow crime data to be investigated at different angles.

### 3.2 Data Fusion

It is impractical to analyse only single source of data if we aim to obtain a better and complete picture of a problem. Therefore, there is a need to combine or fuse different sources of data for achieving the aim. Fusion is defined as merging or combining data or information supplied by multiple sources or same source at different periods of time and exploiting the joint information in tasks such as making decision, numerical estimation, resolving data conflict, building summary, etc. [41,11,42]. Depending on the modality of data, many techniques or methods ranging from mathematical aggregation operators and probabilistic models to machine learning and soft computing approach are developed to handle the different needs in the fusion problems [41,11,42,36]. According to [41], data fusion is performed for three purposes: (a) data pre-processing for improving data quality; (b) models building through aggregations of different data models; and (c) information extraction for summarising and repretating data.

As stated by [11], various application domains have encountered the problem of combining or fusing data from several sources. In particular, [42] addressed the application of data fusion in the areas of economics, biology, education and computer

science. For instance, [51] discussed about a framework for fusing data through a voting like process to adjudicate conflicts among the data supplied by multiple sources. On the other hand, record linkage as and re-identification procedures has been applied in distributed databases to identify multiple records from which are referring to the same object from distributed databases [41,36]. In person authentication problem, different models are built from biometric sources such as fingerprint, hand geometry and face images, the combination of the three models through weight allocation for the purpose of classifying a person correctly [38]. For the purpose of information extraction, Self Organising Maps(SOM) is applied in [29] to visualize the correlations among the multi-modality biomedical data by building summary of data and represented in a fish glyph.

Likewise, [30] reflected that crime data can be aggregated across multi-modal data sources because crime patterns could also be found in different variety of data. Moreover, crime profiling heavily relies on multi-modal data sources such as demographic details, forensic reports, police narrative reports, etc. in order to obtain comprehensive understanding of the criminals, victims and crime. Therefore data fusion could be regarded as a methodology in fusing multi-modal crime data. In addition, it is critical to highlight one of the big data issues - the veracity in crime domain and also for building of crime profiles. Similar to any other domains, there exists data quality issues such as lack of accuracy and reliability due to erroneous or missing data from multi-modal data sources. The analysis that is based on incomplete, erroneous and narrow view of data sources could affect the accuracy of crime profiles created which have an impact on the decisions taken by law enforcement professionals.

### 3.3 The Complementary Processes

Principally, data mining and data fusion could be regarded as two complementary processes in which automated knowledge discovery is achieved via data mining and synthesizing of multi-modal is accomplished by data fusion [45,12]. The complementary processes has been discussed in [45] for automatic target recognition (ATR) processes by using multiple sensors and sources of information in military domain. In addition, [19] mentioned that data mining generally discovers knowledge from single data source and the consequent fusion process could provide the means to achieve global information processing stream for appropriate interpretation and understanding of fused information. Similarly, [41] noticed that data or information fusion is becoming a major need in data mining community however the gap between both fields is still large. Thus, it is apparent that data fusion becomes necessary and essential for obtaining better outcomes in data mining.

As discussed previously, the issue of veracity exists in crime data as the collection and capturing of multi-modal crime data are incomplete and erroneous. Therefore, this could be addressed by data fusion in which the pre-processing of crime data before the stage of data mining could be performed. In addition, the multi-modality in crime data created the other big data issues such as high volume, high variety and high velocity. The complexity in multi-modal crime data requires methodologies that could deal with all these issues and it is vital to consider how mining and fusion processes can be conducted in harmony.

It has been commonly reported in crime data mining literature that the mining of crime data is performed directly on a huge database, in which all data sources are merged and linked into a big integrated database. Nevertheless, as pointed out by [52], the big integrated database may not reveal the distribution of patterns that could separately exist in multiple data sources before the merging process. This is because the merging process could have destroyed the trivial but also valuable patterns that exists in those separated data sources[52]. Thus, this raises the issue of how data can be mined separately at the multi-modal crime data and subsequently be fused to provide summaries and representations of patterns.

As mentioned previously, different crime profilers adopt different profiling frameworks and due to their different profiling experience in their careers, there are always debates in the field on how profiling could be standardized systematically and scientifically. Hence, it is a challenge for the field of artificial intelligence in considering the issues of how data fusion and data mining can be complemented in order to mimic human profilers in a multi-modality environment.

## **4 The Soft Computing Approach**

The final goal of profiling is to understand the underlying relationship between crime and criminal holistically. Nevertheless, profiling processes start with limited amount of clues and pieces of information, thus it requires profilers to be creative in performing exploratory analysis on the available data. Therefore clustering could be a suitable tool for naturally grouping crime data into groups of similar features regarding the crimes or criminals. In this section, we discuss about the very popular clustering technique, Self Organising Maps (SOM) with examples in crime domain and also identify the limitation of SOM in discovering patterns in crime data. In relation to the multi-modality issues discussed previously, we underline our approach, Growing Self Organising Maps (GSOM), an extension of SOM. We present the previous work done in crime domain and extend it by highlighting an important area to deal with one of the issues identified in multi-modal crime data. In that, we suggest that the further examination on crime data in multiple abstractions is necessary.

### **4.1 The Self Organising Maps (SOM)**

Clustering provides a way in analysing similar patterns in data, in particular when the patterns are unknown from the beginning. One of the popular and well known clustering methodologies is the Self Organizing Maps (SOM), which is an unsupervised neural network learning methods that can iteratively organize large input data into clusters in which data within a particular cluster are in high similarities compare to the neighbouring clusters [7]. It is noted that Self Organising Maps (SOM) is effective for visualizing high dimensional or volume data in which the data is mapped into lower dimensional data through dimensionality reduction [27]. In addition, SOM preserves the topological relationship of the data elements and thus providing abstractions for the high dimensional data [27].

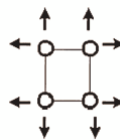
There are many examples of successful applications of SOM in solving crimes. For the applications in violent crimes, [5] discuss the application of SOM in modelling the

behaviour of offenders who commit serious sexual assaults using CRISP-DM methodology. Whereas [31] uses SOM to provide automated homicide crime analysis system based on the parameters identified by the police department. SOM also has been discussed in [32] to demonstrate the clustering on border smuggling activities while [6] aims to recognize burglary offences committed by a network of offenders through SOM. These applications have shown the capability of SOM in clustering multi-dimensional data through mapping into a two dimensional space, thus complexity of the problem has been reduced and at the same time crime patterns has been identified. Nevertheless, the crime patterns obtained from these applications are limited to single abstraction level. It does not have the capability to allow flexible examination on crime patterns due to the rigid structure in SOM.

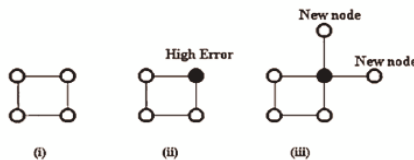
**4.2 The Growing Self Organising Maps (GSOM)**

An extension of the SOM, called the Growing Self Organizing Maps (GSOM) has been developed with the capability of self adapting according to the input data and could better represent clusters [7,8]. Unlike SOM, Growing Self Organizing Maps does not start with a predefined network, instead it is initialized with four nodes as shown in Figure 1. The four nodes are named as the boundary nodes. The entire node generation process begins at the boundary nodes in which each of the nodes are allowed to grow freely into desired directions. Figure 2 shows the process of node generation from boundary nodes. The new node is grown to represent input data using a heuristic approach and the allocation of weight values of nodes during node growth are similar to SOM, which is self organised.

To control the spread of the map, a concept called Spread Factor (SF) is developed for specifying the amount of spread that is needed for the analysis on data. Such characteristic allows clustering to be done hierarchically by gradually adjusting the values of Spread Factor. In fact, Spread Factor takes values from 0 to 1 and is regardless to the dimensions in the data. Therefore, data analysis usually begins with low value and slowly increases based on the further observations of the selected region of data. Thus,



**Fig. 1.** The initial GSOM with four boundary nodes



**Fig. 2.** The generation of new node from the boundary of network



it allows the comparison of results in multiple abstractions of the same data source and also the comparison of results from different data sources with a different number of attributes by mapping them with the same Spread Factor.

The following shows GSOM process, further explanations is described in [8].

1. Initialization phase:

- (a) Initialize the weight vectors of the starting nodes (usually four) with random numbers between 0 and 1.
- (b) Calculate the growth threshold ( $GT$ ) for the given data set of dimension  $D$  according to the spread factor ( $SF$ ) using the formula:

$$GT = -D \times \ln(SF) \quad (1)$$

2. Growing Phase:

- (a) Present input to the network.
- (b) Determine the weight vector that is closest to the input vector mapped to the current feature map (winner), using Euclidean distance. This step can be summarized as: find  $q'$  such that:

$$|\vartheta - \omega_{q'}| \leq |\vartheta - \omega_q| \quad \forall q \in \mathbf{N} \quad (2)$$

where  $\vartheta$ ,  $\omega$  are the input and weight vectors respectively,  $q$  is the position vector for nodes and  $\mathbf{N}$  is the set of natural numbers.

- (c) The weight vector adaptation is applied only to the neighbourhood of the winner and the winner itself. The neighbourhood is a set of neurons around the winner, but in the GSOM the starting neighbourhood selected for weight adaptation is smaller compared to the GSOM (localized weight adaptation). The amount of adaptation (learning rate) is also reduced exponentially over the iterations. Even within the neighbourhood, weights that are closer to the winner are adapt more than those further away. The weight adaptation can be described by:

$$\omega_j(k+1) = \begin{cases} \omega_j(k) & \text{if } j \notin \mathbf{N}_{k+1} \\ \omega_j(k) + LR(k) \times (x_k - \omega_j(k)) & \text{if } j \in \mathbf{N}_{k+1} \end{cases} \quad (3)$$

where the Learning Rate  $LR(k)$ ,  $k \in \mathbf{N}$  is a sequence of positive parameters converging to zero as  $k \rightarrow \infty$ .  $\omega_j(k)$  and  $\omega_j(k+1)$  are the weight vectors of the node  $j$  before and after the adaptation and  $\mathbf{N}_{k+1}$  is the neighbourhood of the winning neuron at the  $(k+1)$ th iteration. The decreasing value of  $LR(k)$  in the GSOM depends on the number of nodes existing in the map at time  $k$ .

- (d) Increase the error value of the winner (error value is the difference between the input vector and the weight vectors).
- (e) When  $TE_i \geq GT$  where  $TE_i$  is the total error of node  $i$  and  $GT$  is the growth threshold. Grow nodes if  $i$  is a boundary node. Distribute weights to neighbours if  $i$  is a non-boundary node.
- (f) Initialize the new node weight vectors to match the neighbouring node weights.
- (g) Initialize the Learning Rate  $LR$  to its starting value.
- (h) Repeat steps (b)-(g) until all inputs have been presented and node growth is reduced to a minimum level.

3. Smoothing phase:
  - (a) Reduce learning rate and fix a small starting neighbourhood.
  - (b) Find winner and adapt the weights of the winner and neighbours in the same waybas in growing phase.

### 4.3 Related Work

In the context of crime data, the usefulness and effectiveness of GSOM has been demonstrated in [48] in which GSOM was used to mine crime data that are distributed in different locations. The crime data sets consist of different types of crime including murder, rape, robbery, assault, burglary, larceny, theft and arson in different states in US. It was shown that when the SF is low, data with same states are grouped together and vice versa. Clusters could be obtained in high or low level of abstractions and could be also be represented in concept hierarchies.

Therefore, this chapter intends to further evaluate the capabilities of GSOM based on the previous research by [48]. There is a need to conduct lower level or detail analysis of crime data especially when the issues regarding multi-modality in crime data has been underlined. The hierarchical clustering by GSOM produces a vertical expansion of data provided that the input data consists of multiple granularities. After observing that the multi-modal crime data consists of such features, we are able to identify patterns in different granularity from the multi-modal crime data. Henceforth, we suggest that profiling of crime or criminals should be further investigated in greater details because there can be very implicit patterns hidden within the multi-modality of crime data.

In comparison to the previous applications of SOM in crime data mining [32,5,31,6], this chapter contributes by providing a novel framework for flexible mining of crime data. Moreover, it also contributes by fusing crime data at different levels because profiling is a process of exploring crime data separately and subsequently synthesizing all information to build a complete profiles. The ability to illustrate crime patterns at different levels of abstraction could facilitate profilers in understanding a crime and criminals.

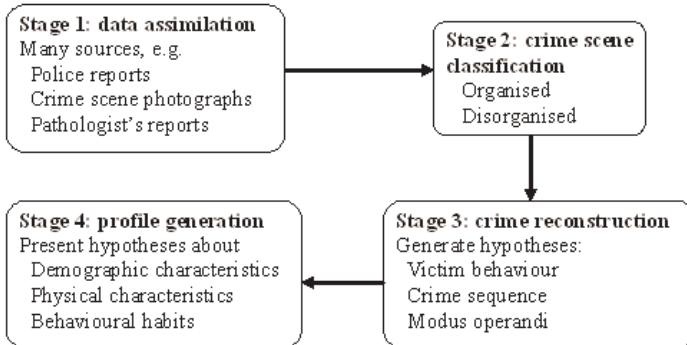
## 5 The Proposed Conceptual Framework

For profiling multi-modal crime data, we propose a conceptual framework that incorporates the two complementary processes - data mining and data fusion. Firstly, an existing profiling framework that is adopted by FBI is discussed and it is followed by a thorough discussion of our proposed conceptual framework, namely Multi-Modal Data Fusion in Crime (MMDF-C). In addition, the comparison our framework with the existing profiling framework adopted by FBI is illustrated and justified for crimes profiling. We also distinguish our proposed conceptual framework from the usual data mining or data fusion processes.

### 5.1 FBI Crime Scene Analysis (CSA)

A popular and well-known paradigm developed by FBI, namely Crime Scene Analysis (CSA) involves 4 stages in profiling a crime [24] is depicted in Figure 3. As explained

by [24], CSA starts with the collection of various data sources as a crime usually has a variety of associated documented materials. Then, it is followed by the classification of different crime scenes into two major categories: (a) organised, which means there exists evidence of proper planning for the crime and (b) disorganised, which means the crime committed is not planned and thus the crime scene is chaotic. Such classification has shown important relationship between the crime scene and psychology of the criminal and thus helping police officer to plan for their interrogation techniques.



**Fig. 3.** The major stages of FBI crime scene profiling, as illustrated by [24]

Crime scene reconstruction is conducted at stage 3 whereby the inference and deduction processes are involved for understanding the series of events between victim and offender. Given the information in stage 1, stage 3 attempts to clarify the modus operandi of an offender so that such information could be used to associate the currently investigated crime to the other similar crimes. Finally in last stage, profilers formulate a description of the offender by taking all the hypotheses together. A typical profile of an offender includes demographic details (race, age, sex, etc.), physical characteristics (height, weight, etc) and behavioural habits (hobbies, social activities, etc.).

Apparently, the profiles generated by CSA are focussing on offenders and therefore do not include the characteristics about crimes and victims. Although we know that the major focus is to identify the offender, we advocate that a profile should also include at least two of the three components, which are the description about the crime, criminal and victim (if any). We suggest that the three components are interrelated and should be included to build a complete and holistic crime profile.

## 5.2 Multi-Modal Data Fusion in Crime (MMDF-C)

Our proposed conceptual framework, namely Multi-Modal Fata Fusion in Crime (MMDF-C) is mainly developed for handling the multi-modal crime data for crime profiling. The development of MMDF-C was inspired by the multi-modal information

processing model developed by [25] and also the FBI CSA discussed by [24]. Moreover, the multi-modal information processing model reported in [25] considered only two audio and visual data sources.

As a result, MMDF-C has expanded, enhanced and enriched the two existing frameworks which originate from two different fields. The novelty of MMDF-C has accommodated the multi-modal data issues that have been discussed in previous sections. For clarifications and thorough explanations, MMDF-C is discussed progressively with the three different levels of depictions.

### 5.2.1 The Global View of MMDF-C

Figure 4 shows the global view of MMDF-C, in which multi-modal crime data are forwarded to the centre for multi-modal data fusion via GSOM. This is similar to the stage 1 in FBI CSA profiling framework, in which profilers are using multiple sources of crime data. The data fusion process conducted with GSOM will produce global clusters and show the different levels of abstraction for crime profiles. Thus, the construction of profiles is corresponded to the generation of crime profiles at stage 4 in FBI CSA.

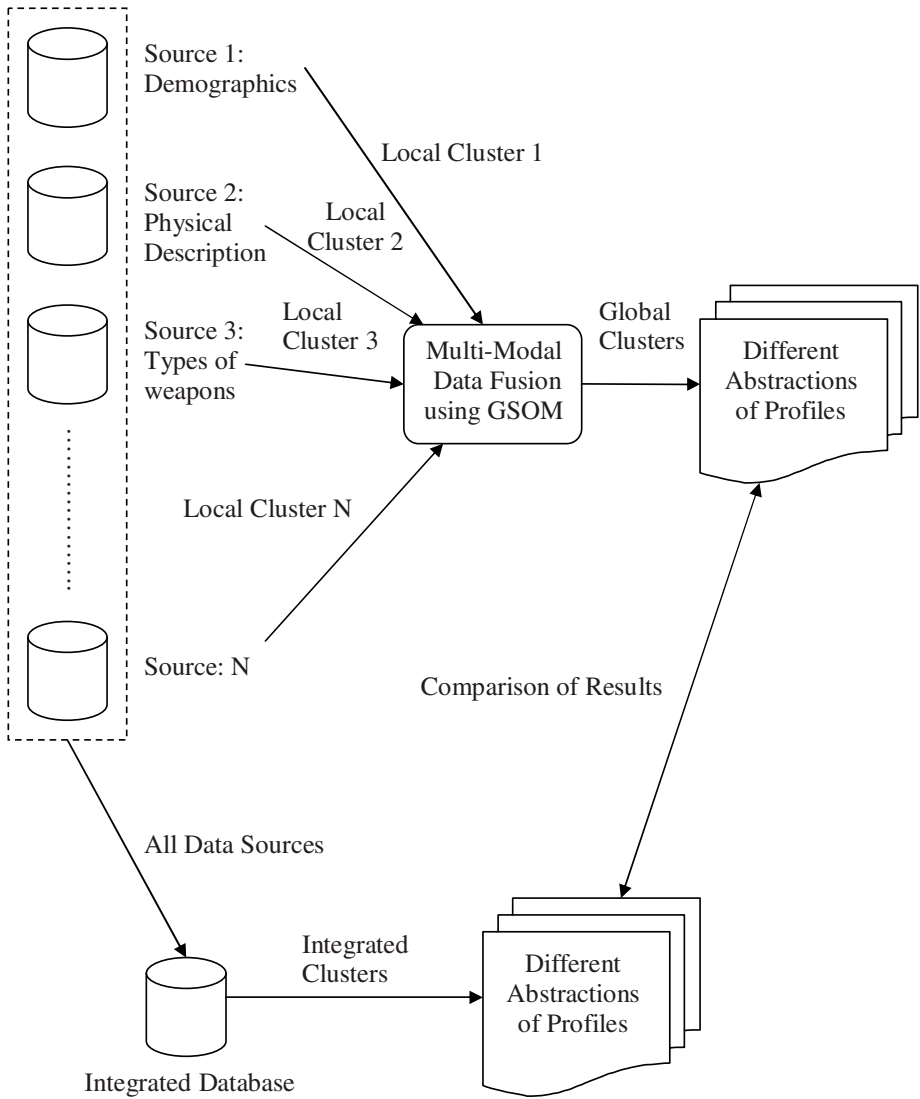
For the purpose of pilot studies, an integrated database is used. The integrated database considers all data sources and integrated clusters are produced to profile crime at different levels of abstractions. The difference between pilot studies and MMDF-C is the clustering is performed without the involvement of categorizing, selecting and fusing of the crime data. Thus, the integrated profiles are compared with profiles generated from multi-modal data fusion with the GSOM.

In addition, we agree with [52] that patterns discovered from single or integrated database may be incomplete and some patterns will be destroyed during the integration process. These patterns are overlooked and could be significant and valuable only if the multi-modal data are examined separately. The variety of multi-modal crime data always portray different patterns and therefore crime data should be examined in multi-modal perspectives for building a holistic crime profile.

### 5.2.2 The Different Phases in MMDF-C

The complementary processes - Data Fusion and Data Mining are depicted in Figure 5. In fact, MMDF-C is further divided into different correlated phases. It is obvious that the stages in this framework are reciprocal and flows from top to the bottom. This framework also includes feedback loops that repeat the data pre-processing, fusion and mining in order to fine tune crime profiles and improve the final profiles.

The framework consists of mainly of two processes - Multi-Modal Data Pre-processing and Data Fusion and Mining. Each of these main processes could be further broken down into several sub-phases. The different stages in MMDF-C are analogous to the two stages in FBI CSA. In particular, stage 2 in CSA is similar to how MMDF-C classifies the multi-modal of crime data according to data structure. Whereas stage 3 and stage 4 in CSA are corresponding to data mining and data fusion processes in MMDF-C.



**Fig. 4.** The global view of MMDF-C

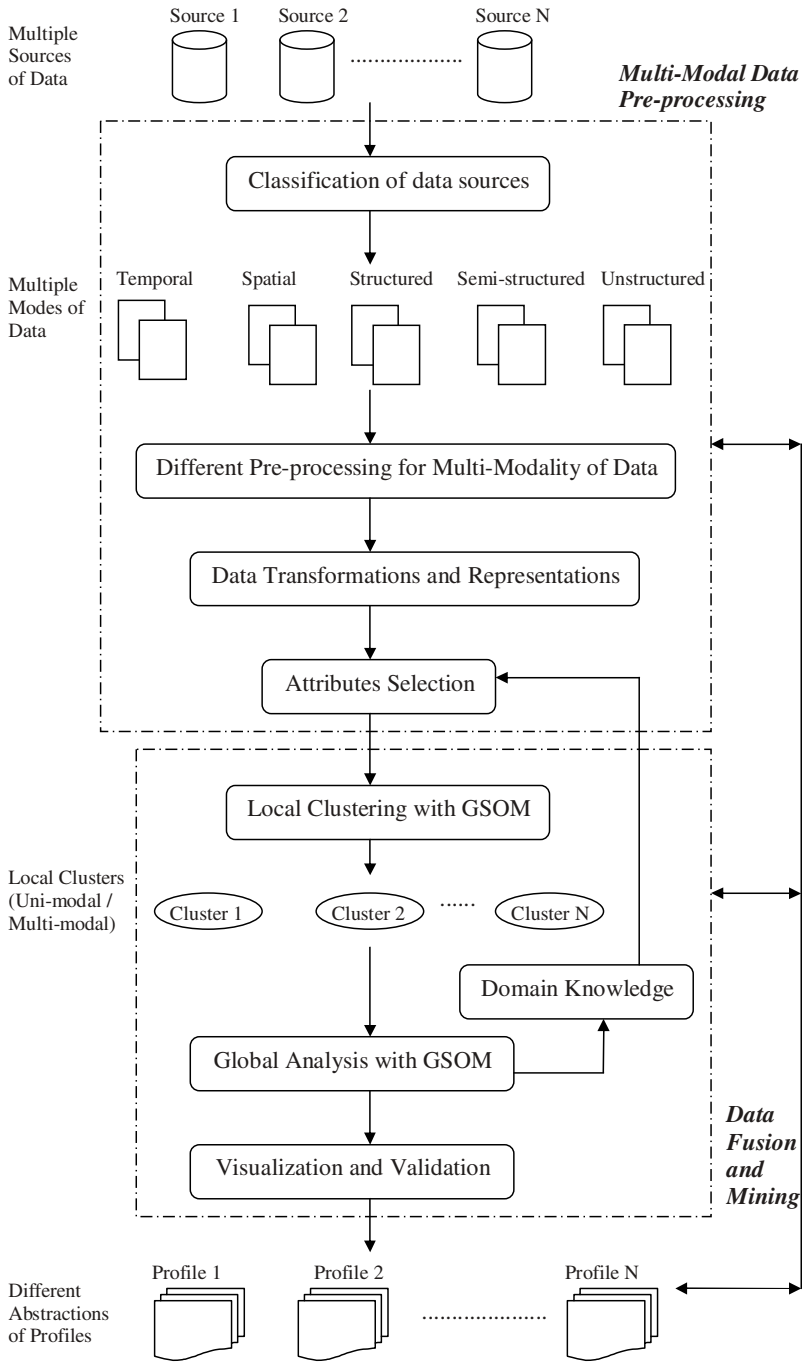


Fig. 5. The different stages in MMDF-C

The Multi-Modal Data Pre-processing includes the following phases:

- Classification of data sources
- Different Pre-processing for multi-modality of data
- Data Transformations and Representations
- Attribute Selection

Crime data could be differentiated according to the five main modalities and therefore different data pre-processing techniques are required to accommodate the five modalities of crime data. The pre-processed data is then transformed and represented. The purpose of transformations is to retain the characteristics of original data and simplify the data for better representations. Attribute selection is performed to extract attributes that are dominant and have higher predictability. Besides, domain knowledge is important and helpful in selecting the suitable and applicable attributes for the next step, Data Fusion and Mining.

The complementary processes involve the following 3 phases:

- Local Clustering with GSOM
- Global Analysis with GSOM
- Visualization and Validation

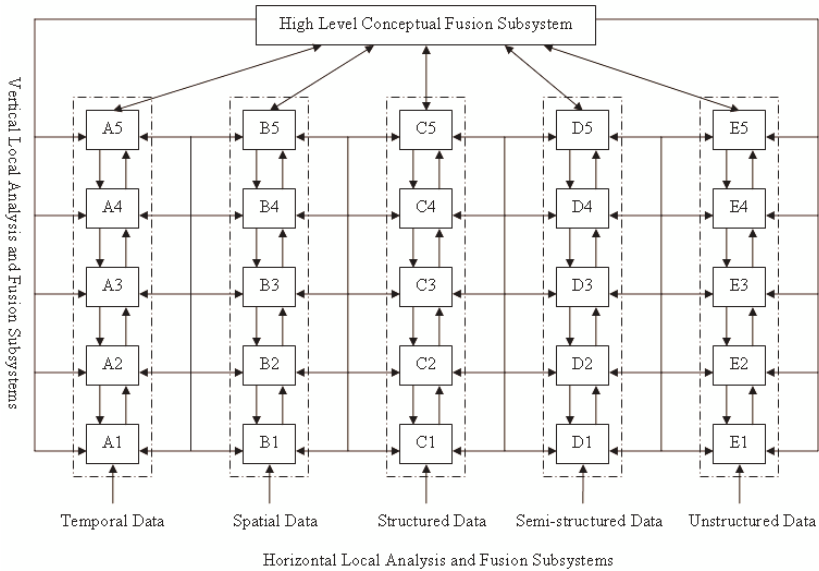
The merging of selected attributes allows local clusterings to be performed on data sources which exist in either uni-modal or multi-modal and the modalities of data sources are depending on how they are pre-processed previously. The local clusters are the inputs for the global analysis in which fusion is performed at global level for obtaining a complete picture of a crime, criminal and victim. Thus, this mimics the way human profilers combine available data to reach a conclusion. The synthesized results are visualized for better understanding of the data and also compared with the integrated database. The feedback from domain knowledge is forwarded back to the stage of attribute selection in order to improve the created profiles.

### 5.2.3 Data Fusion in MMDF-C

The architecture of MMDF-C is depicted in Figure 6. It shows that the fusion processes are flexible in which the of fusions could be performed at different directions. Specifically, the architecture illustrates that fusion can be operated at the three major levels:

- Vertical Local Analysis and Fusion Subsystems (Vertical Processing)
- Horizontal Local Analysis and Fusion Subsystems (Horizontal Processing)
- High Level Conceptual Fusion Subsystem (Global Processing)

In particular, Vertical Local Analysis and Fusion Subsystems involve five different vertical processing for the multi-modality of crime data. Each of the subsystems is processed separately and vertically and they include: (a) Temporal Data Fusion Subsystem; (b) Spatial Data Fusion Subsystem; (c) Structured Data Fusion Subsystem; (d) Semi-structured Data Fusion Subsystem; and (e) Unstructured Data Fusion Subsystem. The different stages for each of the subsystem are also specified as below.



**Fig. 6.** The architecture of Multi-Modal Data Fusion in Crime (MMDF-C)

(A) Temporal Data Fusion Subsystem involves the following five modules:

- A1** Temporal Metadata Derivation Module explores the multidimensionality of date and time so that implicit data could be obtained.
- A2** Time-Date Aggregation Module identifies the different granularities in temporal data and combines the time and date into a representative value.
- A3** Event-Time Distribution Module examines the preliminary distribution of the crime events in the sequence that spans across a period of time.
- A4** Time-Date Difference Module represents the temporal difference between crime events.
- A5** Time-Date Identification Module accounts for ranking and selecting prominent time-date data for high level conceptual fusion.

(B) Spatial Data Fusion Subsystem involves the following five modules:

- B1** Spatial Metadata Derivation Module explores the multidimensionality of spatial data in order to obtain the implicit spatial information.
- B2** Spatial Aggregation Module identifies the different granularities in spatial data and selects the best representative value.
- B3** Distance Module accounts for obtaining the difference between two spatial locations in matrix representation.
- B4** Spatial Distribution Module distributes spatial data to detect the hot spots or crime spots geographically.
- B5** Spatial Identification Module accounts for ranking and selecting prominent spatial data for high level conceptual fusion.



(C) Structured Data Fusion Subsystem involves the following five modules:

- C1** Structured Metadata Derivation Module explores the multidimensionality of structured data in order to obtain the implicit information among the data.
- C2** Model Definition Module defines the implicit information in descriptive and transactional models to allow examinations at different angles.
- C3** Data Type Registration Module classifies the different data structure in the context of data type such as numerical, categorical, nominal, etc.
- C4** Data Aggregation Module aggregates the different granularities found in the previous modules and transforms them into proper representations.
- C5** Structure Data Identification Module accounts for identifying the characteristics among the structured data for high level conceptual fusion.

(D) Semi-structured Data Fusion Subsystem involves the following five modules:

- D1** Content Classification Module classifies the different types of semi-structured input data to identify the different contents and purposes of the input data.
- D2** Attributes Identification Module identifies the important attributes of the semi-structured data.
- D3** Entities Identification Module identifies the related entities that contain in the structure data.
- D4** Semantic Identification Module identifies the implicit meaning between the attributes and entities of the structure data.
- D5** Semi-Structured Data Identification Module recognizes and selects relevant characteristics for high level conceptual fusion.

(E) Unstructured Data Fusion Subsystem involves the following five modules:

- E1** Data Categorization Module categorised unstructured data into different data formats.
- E2** Elementary Processing Module performs feature extraction on the different data formats for object recognition purposes.
- E3** Recognition Module recognised features extracted previously through processing on the different data formats in crime data.
- E4** Configuration Module reconfigures items recognised previously by connecting the items for semantic understanding.
- E5** Unstructured Data Identification Module selects the reconfigured items for high level conceptual fusion processing.

The Horizontal Local Analysis and Fusion Subsystems involve merging or fusing of the five multi-modal data cross-sectionally. Therefore, horizontal processing allows bottom up merging from low level to high level. It involves fusing any modes of data within the different levels of processing. On the other hands, the High Level Conceptual Fusion Subsystem takes input from all of the subsystems either horizontally or vertically for conceptual analysis. This high level conceptual processing involves information fusion where information comes from the local analysis from the original data. Therefore, the high level conceptual fusion subsystem could be regarded as a method to mimic human profiler.

## 6 The Demonstration of MMDF-C

This section demonstrates the functionalities discussed in section 4 and 5 with a small sample of crime data obtained from public records. In fact, we are demonstrating GSOM with three types of data structure which are publicly accessible. Thus, we have used structured data and temporal data to build the profile of crimes. In fact, this section intends to show the capability of hierarchical clustering in GSOM for producing multiple levels of abstraction in crime profiling. The multiple levels of abstraction are the outcomes of information fusion as they are also the summaries and representations of data and therefore serve the purpose of information extraction.

Apparently, there are many modules in MMDF-C but due to limited space and limited availability of crime data, we will only demonstrate and discuss some experimental results of certain modules. Again, we want to emphasize that the purpose of this section is to provide a glimpse of the practical value of the MMDF-C.

### 6.1 Description of Crime Data

The crime data is obtained from LA County, California State in US [2]. The crime type is murder incident that happened from January to May in 2007, which contains 124 releasable cases. Figure 7 depicted the distribution of the variety of murder cases. As this is a public record database, only limited sources of crime data are made available. The website in LA County Murder includes data sources about murder incident, victim, suspect and vehicle.

There are some missing and incomplete data in the sample data and we have dealt with the issues by doing some data pre-processing. For instance, some values were replaced with the mean values of the data source. Moreover, for the suspect data source, the only available information include name, age, gender and race. There is no information about the physical characteristics. Henceforth, we decided to use murder incident

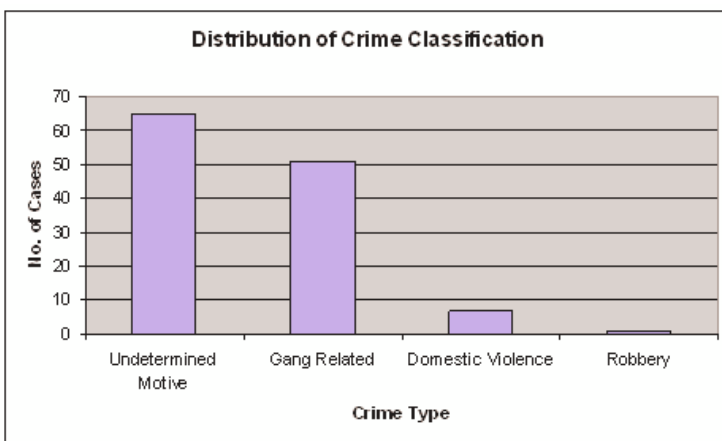


Fig. 7. The distribution of crime type from January to May 2007

and victim as input data sources and suspect and vehicle data are omitted. Table 3 shows the number of cases in terms of the case status and the different crime classification of the sample data.

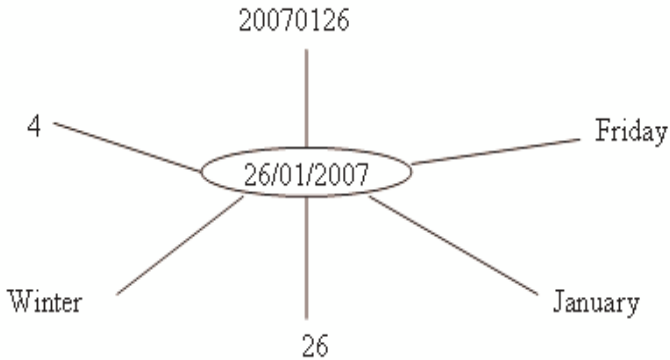
**Table 3.** The number of cases by case status and crime classification

<b>Crime Type \ Case Status</b>	<b>Undetermined Motive</b>	<b>Gang Related</b>	<b>Domestic Violence</b>	<b>Robbery</b>	<b>Total Cases</b>
Suspect Unknown	56	46	0	1	<b>103</b>
Suspect Deceased	0	0	2	0	<b>2</b>
Suspect Known (In Custody)	5	4	5	0	<b>14</b>
Suspect Known (Warrant Issued)	4	1	0	0	<b>5</b>
<b>Total Cases</b>	<b>65</b>	<b>51</b>	<b>7</b>	<b>1</b>	<b>124</b>

## 6.2 Different Metadata Derivations

Deriving metadata is important because patterns of certain data are implicit. For instance, [30] mentioned that weather season can be associated with vehicles theft during weekdays as more people pre-heat their cars during weekdays mornings before they go to work and thus create the opportunity for vehicle theft. Therefore crime data should also be derived to obtain different granularities. In [47], it has been discussed that metadata could be derived from date, vehicle identification numbers, names, addresses and so on. Additionally, [30] suggested that crime data should be examined creatively by transcending the usual analytical boundaries, where only single type of data or single granularity of data or information.

Within the two data sources (murder incident and victim information) that we examined, we further categorise them into temporal and structured data. To clarify the Temporal Metadata Derivation Module, as discussed in section 5.2.3, we can perform metadata derivation on temporal data that are obtained from the murder incident data source. The metadata derived from the temporal crime data, especially the date of crime, is depicted in Figure 8. In our proposed framework, we suggest that the different granularities of date information and time information could then aggregated in the next module.



**Fig. 8.** The metadata derived from date

### 6.3 Pre-processing and Data Transformation

At the stage of pre-processing and data transformation for murder incident and victim data sources, we have considered the m-of-n Remapping and Scaling Transformation discussed in [37]. Table 4 shows the attributes and the values as well as the labels for the attributes. We have pre-processed the temporal data, murder incidents and victim information separately to show the multi-modal crime data in the facets of data sources and data structures.

### 6.4 Experimental Results

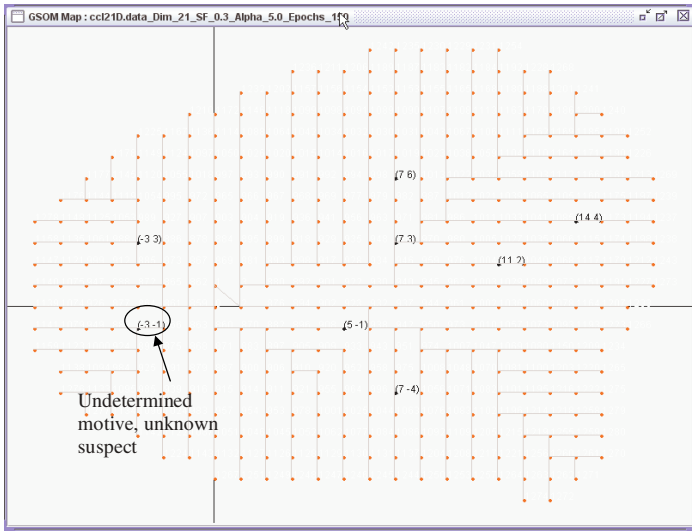
This section presents the experimental results and to show the crime patterns that has been discovered and visualized with GSOM. We initially discuss the results of temporal data and then structured data which consists of murder incident and victim information. The GSOM was used to perform clustering separately on the crime data based on parameterising the different values of Spread Factor (SF). Thus, the results has shown that GSOM is able to generate multiple levels of abstraction for crime patterns. These crime patterns could be viewed as the crime profiles and victim profiles, which are also outcomes of data fusion that serve the purpose of information extraction.

With murder incident, we have considered three attributes, namely crime classification, case status and the crime location. In Figure 9, it could be seen that there are eight clusters obtained with spread factor (SF) set at 0.3. To explain the different levels of abstraction for crime patterns in murder incident, we could take an example of the cluster at coordinate point (-3, 1). The crime pattern consists of murder cases with undetermined motive and unknown suspect.

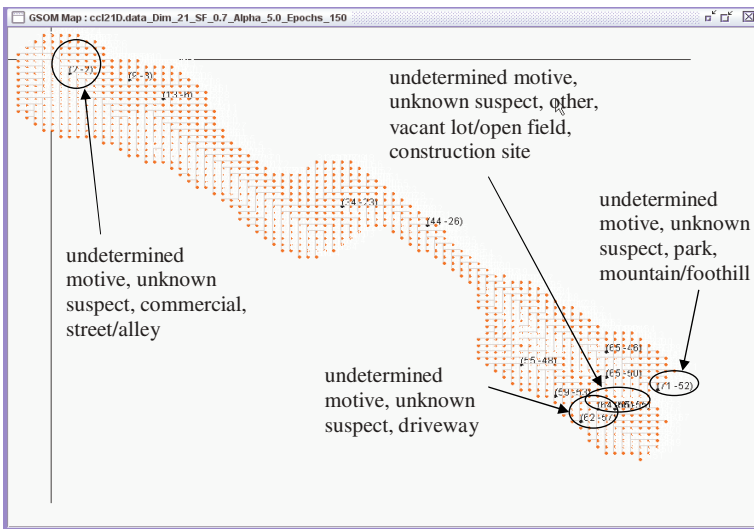
When spread factor (SF) was increased to 0.7 as seen in Figure 10, we noticed that more clusters are formed. It consists of murder cases with undetermined motive and unknown suspect that occurred at locations such as commercial, street/alley, driveway, park, mountains/foothills, other, vacant lot/open field and construction site.

**Table 4.** Data Mapping for the Attributes

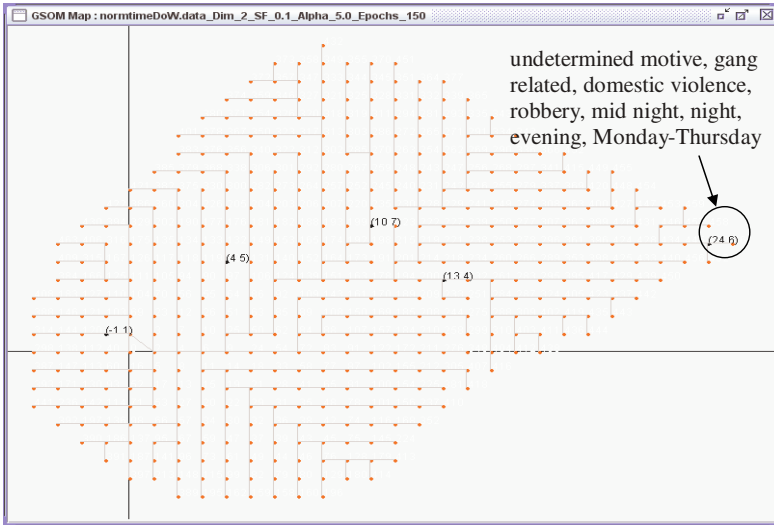
Data Source	No	Attributes	Values	Labels
Murder Incident	1	Crime Classification	Undetermined Motive	Obtain from m-of-n remapping
			Domestic Violence	
			Gang Related	
			Robbery	
	2	Case Status	Suspect (s)Unknown	
			Suspect (s) Known – In Custody	
			Suspect (s) Known – Warrant Issued	
			Suspect Deceased	
	3	Location	Residence	
			Commercial	
			School grounds	
			Driveway	
			Sidewalk	
			Street / Alley	
			Freeway / Highway	
			Other	
			Park	
			Mountains / Foothills	
			Rural / Isolated	
	Construction Site			
	Vacant lot / Open Field			
4	Time	Morning	1	
		Afternoon	2	
		Evening	3	
		Night	4	
		Mid-Night	5	
5	Date	Monday	1	
		Tuesday	2	
		Wednesday	3	
		Thursday	4	
		Friday	5	
		Saturday	6	
		Sunday	7	
Victim Information	1	Cause of Death	Gun Shot Wound(s)	1
			Stabbing	2
			Undetermined	3
			Blunt Force Trauma	4
			Strangulation/ Asphyxiation	5
	2	Gender	Male	1
			Female	0
	3	Race	White	1
			Black	2
			Hispanic	3
Asian / Pacific Islander			4	
Others			5	
	4	Age	Numerical (e.g. 27)	Scaling Transformation



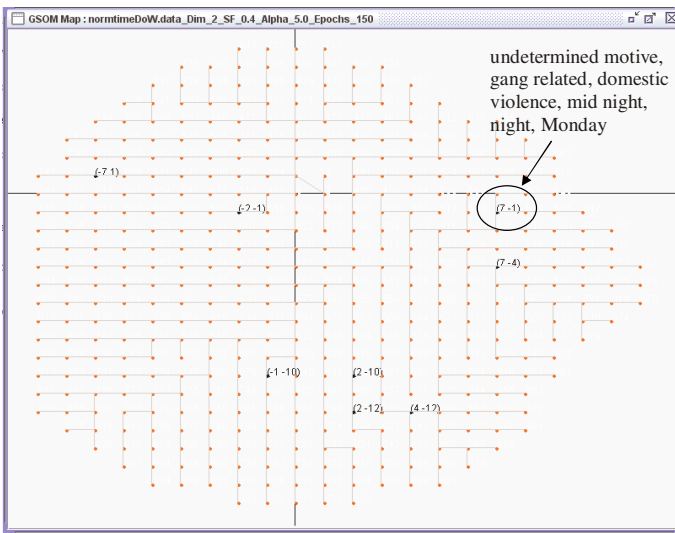
**Fig. 9.** The crime pattern visualization of murder incident when SF = 0.3



**Fig. 10.** The crime pattern visualization of murder incident when SF = 0.7

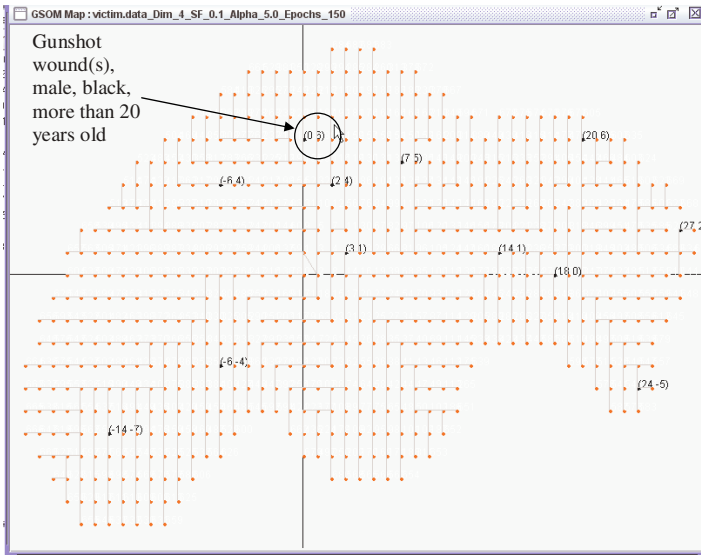


**Fig. 11.** The crime pattern visualization of temporal analysis when SF = 0.1

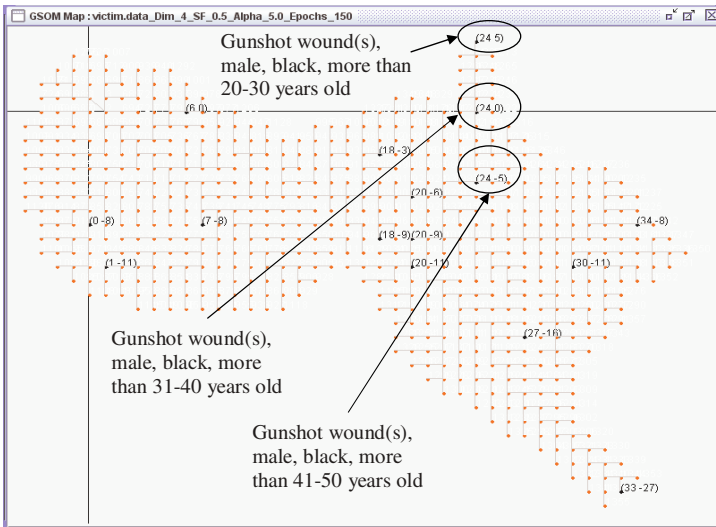


**Fig. 12.** The crime patterns visualization of temporal analysis when SF = 0.4

With temporal data, we simply decided to derive the metadata by representing time in terms of morning, afternoon, evening, night and midnight as well taking day of the week (DoW) for date. The temporal analysis is able to illustrate the occurrence of crime at multiple levels of granularity. In addition, they could be correlated to crime type, as shown in Figure 11 and Figure 12.



**Fig. 13.** The crime pattern visualization of victim when SF = 0.1



**Fig. 14.** The crime pattern visualization of victim when SF = 0.5

Whereas with the victim data source, the crime patterns at high level of abstraction has shown only information about the suspect cause of death. The spread factor (SF) was increased progressively to show the crime patterns in terms of its gender, race and age group. The crime patterns from victim information, or it could be the victim profiles, are illustrated in Figure 13 and Figure 14.



## 7 Conclusion

We have described and identified the multi-modality issues in crime domain. Further more, we proposed our framework, MMDF-C in relation to the identified issues in crime data and also for crime profiling. We have also demonstrated the capabilities of GSOM for extracting crime patterns and addressing the multi-modality issues. In brief, the proposed framework and adoption of GSOM to visualise and represent multi-modal crime data have demonstrated that multi-modal crime data has been fused by parameterising the different values of spread factor (SF).

These initial experiments only demonstrate a small proportion of the overall conceptual model, as discussed in section 5. Nevertheless, it provides a glimpse of the use of artificial neural networks based techniques to support the proposed framework and to perform crime profiling computationally. Besides, the artificial neural networks, particularly GSOM, is a soft computing approach that mimic human profiling processes. Although the crime patterns shown are different from the actual crime profiles generated by human profilers, it has proven that it is possible to build a systematic approach for crime profiling.

In future, the crime profiles obtained by GSOM could be further enhanced if more crime data could be the inputs. Specifically, the proposed framework will look into the unstructured and semi-structured crime data. That means the future research will also consider the text mining techniques on unstructure data sources such as narrative reports, social media, etc. Thus, this opens up more opportunities in data mining and data fusion based on results obtained from different data sources and data structures.

## References

1. FBI Uniform Crime Reporting Systems, <http://www.fbi.gov/ucr/ucr.htm>
2. Los Angeles County Murder Cases, <http://www.lacountymurders.com/caseinfo2.cfm>
3. Public Practice Local Crime Profile, [http://www.publicpractice.net/crime\\_portrait.htm](http://www.publicpractice.net/crime_portrait.htm)
4. Adderley, R., Musgrove, P.: Police crime recording and investigation systems a user's view. *International Journal of Police Strategies & Management* 24(1), 100–114 (2001)
5. Adderley, R., Musgrove, P.B.: Data mining case study: Modeling the behavior of offenders who commit serious sexual assaults. In: *KDD*, pp. 215–220 (2001)
6. Adderley, R., Musgrove, P.B.: Modus operandi modelling of group offending: A data mining case study. *International Journal of Police Science and Management* 5(4), 265–276 (2003)
7. Alahakoon, D., Halgamuge, S.K., Srinivasan, B.: A self growing cluster development approach to data mining. In: *IEEE Conference Systems, Man and Cybernetics*, pp. 2901–2906 (1998)
8. Alahakoon, D., Halgamuge, S.K., Srinivasan, B.: Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks* 11(3), 601–614 (2000)
9. Baumgartner, K.C., Ferrari, S., Salfati, C.G.: Bayesian network modeling of offender behavior for criminal profiling. In: *IEEE Conference on Decision and Control*, pp. 2702–2709 (2005)

10. Bekerian, D.A., Jackson, J.L.: Chapter12 - critical issues in offender profiling. In: Jackson, J.L., Bekerian, D.A. (eds.) *Offender Profiling: Theory, Research and Practice*, pp. 209–220. John Wiley & Sons (1997)
11. Bloch, I., Hunter, A., Appriou, A., Ayoun, A., Benferhat, S., Besnard, P., Cholvy, L., Cooke, R., Cuppens, F., Dubois, D., Fargier, H., Grabisch, M., Kruse, R., Lang, J., Moral, S., Prade, H., Saffiotti, A., Smets, P., Sossai, C.: Fusion: General concepts and characteristics. *International Journal of Intelligent Systems* 16(10), 1107–1134 (2001)
12. Brown, D.E.: Data mining, data fusion, and the future of systems engineering. In: *IEEE International Conference on Systems, Man and Cybernetics*, pp. 26–30 (2002)
13. de Bruin, J.S., et al.: Data mining approaches to criminal career analysis. In: *International Conference on Data Mining (ICDM)*, pp. 171–177 (2006)
14. Charles, J.: Ai and law enforcement. *IEEE Intelligent Systems* 13(1), 77–80 (1998)
15. Chen, H., Chung, W., Qin, Y., Chau, M., Xu, J.J., Wang, G., Zheng, R., Atabakhsh, H.: Crime data mining: An overview and case studies. In: *National Conference on Digital Government Research (dg.o)*, pp. 1–5 (2003)
16. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: A general framework and some examples. *IEEE Computer* 37(4), 50–56 (2004)
17. Chen, M., Han, J., Yu, P.S.: Data mining: An overview from a database perspective. *IEEE Transaction on Knowledge and Data Engineering* 8(6), 866–883 (1996)
18. Chu, H.C., Deng, D.J., Park, J.H.: Live data mining concerning social networking forensics based on a facebook session through aggregation of social data. *IEEE Journal on Selected Areas in Communications* 29(7), 1368–1376 (2011)
19. Dasarathy, B.V.: Information fusion, data mining, and knowledge discovery. *Information Fusion* 4(1), 1 (2003)
20. Elmaghraby, A.S., Kantardzic, M.M., Wachowiak, M.P.: Data mining from multimedia patient records. In: *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*. *Massive Computing*, vol. 6, ch. 16, pp. 551–595. Springer, US (2006)
21. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. *Communications of The ACM* 39(11), 27–34 (1996)
22. Geradts, Z., Keijzer, J.: The image-database rebezo for shoeprints with developments on automatic classification of shoe outsole designs. *Forensic Science International* 82(1), 21–31 (1996)
23. Helbicha, M., Hagenauera, J., Leitnerb, M., Edwardsc, R.: Exploration of unstructured narrative crime reports: an unsupervised neural network and point pattern analysis approach. *Journal of Cartography and Geographic Information Science* 40(4), 1–11 (2013)
24. Howitt, D.: *Forensic and Criminal Psychology*. Pearson Education (2002)
25. Kasabov, N.: Evolving systems for integrated multi-modal information processing. In: *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machine*, ch. 13, pp. 257–271. Springer, London (2003)
26. Kocsis, R.N.: An empirical assessment of content in criminal psychological profiles. *International Journal of Offender Therapy and Comparative Criminology* 47(1), 37–46 (2003)
27. Kohonen, T.: *Self Organizing Maps*. Springer (2001)
28. Lin, S., Brown, D.E.: Criminal incident data association using the OLAP technology. In: Chen, H., Miranda, R., Zeng, D.D., Demchak, C.C., Schroeder, J., Madhusudan, T. (eds.) *ISI 2003*. LNCS, vol. 2665, pp. 13–26. Springer, Heidelberg (2003)
29. Martin, C., grosse Deters, H., Nattkemper, T.W.: Fusion biomedical multi-modal data for exploratory data analysis. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) *ICANN 2006*. LNCS, vol. 4132, pp. 798–807. Springer, Heidelberg (2006)
30. McCue, C.: *Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis*. Butterworth-Heinemann (2007)

31. Memon, Q.A., Mehboob, S.: Crime investigation and analysis using neural nets. In: IEEE 7th International Multi Topic Conference (INMIC), pp. 346–350 (2003)
32. Mena, J.: Investigative Data Mining for Security and Criminal Detection. Butterworth-Heinemann (2003)
33. Oatley, G., Ewart, B., Zeleznikow, J.: Decision support systems for police: Lessons from the application of data mining techniques to soft forensic evidence. *Artificial Intelligence and Law* 14(1-2), 35–100 (2006)
34. Pastra, K., Saggion, H., Wilks, Y.: Extracting relational facts for indexing and retrieval of crime-scene photographs. *Knowledge-Based Systems* 16, 313–320 (2003)
35. Pinizzotto, A., Finkel, N.: Criminal personality profiling: An outcome and process study. *Law and Human Behaviour* 14(3), 215–233 (1990)
36. van der Putten, P., Kok, J.N., Gupta, A.: Why the information explosion can be bad for data mining, and how data fusion provides a way out. In: 2nd SIAM International Conference on Data Mining (SDM) (2002)
37. Pyle, D.: *Data Preparation for Data Mining*. Morgan Kaufmann (1999)
38. Ross, A., Jain, A.K., Qian, J.-Z.: Information fusion in biometrics. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 354–359. Springer, Heidelberg (2001)
39. Siegel, L.J.: *Criminology: Theories, Patterns, and Typologies*. Thompson Wadsworth (2007)
40. Strano, M.: A neural network applied to criminal psychological profiling: An italian initiative. *International Journal of Offender Therapy and Comparative Criminology* 48(4), 495–503 (2004)
41. Torra, V.: Trends in information fusion in data mining. In: Torra, V. (ed.) *Information Fusion in Data Mining*. STUDEFUZZ, vol. 123, pp. 1–6. Springer, Heidelberg (2003)
42. Torra, V., Narukawa, Y.: *Modelling Decisions: Information Fusion and Aggregation Operators*. Springer, Berlin (2007)
43. Turvey, B.: *Criminal Profiling: An Introduction to Behavioral Evidence Analysis*. Academic Press (1999)
44. Fayyad, U., Piatetsky-Shapir, G., Smyth, P.: From data mining to knowledge discovery in databases. In: American Association for Artificial Intelligence (AAAI), pp. 37–54 (1996)
45. Waltz, E.L.: Information understanding: Integrating data fusion and data mining processes. In: IEEE International Symposium on Circuits and Systems (ISCAS 1998), pp. 553–556 (1998)
46. Wang, G., Chen, H., Xu, J., Atabakhsh, H.: Automatically detecting criminal identity deception: An adaptive detection algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans* 36(5), 988–999 (2006)
47. Westphal, C., Blaxton, T.: *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley & Sons (1998)
48. Wickramasinghe, L.K., Alahakoon, L.D.: Dynamic self organizing maps for discovery and sharing of knowledge in multi agent systems. *International Journal on Web Intelligence and Agent Systems* 3(1), 31–47 (2005)
49. Xu, J., Chen, H.: Criminal network analysis and visualization. *Communications of the ACM* 48(6), 101–107 (2005)
50. Xue, Y., Brown, D.E.: A decision model for spatial site selection by criminals: A foundation for law enforcement decision support. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 33(1), 78–85 (2003)
51. Yager, R.R.: A framework for multi-source data fusion. *Information Sciences* 163, 175–200 (2004)
52. Zhang, S., Zhang, C., Wu, X.: *Knowledge Discovery in Multiple Databases*. Springer (2004)

# Anthropometric Measurement of North-East Indian Faces for Forensic Face Analysis

Kankan Saha<sup>1</sup>, Mrinal Kanti Bhowmik<sup>1</sup>, and Debotosh Bhattacharjee<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Tripura University (A Central University),  
Suryamaninagar - 799022, Tripura, India

kankansaha@yahoo.com, mkb\_cse@yahoo.co.in

<sup>2</sup>Department of Computer Science and Engineering,  
Jadavpur University, Kolkata - 700032, India

debotosh@ieee.org

**Abstract.** This chapter presents a study of the facial structural differences between the various tribes and non-tribes of the north-eastern region of India. Distances between the various facial feature points are measured using the face images of the newly created database, named as, the Department of Electronics and Information Technology-Tripura University (DeitY-TU) face database. After careful observation of the human facial structure and conducting primary anthropometric measurements, a comparative study between different Mongolian tribes and the non-tribes have been done for face identification through the determination of resemblance, which may be useful in determining the facial characteristics of the criminals and terrorists of the north-eastern region of India, and also for strengthening forensic analysis to stop illegal emigrations.

**Keywords:** Mongolian faces, north-east Indian faces, DeitY-TU face database, fiducial point, anthropometric measurement, forensic application, face identification, ethnic group detection.

## 1 Introduction

Human face is an intriguing subject which has gained attention of countless artists, poets and scientists. Recognizing the face automatically has become one of the most active and widely used techniques because of its reliability in the process of verifying a person's identity. It is becoming more important in terms of security and privacy as face images can be acquired with or without any cooperation from the person of interest. Security, now-a-days, is given the top priority to counter the possible threats from terrorists and criminals, which includes improvising new dimensions of technological advancement for forensic investigations as well. Forensic science is the application of one or more scientific branches to investigate and establish facts of interest in relation to criminal or civil law [1]. Identifying an individual is generally based on a comparison of an unknown and a known and is used when there is someone to compare the evidence against or it is used to create the potential for comparing the evidence against a suspect in the future.

Considering the terrorist activities in India, the north-eastern region occupies one of the prime positions [2]. In this context, we may contribute an alternative perspective to increase the security and its developments for the north eastern states of India. So our ongoing research on face identification may be useful for the illegal migration as well as information security enhancement in homeland security. The primary aim of this study is to conduct anthropometric measurements on the different face images of the different north-eastern region of India for identifying the region based human faces.

In the field of face recognition, detection of the most important facial feature points and measurement of the distances between them are being developed by the application of anthropometric study. Anthropometry was first introduced to the field of forensics to identify individual criminals by their physical characteristics. The purpose of anthropometric analysis is to study the variation of human physical characteristics. It is also a key technique to find out the differences and similarities among numerous races.

At present, there are no available databases for the north-east Indian people. Recently, we are developing a database named as the Department of Electronics and Information Technology-Tripura University (DeitY-TU) face image database, which contains images of a different tribe and non-tribe people of different races, especially people belonging to Mongolian origin [3, 4] obtained from the north-eastern region of India. This region has been occupied by several streams of the Mongoloid people who came from the north and the east at different periods [5]. The diverse Mongoloid groups in the course of time settled down in different habitats and ecological settings of the north eastern region crystallized into separate entities which are referred to as tribes today [5, 6].

In this chapter, we have already collected face images of different Mongolian people like: Tripura, Reang, Chakma, Debbarma, Jamatia, Darlong, Mog, Halam etc. from the different states, and using these images we have measured the anthropometric distances between several feature points in the face, and also tried to find out the facial structural differences between the different tribe and non-tribe male/female people of the different north-eastern states of India. The aim of this study is to carry out a comparison of anthropometric values of Mongolians in the north-eastern region of India of similar socioeconomic status with special reference to five states: Assam, Mizoram, Tripura, Nagaland and Manipur.

The rest of this chapter is divided into seven sections covering existing techniques for the measurement of anthropometric values, anthropometry in forensics, people of the north-eastern region of India, creation of the DeitY-TU face database, anthropometric measurements of the face images, analysis and observation of the findings, and conclusion.

## **2 Existing Anthropometric Measurement Techniques**

Popular methods to measure craniofacial anthropometry include direct measurements on the surface of the skin, radiographic cephalometry and photographic approaches. The direct measurement method has several advantages, such as non-invasiveness,

technical simplicity and low cost. Though, there is the risk of examiner subjectivity and as a result it can produce poor outcome and may not be completely reliable every time [7]. Again, repeated measures are not possible always as all the subjects may not be available for the next time. Radiographic cephalometry is suitable for the observation of hard tissue such as bone, and can assess many points, angles and planes [8]. But, it is a relatively high cost solution, and it involves a threat of exposing the subjects to radiation during measurement. For these reasons, the photographic approach gains more preference over the other methods and evolves as a useful method for measuring facial anthropometric parameters. Photographs can be obtained easily and then, can be stored permanently for future use [8]. This approach using photographs is suitable for the analysis of facial features and is adaptable to meet the specific measurement needs of different investigators [9].

In [10], Farkas conducted 25 measurements on head and face based landmark points to examine three racial groups: North American Caucasian, African-American, and Chinese. Im et al. [11] represented about the study of the genetic effect and quantitative trait locus (QTL) of seven traits, based on anthropometric measurements over a population of Mongolian race. Luximon et al. [12] used 3D scanning technology to study the use of traditional facial anthropometry. In traditional anthropometry, a lot of numerical dimensions are measured in order to portray the different face shapes. Li et al. [13] presented a novel pose estimation method by improving the traditional geometrical design method for determining feature characteristic. For this, they built an isosceles triangular model of face based on the four eye-corners and subnasal.

Ngeow and Aljunid [14] established the craniofacial anthropometric norms of the young adult (18-25 years) Malaysian Indian. They conducted 22 linear measurements for twice, from 28 landmarks over six craniofacial regions using standard anthropometric instruments. A third reading was taken if the earlier two measurements were not consistent enough. They adopted the methodology and evaluation of indices of the craniofacial region from [15]. Landmarks were marked directly on the skin to avoid errors in locating them.

Jahanshahi et al. [16] used classic cephalometry for conducting some cross-sectional studies to compare the face shapes in Fars and Turkman ethnic groups of normal newborns and 17–20 years old males and females in Gorgan (North of Iran).

Sohail and Bhattacharya [17] presented an automatic technique to detect 18 facial feature points, which were mostly around the eyes, eyebrows, nose and mouth. For this, they used a statistically developed anthropometric face model. They isolated the different facial feature regions and located the different regions by using the distance between the two eye centers as the principal parameter. They conducted these measurements on 300 frontal face images of more than 150 subjects and used the obtained proportions to build the anthropometric face model.

Although anthropometric measurement of the face provides useful information about the facial structure, it has rarely been used in automatic detection and localization of different facial features [17]. In the field of forensics, uncertainty in the anthropometric measurements of different facial proportions has become the focus of attention, as forensic results need to be accurate and reliable. Variations in facial

anthropometric measurements may happen from the use of different operators, due to different subjects, and even due to taking multiple measurements of different photographs of the same subject [1].

### 3 Anthropometry in Forensics

Anthropometry is the systematic collection and correlation of measurements of the physical sizes and shapes of the human body. It can also be depicted as a hallmark technique that deals with the study of body proportion and absolute dimensions that vary widely with age and sex within and between racial groups.

Face anthropometry is also an important technique. Face shape is dynamic, due to the many degrees of changes that occur due to the various expressions and movements of the human face. Again, the variability of face shape is also highly limited by both genetic and biological constraints [18]. Anthropometric studies of a human face have a long history but most of those are limited to linear measurements taken directly by using calipers and tapes on the human face [17]. It involves making accurate and standard measurements, so that the various differences among the human face shapes can be described objectively.

Anthropometry, first used in the 19<sup>th</sup> century, has since been substituted with more perceptive methods of identification. It achieves varying results, which is an indication that more research is needed to ensure a reliable forensic identification method, but still it is being used by many researchers for various purposes like, face recognition, facial structure analysis etc. because of its undeniable advantages. Anthropometric measurements are portable, non-invasive, inexpensive, useful in field studies, and comprehensible to communities at large [19]. They generate data that can be evaluated numerically and used to compare across populations.

From a forensic perspective, facial recognition is important and powerful tool in various scenarios [20] like, (1) searching faces from a crowd scene, which have been obtained from a given database; (2) picking out the best face match for a probe image from previously acquired face images; (3) producing supporting evidence to support or reject the assumption that a person in an image is the suspect in custody.

The most common recognizable way to identify someone is from their face and therefore, the methods of identification that involves face, are all very important to forensic science [1]. Whether the evidence available is from a video recording, still image or eye witness, the use of facial identification procedures is vital to the investigation of crime.

Anthropometry can be used for forensics in certain circumstances to facilitate comparison of a photo of a suspect with the potential criminal disclosed in surveillance video recording [1]. Though, anthropometry does not provide the same success rate in identifying a subject as DNA or fingerprinting, but these evidences are not always left at crime scenes. Sometimes the only evidence available relating to an crime is from surveillance videos and research was needed to provide acceptance to anthropometry as a viable method of identification or as a helping tool for exploring facial characteristics of the individuals for forensic identification.

But an automatic anthropometric measurement technique for human faces requires accurate detection and identification of the various facial feature points. Identification of facial feature points plays an important role in many facial image applications like human computer interaction, video surveillance, face detection, face recognition, facial expression classification, face modeling and face animation. A large number of approaches have already been attempted towards addressing this problem, but complexities added by circumstances like inter-personal variation (i.e. gender, race), intrapersonal changes (i.e. pose, expression) and inconsistency of acquisition conditions (i.e. lighting, image resolution) have made the task quite difficult and challenging.

In a study by Farkas et al. [10] based on measurements found reliable in one of his previous research, landmarks that were able to be seen clearly on the photo were used to create age progression photographs for missing children using anthropometry.

Forensic anthropometry has traditionally been considered to be controversial and unproven method of identification [1]. The general feeling within the forensic science community is that there are too many factors which make this method subjective and that even when high quality photographs taken in a controlled setting are available, factors such as lighting, head position, camera position, and operator experience may all contribute to the inaccuracy of this technique for identification purposes.

## **4 North-East India and Its Inhabitants**

North East India is the homeland of a large number of ethnic groups who came from different directions at different historic times. These groups belong to different racial community, speak different languages and have mixed socio-cultural traditions [5]. This region has been occupied by several streams of the Mongoloid people who came from the north and the east at different periods. The Australoids came to this region before the coming of the Mongoloids who partially or fully absorbed the Australoid strains [21, 22]. The physical features of different tribes of North East India suggest that the Australoid elements are present in some of the tribes. It has been stated that long ago one section of the Indo-Mongoloids spread over the whole of the Brahmaputra valley, North Bengal and East Bengal (now Bangladesh) giving rise to various tribal groups inhabiting this region [23]. The diverse Mongoloid groups in the course of time settled down in different habitats and ecological settings of the north eastern region crystallized into distinct entities which are referred to as tribes today [6, 24].

The overwhelming majority of the people living in North East India are Hindus (60.93%). The second largest religious group is the Muslims, who constitute 21.55% of the total population of North East India. The Christians constitute 13.63% of the total population of North East India. The Buddhists and Jains are not dominant in any of this region of India. North East India represents a sort of ethnological transition zone between India and neighboring China, Tibet, Burma and Bangladesh.

In the seven states of North East India, the percentages of tribal population vary significantly. In the states of Assam, Manipur and Tripura, the percentages of tribal population to the total population of the respective states are 12.82%, 34.41% and 30.95%. In Arunachal Pradesh, Meghalaya, Mizoram and Nagaland the percentages of tribal population to the total population of the respective states are quite high. In Mizoram, the tribals constitute 94.75% of the total state's population.



## 5 Department of Electronics and Information Technology- Tripura University (DeitY-TU) Face Database

All the anthropometric measurements are being conducted on the frontal neutral face images of the DeitY-TU face database, which is a visual face image database under development. A database is being created with the face images of the different tribe as well as non-tribe people of the seven North-Eastern states of India. In this database, faces have been captured in four illumination conditions and with eight expressions and images are being clicked concurrently from five different viewpoints [3, 4].

**Table 1.** DeitY-TU face image database statistics

State	Total Images	No. of Images per Person	Total Persons	Male	Female	Tribe	Non-tribe
Mizoram	10,640		112	62	50	112	0
Assam	10,165		107	46	61	14	93
Tripura	9,500	95	100	49	51	34	66
Nagaland	9,595		101	57	44	101	0
Manipur	9,880		104	80	24	29	75
Total	49,780	95	524	294	230	290	234

**Table 2.** Tribes and non-tribes of north-east Indian states collected from the DeitY-TU face database

State	Tribes	Non-Tribes
Mizoram	Bawitlung, Bawlte, Chawngthu, Chenkual, Fanai, Hrangchal, Jinhlong, Khawlhiring, Laichhak, Miller, Ralte, Sailo, Zadeng	-
Assam	Borgohain, Kachari, Rabha, Lalung, Basumatary (BodoKachari, Koch)	Bora, Baishya, Bordoloi, Das, Dewan, Hazarika
Tripura	Debbarma, Reang, Jamatia, Darlong, Tripura, Chakma, Mog, Rupini	Dey, Saha, Bhowmik, Majumder, Chakraborty, Das, Nandi, Biswas, Debnath, Mahajan, Shil
Nagaland	Khiamniungan, Rengma, Aonaga, Angami, Phom, Lothanagal, Suminaga, Tangkhul, Mao, Sangtam, Konyak, Zeliang, Pochury	-
Manipur	Kabui, Tangkhul, Liangmai, Kom, Thadou, Rongmei, Anal, Maring, Liangmai, Poumai, Mao, Naga	Meitei

All images are being captured, against a homogeneous black background of 8.7 ft×6.5ft dimension to prevent light reflection. Till now, 49,780 images of total 524 persons have been collected from five north-eastern states: Mizoram, Assam, Tripura, Nagaland and Manipur. An overall scenario of the collected images with the number of males, females, tribes, and non-tribes from the different states for the DeitY-TU face database is shown in Table 1. Table 2 shows the different tribe and non-tribe people collected from these states. Some frontal sample images with the neutral expression and full illumination are shown in Fig. 1.



**Fig. 1.** Sample neutral frontal face images in full illumination condition of DeitY-TU face database

## 6 Anthropometric Measurements of DeitY-TU Face Images

### 6.1 Selection of Image Data

Here, facial anthropometric analysis of the different races including the Mongolian people have been done using the visual face image database being collected from the seven north-eastern states of India. However, the faces have been captured with multiple variations of illumination, expression and pose, yet for anthropometric measurements, we have considered only the neutral frontal face images captured in full illumination condition. The frontal images of this database are of 1936×1288 pixels dimension and have been clicked from a distance of 4.5ft from the user. To have an

apparent perceptible of the facial differences between the tribe and non-tribe males as well as females, we tried to select at least 10 male tribes, 10 female tribes, 10 male non-tribes, and 10 female non-tribes from each of the five states. But, due to unavailability of sufficient image data in all the states, we have adjusted our selection as shown in Table 3. After careful review, photographs of 200 subjects were considered for distance measurement and analysis.

**Table 3.** Number of persons from different states for which anthropometric measurements have been conducted

State	Male Tribe	Female Tribe	Male Non-Tribe	Female Non-Tribe
Mizoram	20	20	0	0
Assam	6	7	14	13
Tripura	10	10	10	10
Nagaland	20	20	0	0
Manipur	10	5	10	15
Total	66	62	34	38

## 6.2 Selection of Landmark Points and Distances

The most important facial features those are responsible for constructing the basic structure of the human face, are eyes, nose, mouth, eyebrows etc. In this chapter, based on the above mentioned facial parts, we have selected 4 unilateral, and 8 bilateral, consisting a total of 20 landmark points, which will be included in subsequent anthropometric studies. Bilateral landmarks are located on both sides of the face. All the 20 landmark points are shown in Fig. 2.

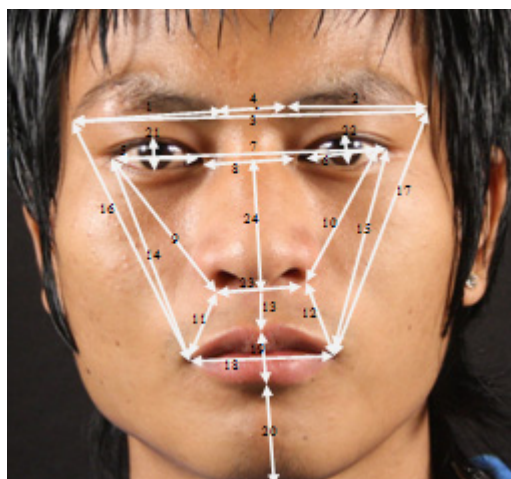


**Fig. 2.** Sample image with twenty landmark points of DeitY-TU database

From these landmarks, a total of 24 linear measurements were selected for comparison of images, which are listed in Table 4 and a sample image with the distances to be measured is shown in Fig. 3.

**Table 4.** Anthropometric measurements employed in this study

Sl. No.	Distances	Codes used for the Distances
1.	right eyebrow right corner to right eyebrow left corner	REbR-REbL
2.	left eyebrow right corner to left eyebrow left corner	LEbR-LEbL
3.	left eyebrow right corner to right eyebrow left corner	LEbR-REbL
4.	left eyebrow left corner to right eyebrow right corner	LEbL-REbR
5.	right eye right corner to right eye left corner	RER-REL
6.	left eye right corner to left eye left corner	LER-LEL
7.	left eye right corner to right eye left corner	LER-REL
8.	left eye left corner to right eye right corner	LEL-RER
9.	right nose corner to right eye right corner	RN-RER
10.	left nose corner to left eye left corner	LN-LEL
11.	right nose corner to right mouth corner	RN-RM
12.	left nose corner to left mouth corner	LN-LM
13.	sub-nasal point to upper lip outer middle	SN-ULOM
14.	right eye right corner to right mouth corner	RER-RM
15.	left eye left corner to left mouth corner	LEL-LM
16.	right eyebrow right corner to right mouth corner	REbR-RM
17.	left eyebrow left corner to left mouth corner	LEbL-LM
18.	right mouth corner to left mouth corner	RM-LM
19.	upper lip outer middle to lower lip outer middle	ULOM-LLOM
20.	lower lip outer middle to chin middle	LLOM-CM
21.	left eye upper lid midpoint to left eye lower lid midpoint	LEUM-LELM
22.	right eye upper lid midpoint to right eye lower lid midpoint	REUM-RELM
23.	left end point to right end point of nose	LN-RN
24.	nose height	NH



**Fig. 3.** Sample face image of DeitY-TU face database with 24 anthropometric distances

### 6.3 Measurement of Distances

For measurement of the distances, first task is to identify the landmark points, for which we have used morphological operator and Harris corner detection method [25]. In this approach, we have first generated a unique background using dilation (morphological) operation, so that the corner detection algorithm doesn't detect any corner in the background i.e. outside the face. Dilation is a powerful operator for extracting features from an image e.g. filling holes and broken areas. Here it is used to remove the light objects from the background. After that, we have used a corner detection technique called the 'Harris corner detector', which is an interest point detector and is strongly invariant to scale and illumination variation. Then, the intermediate distance between obtained corners are measured. These distances are then compared with the predetermined distances between the corresponding feature points. These predetermined distances are the manually calculated average distances between the feature points. The automatically measured distances those come within a considerable range of the corresponding predetermined distances ensures the detection of a desired feature point [26].

**Unique Background Creation.** In face images, it may happen in some cases that some background light is present. Presence of background light may create a problem in the corner detection process as the corner detection method will detect the corners from the whole input image i.e. from both the face region as well as the background of the face also. Therefore, to eradicate the possibility of getting any unwanted corner from the background, it becomes essential to create a unique background before starting the detection of corners. For this purpose, the 'dilation' morphological operation has been applied in this work.

Morphology is a broad set of image processing operations that apply a structuring element to an input image, creating an output image of the same size. In a morphological operation, the value of each pixel in the output image is based on a comparison of the corresponding pixel in the input image with its neighbours. The most basic morphological operations are dilation and erosion. In this chapter, we have applied dilation operation to generate the images with a unique background. Dilation is one type of operation that grows or thickens objects in an image [27]. In this operation, the value of the output pixel is the maximum value of all the pixels in the input pixel's neighbourhood. For e.g. in a binary image, if any of the pixels is set to the value 1, the output pixel is also set to 1. In our work, we have applied dilation operation on gray scale images. Mathematically, dilation is defined in terms of a set operation. The dilation of  $A$  by  $B$ , denoted  $A \oplus B$ , is defined as:

$$A \oplus B = \{z \mid \hat{(B)}_z \cap A \neq \emptyset\} \quad (1)$$

where,  $\emptyset$  is the empty set, and  $B$  is the structuring element. In words, the dilation of  $A$  by  $B$  is the set consisting of all the structuring element origin locations where the reflected and translated  $B$  overlaps at least some portion of  $A$ .

**Corner Detection.** A corner is defined as the intersection of two edges. Corners can also be defined as points for which there are two dominant and different edge directions in a local neighbourhood of the point. An interesting point can be a corner, but it can also be, for example, an isolated point of local intensity maximum or minimum, line endings, or a point on a curve where the curvature is locally maximal. The main advantages of a corner detector are its ability to detect the same corner in multiple similar images, under conditions of different lighting, translation, rotation and other transforms.

The Corner Detection block finds corners in an image using the Harris corner detection, the minimum eigenvalue, or local intensity comparison method. The block finds the corners in the image based on the pixels that have the largest corner metric values [28]. A simple approach to corner detection in images is using a correlation, but this gets computationally very expensive and suboptimal.

Harris Corner Detector is one of the promising tools to analyze the corner points. It is based on the autocorrelation of image intensity values or image gradient values. The gradient covariance matrix is given by:

$$M = \begin{pmatrix} A & C \\ C & B \end{pmatrix} \tag{2}$$

where,

$$A = (I_x)^2 \otimes w,$$

$$B = (I_y)^2 \otimes w, \text{ and}$$

$$C = (I_x I_y)^2 \otimes w$$

$I_x$  and  $I_y$  are the gradients of the input image,  $I$  in the X and Y direction, respectively. The symbol  $\otimes$  denotes a convolution operation. The coefficients have been used for separable smoothing filter parameter to define a vector of filter coefficients. The block multiplies this vector of coefficients by its transpose to create a matrix of filter coefficients,  $w$ .

The Harris corner detection method avoids the explicit computation of the eigenvalues of the sum of squared differences matrix by solving for the following corner metric matrix,  $R$ :

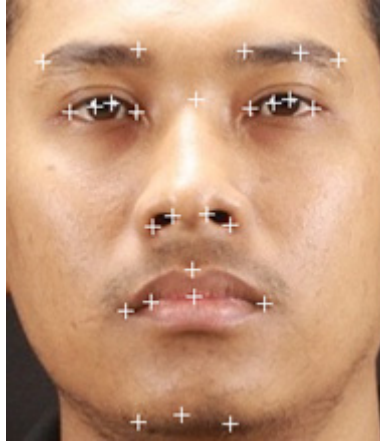
$$R = AB - C^2 - k(A + B)^2 \tag{3}$$

The variable  $k$  corresponds to the sensitivity factor. We can specify its value using the Sensitivity factor ( $0 < k < 0.25$ ) parameter. The value of  $k$  has to be determined empirically, and in this literature we have used the value 0.04. The smaller the value

of  $k$ , the more likely it is that the algorithm can detect sharp corners. On the basis of  $R$  the pixels are classified as follows:

$R > 0$ : Corner pixel,  $R \sim 0$ : pixel in flat region,  $R < 0$ : Edge pixel

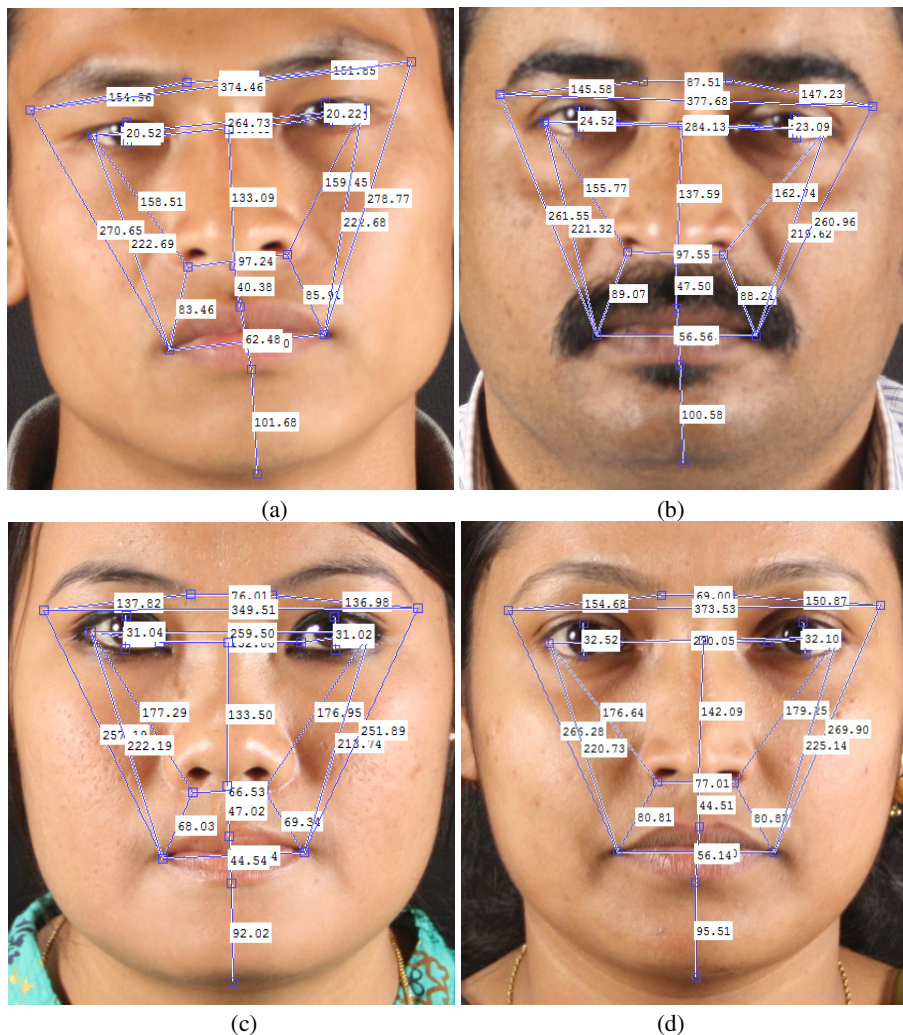
A sample image of the DeitY-TU face database has been shown in Fig. 4 with all the detected corners. Only the cropped face part is shown here.



**Fig. 4.** Sample DeitY-TU face of a male tribe with the detected corners

**Determination of the Desired Corners.** After the corners have been detected, our first task is to find out the different landmark points. For that, we first need to find out the lower and upper limits of the distances between the landmark points manually. For this purpose, manual distance calculation has been done over 20 images of different classes. For each distance, we find out the highest and lowest values and use these values as the limits or as a range for that particular distance, which leads to the determination of the desired corners or landmark points. However, this approach, for detecting the exact location of all the desired landmark points, has not been tested with a large number of dataset.

**Distances Obtained.** The obtained measurements for all the 24 distances of the 200 individuals are sub-divided into four groups: male tribes, male non-tribes, female tribes and female non-tribes, and the total number of images used for these groups of people are 66, 34, 62, and 38 respectively as shown in Table 3. Fig. 5 shows four sample images of these groups along with the measured values of the 24 distances. Here all the distances are Euclidian distances and have been measured in terms of pixels. Average values of these distances have been measured for the four groups separately for each of the five states (three states for the non-tribes) and are shown in tables 5 – 8 respectively. All the highest and lowest values for each distance are shown in bold and italic respectively.



**Fig. 5.** Sample images with 24 anthropometric distances measured for (a) male tribe, (b) male non-tribe, (c) female tribe and (d) female non-tribe

**Table 5.** Average distances calculated for the male tribes of the five different states

Sl. No.	Distances	Assam	Manipur	Mizoram	Nagaland	Tripura
1	REbR-REbL	166.5983	168.686	165.7736	<b>180.5791</b>	148.8273
2	LEbR-LEbL	162.7067	166.35	165.3036	<b>181.4762</b>	148.8613
3	LEbR-REbL	408.9767	411.046	399.9295	<b>440.418</b>	376.3247
4	LEbL-REbR	<b>83.32667</b>	80.901	71.63727	83.208	81.08333
5	RER-REL	74.58333	<b>91.107</b>	89.83636	82.398	68.39267
6	LER-LEL	74.34167	<b>90.372</b>	89.41182	82.454	68.11



**Table 5.** (continued)

7	LER-REL	304.4383	<b>323.209</b>	313.1591	320.993	282.1187
8	LEL-RER	155.9567	138.702	134.51	<b>156.966</b>	145.868
9	RN-RER	184.8867	196.785	189.7627	<b>204.093</b>	174.1513
10	LN-LEL	185.8267	193.993	188.6182	<b>206.268</b>	173.8733
11	RN-RM	90.84833	98.498	96.84318	<b>105.297</b>	97.356
12	LN-LM	90.31667	98.696	96.41545	<b>104.418</b>	97.68133
13	SN-ULOM	47.15833	<b>60.34</b>	50.89	54.9699	51.06
14	RER-RM	249.0083	259.556	251.8714	<b>271.9422</b>	239.9767
15	LEL-LM	247.5717	257.462	249.5945	<b>271.8606</b>	238.798
16	REbR-RM	300.6833	314.285	307.89	<b>326.993</b>	287.7393
17	LEbL-LM	292.545	310.796	306.9427	<b>325.0388</b>	285.7513
18	RM-LM	172.3917	175.682	172.3805	<b>182.427</b>	164.2893
19	ULOM- LLOM	68.155	65.702	66.53182	<b>69.9725</b>	68.59333
20	LLOM-CM	118.6967	117.824	114.9709	<b>125.3249</b>	103.5507
21	LEUM- LELM	30.12667	30.227	29.36773	<b>32.8001</b>	26.70067
22	REUM- RELM	29.705	29.903	29.01636	<b>32.7034</b>	26.11533
23	LN-RN	<b>104.7717</b>	100.89	104.5186	102.9178	98.1
24	NH	153.1683	147.439	147.4314	<b>162.988</b>	143.35

**Table 6.** Average distances calculated for the male non-tribes of the three different states

Sl. No.	Distances	Assam	Manipur	Tripura
1	REbR-REbL	168.1693	<b>172.3871</b>	155.2753
2	LEbR-LEbL	167.0171	<b>172.8293</b>	153.1347
3	LEbR-REbL	407.2993	<b>413.7254</b>	378.8153
4	LEbL-REbR	<b>75.60857</b>	70.34429	73.11933
5	RER-REL	83.69	<b>94.04429</b>	75.894
6	LER-LEL	82.91286	<b>94.07179</b>	75.558
7	LER-REL	300.81	<b>324.6482</b>	284.476
8	LEL-RER	<b>142.2286</b>	136.8936	135.5773
9	RN-RER	182.5879	<b>204.0354</b>	167.294
10	LN-LEL	181.1593	<b>204.2686</b>	166.0933
11	RN-RM	96.79143	<b>100.52</b>	90.624
12	LN-LM	96.48929	<b>100.5175</b>	90.60467
13	SN-ULOM	55.23714	<b>58.64143</b>	46.30467
14	RER-RM	247.5514	<b>269.1932</b>	224.73
15	LEL-LM	243.0786	<b>268.2468</b>	222.9647
16	REbR-RM	300.235	<b>331.9137</b>	269.274
17	LEbL-LM	298.3529	<b>329.6093</b>	267.1453

**Table 6.** (continued)

18	RM-LM	173.7964	<b>176.2614</b>	164.918
19	ULOM-LLOM	68.345	<b>68.37679</b>	60.444
20	LLOM-CM	<b>117.3807</b>	112.2532	111.502
21	LEUM-LELM	31.44357	<b>33.73714</b>	31.34867
22	REUM-RELM	31.04857	<b>33.13036</b>	30.75133
23	LN-RN	100.8214	<b>133.5311</b>	91.03067
24	NH	149.2029	<b>158.9457</b>	137.8753

**Table 7.** Average distances calculated for the female tribes of the five different states

Sl. No.	Distances	Assam	Manipur	Mizoram	Nagaland	Tripura
1	REbR-REbL	154.9843	156.106	152.5923	<b>167.5572</b>	136.5762
2	LEbR-LEbL	155.1571	156.986	151.8659	<b>167.2911</b>	138.8562
3	LEbR-REbL	389.0586	387.39	378.9532	<b>419.6009</b>	349.9323
4	LEbL-REbR	81.52571	73.386	76.50773	<b>84.0473</b>	76.51
5	RER-REL	75.17143	<b>87.448</b>	86.125	79.3188	67.44769
6	LER-LEL	74.45714	<b>87.102</b>	85.48636	79.1528	67.57385
7	LER-REL	292.0943	303.486	300.4777	<b>315.6656</b>	266.8977
8	LEL-RER	143.2	130.458	129.9445	<b>157.8811</b>	132.4977
9	RN-RER	177.5357	189.236	182.55	<b>198.2087</b>	166.5392
10	LN-LEL	180.6029	187.23	180.9836	<b>198.5761</b>	166.7777
11	RN-RM	93.30286	<b>99.68</b>	91.76364	97.6943	85.59385
12	LN-LM	93.42	<b>98.44</b>	91.065	97.9060	85.40077
13	SN-ULOM	48.05714	50.315	47.19455	<b>51.4861</b>	42.88077
14	RER-RM	237.8471	245.535	239.5614	<b>256.8499</b>	220.8631
15	LEL-LM	239.1886	242.2725	237.3	<b>255.1904</b>	218.9992
16	REbR-RM	290.4929	295.815	294.4673	<b>314.8386</b>	268.0831
17	LEbL-LM	295.4829	296.2525	292.7682	<b>313.4929</b>	268.4215
18	RM-LM	167.7729	171.565	160.7286	<b>174.4018</b>	147.4692
19	ULOM-LLOM	<b>75.90143</b>	62.1475	60.96409	66.0724	58.13769
20	LLOM-CM	104.2643	104.52	106.7527	<b>112.1954</b>	97.5
21	LEUM-LELM	31.43857	29.285	28.93773	<b>33.5458</b>	28.50462
22	REUM-RELM	31.03286	28.5975	28.64864	<b>33.5693</b>	28.15923
23	LN-RN	90.87143	88.7975	94.25409	<b>94.9848</b>	81.69923
24	NH	142.0129	137.435	139.6314	<b>153.2471</b>	133.3392

**Table 8.** Average distances calculated for the female non-tribes of the three different states

Sl. No.	Distances	Assam	Manipur	Tripura
1	REbR-REbL	152.25	<b>159.43</b>	147.0588
2	LEbR-LEbL	150.1369	<b>158.698</b>	148.1494
3	LEbR-REbL	386.5238	<b>389.2773</b>	366.5118
4	LEbL-REbR	<b>85.74462</b>	72.56133	73.48529
5	RER-REL	78.42462	<b>85.42733</b>	73.13941
6	LER-LEL	77.82538	<b>85.61333</b>	73.41588
7	LER-REL	297.0415	<b>303.664</b>	280.49
8	LEL-RER	<b>141.1908</b>	134.0133	134.3394
9	RN-RER	181.7023	<b>189.7347</b>	165.71
10	LN-LEL	181.8238	<b>189.7547</b>	169.6365
11	RN-RM	87.00692	<b>93.72467</b>	87.32
12	LN-LM	87.45769	<b>92.96133</b>	87.47118
13	SN-ULOM	45.60692	<b>48.68933</b>	43.96765
14	RER-RM	234.1054	<b>246.5713</b>	221.0918
15	LEL-LM	234.3115	<b>244.542</b>	221.0241
16	REbR-RM	286.4254	<b>300.7587</b>	266.3388
17	LEbL-LM	286.5523	<b>299.5113</b>	269.7194
18	RM-LM	<b>166.03</b>	165.974	155.0988
19	ULOM-LLOM	<b>65.27692</b>	64.034	64.28882
20	LLOM-CM	97.73231	<b>106.996</b>	101.9
21	LEUM-LELM	<b>34.20154</b>	32.15667	31.35647
22	REUM-RELM	<b>33.88231</b>	31.82533	30.91824
23	LN-RN	87.30538	<b>91.52067</b>	84.38765
24	NH	<b>150.38</b>	141.6693	134.3594

## 7 Analysis and Observations

According to the observations of the data values of anthropometric measurements of the tribe and non-tribe faces of NE India, shown in the tables from 5 to 8, we find that there are huge differences between the distances of the fiducial points of tribes and non-tribes that have been found. All these data shows how the tribe faces differ from the non-tribe faces within the same states as well as the differences between the tribes or non-tribes of different states.

### 7.1 Comparison between the Intrastate Male Tribes and Non-tribes

If we compare the different distances obtained for the male tribes and male non-tribes as shown in Tables 5 and 6, we can see that; length of the eyebrows (distance no. 1 and 2) of the male tribes are relatively shorter than the male non-tribes for the three states: Assam, Manipur, and Tripura, though, for Manipur, the difference is not that

major. Moreover, the distances between the outer endpoints of the eyebrows (distance no. 3) are almost similar for the male tribes and non-tribes of each of the three states separately, but the distance between the inner endpoints of the eyebrows (distance no. 4) is significantly larger for the male tribes in comparison to the non-tribes of the three states (Assam: male tribe – 83.32667, male non-tribe – 75.60857; Manipur: male tribe – 80.901, male non-tribe – 70.34429; Tripura: male tribe – 81.08333, male non-tribe – 73.11933). So, this indicates that the eyebrows of the male tribes are shorter in length and are a bit far away from each other compared to that of the male non-tribes.

Length of the eyes, (distance no. 5 and 6), are also shorter for the male tribes than the male non-tribes for the people of Assam and Tripura, but there is no significant difference for the Manipuri tribe and non-tribe males. Similar to the eyebrows, the distances between the outer endpoints of the eyes, (distance no. 7), also does not differ much for the male tribes and non-tribes of the three states, and the distance between the inner endpoints of the eyes, (distance no. 8), is also larger for the male tribes in case of Assam and Tripura, but for Manipur, it is almost same. Again, thickness of the eyes, denoted by distance no. 21 and 22 are almost similar for the tribe and non-tribe males of Assam and Manipur, but for Tripura, eyes of the male tribes are thinner than the male non-tribes (Tripura: male tribe – 26.70067, 26.11533, non-tribe – 31.34867, 30.75133).

The distance between the sub-nasal point to the upper lip outer middle (distance no. 13: SN-ULOM), i.e. the space amid the nose and mouth for the male tribes are larger, almost similar, and shorter for Assam, Manipur and Tripura respectively in comparison to the male non-tribes.

No significant difference is observed between the male tribes and non-tribes for the width of the mouth (distance no. 18: RM-LM), and thickness of the lips (distance no. 19: ULOM-LLOM). Still, the lips seem to be thicker for the male tribes compared to the male non-tribes of Tripura (male tribe – 68.59333, male non-tribe – 60.444).

The distance from the lower lip outer mid-point to the chin (distance no. 20: LLOM-CM) is observed to be same for the tribe and non-tribe males of Assam, but is slightly higher for the Manipuri male tribes, and radically lower for the Tripuri male tribes in comparison to the male non-tribes of particular states.

Width and height of the nose (distance no. 23 and 24 respectively) for Assamese tribe and non-tribe males are almost similar. Substantial difference is noticed for the Manipuri males, as the width of the nose of the tribes is less to a greater extent than the non-tribes (Manipur: male tribes – 100.89, male non-tribe – 133.5311), and height is also lesser for the male tribes. For the Tripura males, both the nose width and height are higher for the tribes compared to the non-tribes.

## **7.2 Comparison Between the Intrastate Female Tribes and Non-tribes**

From Tables 7 and 8, we have constructed a comparison between the structural differences of the female tribes and non-tribes.

Length of the eyebrows (distance no. 1 and 2) of the female tribes and non-tribes are similar for Assam and Manipur, but the Tripuri female tribes have smaller eyebrows compared to the non-tribes (Tripura: female tribes – 136.5762, 138.8562, female non-tribes – 147.0588, 148.1494). Again, the distances between the outer endpoints of the eyebrows (distance no. 3) of the female tribes and non-tribes of Assam and Manipur are similar, and for Tripura, it is much lesser for the female tribes than the non-tribes. The distances between the inner endpoints of the eyebrows (distance no. 4) are similar for the three states: Assam, Manipur and Tripura. So, it is seen that, unlike the male tribes, the eyebrows of the female tribes are not far away from each, but are not that much stretched out towards the outer ends as compared to the female non-tribes.

Slight variations are noticed for the length of the eyes (distance no. 5 and 6) between the female tribes and non-tribes of Tripura as for the female tribes have eyes with a shorter length than the female non-tribes (female tribes – 67.44769, 67.57385, female non-tribes – 73.13941, 73.41588). The distances between the outer endpoints of the eyes (distance no. 7) too are similar for Assam and Manipur, but are significantly shorter for the female tribes than the non-tribes of Tripura; and the distances between the inner endpoints of the eyes (distance no. 8) are almost similar for these three states. Again, thickness of the eyes (distance no. 21 and 22) are almost similar yet a bit lesser for the female tribes than the non-tribes of the three states.

The distance from the sub-nasal point to the upper lip outer mid-point (distance no. 13) is similar for both the female tribes and non-tribes of all the three states. The width of the mouth (distance no. 18: RM-LM) is similar for the female tribes and non-tribes of Assam, but is slightly higher for the Manipuri female tribes and lesser for the Tripuri female tribes compared to the female non-tribes of the corresponding states. Thickness of the lips (distance no. 19) are similar for the Manipuri female tribes and non-tribes, but lips are thicker for the Assamese female tribes and thinner for the Tripuri female tribes than the female non-tribes of particular states.

For Manipuri female tribes and non-tribes, the distance between the lower lip outer mid-point and chin (distance no. 20) is almost similar, but it is found that for Assam it is higher and for Tripura it is lower for the female tribes compared to the non-tribes.

The width of the nose (distance no. 23) is similar for all the female tribes and non-tribes of the three states, but the nose height (distance no. 24) for Assamese female tribes is significantly lesser than the non-tribes, though it is similar for the female tribes and non-tribes of other two states: Manipur and Tripura.

So, in case of interstate comparison, it can be noticed from these average distances that the Nagaland people (both the tribe males and females) have eyebrows with greater length and for the Tripura people, it is comparatively lesser than all the other states. Again, for eyes, it can be seen that it is higher for the Manipur people and lower for Tripura people. Similarly, the other structural differences in facial construction are also easily observable for the interstate tribe and non-tribe males and females, as all the highest and lowest values for each of these 24 distances are marked with bold and italic respectively in the tables from 5 to 8.

## 8 Conclusion

In this chapter, the anthropometric distances between the various facial feature points of the DeitY-TU face database have been measured in terms of pixels. Observation of the results obtained shows significant differences between the male and female Mongolians tribes and the non-tribes, which may be useful for forensic investigation in determining the origin of the terrorists and criminals of the north-eastern part of India. In the future, our aim is to conduct the experiments on larger dataset and improve the automatic process for detection of accurate facial feature points and thus, increasing the accuracy of the facial anthropometric measurements.

**Acknowledgement.** Authors would like to thank Prof. Phalguni Gupta of IIT Kanpur for his valuable suggestion regarding camera and light setup, and also to Prof. Barin Kumar De, Professor of Physics of Tripura University (A Central University), for his kind support to carry out this research work. This work presented here is being conducted in the Biometrics Laboratory of Computer Science & Engineering Department of Tripura University, under the research project entitled, “Creation of a Visual Face Database of North-Eastern People & Implementation of Techniques for Face Identification” (Vide No. 12(2)/2011-ESD, dated 29/03/2011) supported by the grant from the Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology (MCIT), New Delhi, Govt. of India.

## References

1. Kleinberg, K.F.: Facial anthropometry as an evidential tool in forensic image comparison. Thesis Report, University of Glasgow (2008)
2. Srivastava, D.: Terrorism & Armed Violence in India - An Analysis of Events in 2008. IPCS Special Report 71, Institute of Peace and Conflict Studies (2009)
3. Saha, K., Debnath, R., Bhowmik, M.K., Bhattacharjee, D., Nasipuri, M.: North-East Indian Face Database: Its Design and Aspects. In: 4th International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom 2012). LNEE, pp. 450–456. Springer (2012)
4. Saha, K., Saha, P., Bhowmik, M.K., Bhattacharjee, B., Nasipuri, M.: North-East Indian Face Database: Capturing, Performance Evaluation and Application. In: SPIE-IS&T Electronic Imaging 2013, vol. 8659, pp. 86590Q-1–86590Q-6. SPIE Digital Library, San Francisco (2013)
5. Ali, A.N.M.L., Das, I.: Tribal Situation in North East India. *Studies on Tribes and Tribals* 1(2), 141–148 (2003)
6. Bhagabati, A.C.: Tribal transformation in Assam and North East India: An appraisal of emerging ideological dimensions. Presidential Address, Section of Anthropology and Archaeology. In: 75th Indian Science Congress, Pune. Indian Science Congress Association, Calcutta (1988)
7. Hunter, A.G.: Craniofacial anthropometric analysis in several types of chondrodysplasia. *Am. J. Med. Genet.* 65, 5–12 (1996)

8. Allanson, J.E.: Objective techniques for craniofacial assessment: what are the choices? *Am. J. Med. Genet.* 70, 1–5 (1997)
9. Bishara, S.E., Cummins, D.M., Jorgensen, G.J., Jakobsen, J.R.: A computer assisted photogrammetric analysis of soft tissue changes after orthodontic treatment. Part I: Methodology and reliability. *Am. J. Orthod. Dentofacial Orthop* 107(6), 633–639 (1995)
10. Farkas, L.G.: *Anthropometry on the Head and Face*, 2nd edn. Raven Press, New York (1994)
11. Im, S.-W., Kim, H.-J., Lee, M.K., Yi, J.-H., Jargal, G., Sung, J., Cho, S., Kim, J.-I.: Genome-wide linkage analysis for ocular and nasal anthropometric traits in a Mongolian population. *Experimental and Molecular Medicine* 42(12), 799–804 (2010)
12. Luximon, Y., Ball, R., Justice, L.: The Chinese face: A 3D anthropometric analysis. In: Horvath, I., Mandorli, F., Rusak, Z. (eds.) *Proceedings of the TMCE, Ancona, Italy* (2010)
13. Hua-ming, L., Ming-quan, Z., Guo-hua, G.: Rapid pose estimation of Mongolian faces using projective geometry. In: *33rd Applied Imagery Pattern Recognition Workshop (AIPR 2004)*, pp. 171–176. IEEE Computer Society (2004)
14. Ngeow, W.C., Aljunid, S.T.: Craniofacial Anthropometric Norms of Malaysian Indians. *Indian J. Dent. Res.* 20(3), 313–319 (2009)
15. Hajnis, K., Farkas, L.G., Ngim, R.C.K., Lee, S.T., Venkatadri, G.: Racial and ethnic morphometric differences in the craniofacial complex. In: Farkas, L.G. (ed.) *Anthropometry of the Head and Face*, pp. 201–218. Raven Press, New York (1994)
16. Jahanshahi, M., Golalipour, M.J., Heidari, K.: The effect of ethnicity on facial anthropometry in Northern Iran. *Med. J.* 49(11), 940 (2008)
17. Sohail, A.S.M., Bhattacharya, P.: Detection of Facial Feature Points Using Anthropometric Face Model. In: *Signal Processing for Image Enhancement and Multimedia Processing, Multimedia Systems and Applications Series*, vol. 31, pp. 189–200 (2008)
18. Batista, J.P.: Locating Facial Features using an Anthropometric Face Model for Determining the Gaze of Faces in Image Sequences. In: Kamel, M.S., Campilho, A. (eds.) *ICIAR 2007. LNCS*, vol. 4633, pp. 839–853. Springer, Heidelberg (2007)
19. Moreno, L.A., Joyanes, M., Mesana, M.I., Gross, M.G., Gil, C.M., et al.: Harmonization of Anthropometric Measurements for a Multicenter Nutrition Survey in Spanish Adolescents. *J. Nutrition* 19(6), 481–486 (2003)
20. Peacock, C., Goode, A.: Automatic forensic face recognition from digital images. *Sci. Justice* 44(1), 29–34 (2004)
21. Das, B.M.: *Ethnic Affinities of the Rabhas*. Gauhati University, Guwahati (1960)
22. Das, B.M.: Anthropometry of the Tribal groups of Assam, India. In: Field, H. (ed.) *Field Research Projects, Coconut Grove* (1970)
23. Chaterji, S.K.: Kirata–Jana-Kriti. In: *The Asiatic Society*, pp. xxii+187 (1974)
24. Bhagabati, A.C.: Social formation in North East India. *Bulletin of the Department of Anthropology* VI, 9–29 (1992)
25. Harris, C., Stephens, M.J.: A combined corner and edge detector. In: *4th Alvey Vision Conference, Manchester, UK*, pp. 147–151 (1988)
26. Bhowmik, M.K., Majumder, G., Das, A., Saha, K., Bhattacharjee, D.: Human Eye Detection Using Harris Corner Detector. *J. Tripura Mathematical Society* 14, 16–23 (2012)
27. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: *Digital Image Processing Using Matlab*. Prentice-Hall, Inc., NJ (2003) ISBN 0130085197
28. Ramezanpour, M., Azimi, M.A., Rahmati, M.: A New Method for Eye Detection in Color Images. *Journal of Advances in Computer Research* 1(2), 55–61 (2010)

# Hand Biometrics in Digital Forensics

Asish Bera<sup>1</sup>, Debotosh Bhattacharjee<sup>2</sup>, and Mita Nasipuri<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering,  
Haldia Institute of Technology, Haldia-721657, India

<sup>2</sup>Department of Computer Science and Engineering,  
Jadavpur University, Kolkata-700032, India  
{asish.bera, mitanasipuri}@gmail.com,  
debotosh@ieee.org,

**Abstract.** Digital forensic is now an unavoidable part for securing the digital world from identity theft. Higher order of crimes, dealing with a massive database is really very challenging problem for any intelligent system. Biometric is a better solution to win over the problems encountered by digital forensics. Many biometric characteristics are playing their significant roles in forensics over the decades. The potential benefits and scope of hand based modes in forensics have been investigated with an illustration of hand geometry verification method. It can be applied when effective biometric evidences are properly unavailable; gloves are damaged, and dirt or any kind of liquid can minimize the accessibility and reliability of the fingerprint or palmprint. Due to the crisis of pure uniqueness of hand features for a very large database, it may be relevant for verification only. Some unimodal and multimodal hand based biometrics (e.g. hand geometry, palmprint and hand vein) with several feature extraction, database and verification methods have been discussed with 2D, 3D and infra-red images.

**Keywords:** Forensics, fusion, hand biometrics, multibiometrics.

## 1 Introduction

Omnipresence requirement of security concerned applications in different domains is extremely indispensable and ineluctable in this digital age to protect the assets and properties from unauthorized access or identity theft. As the world population is rising day after day in higher magnitude and therefore the private information is increasing proportionately. The size of the database is becoming ultra large and its truly essential terabytes to store data in digitized versions. Studies said about 95% data have been stored in digital format through worldwide and more than 50% of them are not printed out. So, it is very difficult to handle and operate with such a voluminous data, and that creates an easy path for the criminals to gain the benefits from the common people. Digital crime is growing massively through the computers, hard drives, USBs, disks, networks and other hand held digital devices such as mobile phones, PDAs etc. [4]. The technocrat criminals use very smart and modernized techniques for their bad intention. According to the FBI in 2012, more than 8 billion of the different property



loot was recovered [34]. Digital forensic is an alternative approach to overcome this subtle situation. Digital forensic is a discipline of forensic science that deals with digital data, methods for establishing the identity of a criminal through proper investigations. Several Computational Intelligence methods (fuzzy method, artificial intelligence, rough sets, genetic algorithms etc.) and pattern recognition techniques are serving in this field. Forensics bears very old historical background, and it dates back to the ancient roman era. Digital forensic was introduced in 1980s. The advancement of digital forensics is not commensurate and sophisticated enough with the development of criminology. In this dynamic field, the literary contribution is not at a satisfactory level and hence it desperately requires requisite attentions. It is still believed that, this is a developing area of research seeking robust solution and technology to save our planet from the crime and terrorism. Biometric technology is a key constituent of modern forensic and surveillance technology. Data from different digital resources (image, audio and video) of related biometric modes are accumulated for scientific investigations [23]. Fingerprint and DNA are two well-known and matured techniques and other distinguished modes with different features and tools are utilized as evidences. Hand biometrics is one of such relevant modality. Here, the future prospective of different hand based modalities (e.g. hand geometry, palmprint, hand vein etc.) in forensic have been addressed. Among all hand based features available, palmprint carries the most significance and hand-bacteria identification is a new promising area [30]. Hand geometric designs are thoroughly studied, and its implications have been sorted out. Dorsal hand vein pattern and handprint are also reported.

This chapter is organized as follows: Section 2 describes the biometrics and multi-biometrics. Section 3 is presented with digital forensic techniques and the scope of biometrics in digital forensics. Section 4 is contained with the details about various hand based biometric modes for forensics. Finally, conclusion is drawn in Section 5.

## 2 Biometrics

‘Biometric’ refers to the different physiological (e.g. face, fingerprint, iris, retina, DNA, hand geometry, etc.) and behavioral (e.g. voice, gait, signature, keystroke etc.) uniqueness of a person acquired from different human organs. It is an automated pattern recognition system that allows access grant only to the enrolled users. These inherent human properties can’t be easily stolen, spoofed or shared by third party, and users need not memorize them. Thus, biometric is considered as the best substitution of conventional knowledge based (password) and token based (ID card) secured system where possibility of identity thievery is high. The foremost job of a biometric system is to discriminate whether a claimed identity is legitimate or not. The major properties of a biometric system are uniqueness, universality, stability, measurability, acceptability and performance [1]. Some distinct properties from any particular mode of an individual are extracted and stored in the database as feature templates. Feature vectors are compared against the stored templates, and depending on a predefined threshold value matching is performed. The matching score determines whether the claimant is a valid or unauthorized user. A user can be tested for identification

(one-to-many) or verification (one-to-one) depending on the application necessity by applying different classification algorithms (e.g. Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (SVM) etc.). The performance of a biometric system is measured by the following parameters: False Accept Rate (FAR), False Reject Rate (FRR) and Equal Error Rate (ERR). The parameters are plotted in Receiver Operating Characteristic (ROC) curve. The FRR and FAR should be minimum in forensic and highly secured environment, respectively.

A biometric system works in mainly four modules: sensor, feature extraction, database and decision module [1, 18].

**Sensor Module:** Raw images from respective mode(s) are collected by the forensic experts using high quality cameras or scanners from the evidences at the place of crime occurred. Image resolutions for different modes are also been standardized (e.g. 500 ppi for palmprint). A set of different images are collected at different angle and pose. Noise could be associated with images and environmental factors may degrade the quality of images.

**Feature Module:** After noise removal and image quality enhancement, certain uniform and unique features are extracted from the underlying trait(s) for commencing the research. Features are stored in either encrypted or latent form, to protect their identity from intruders.

**Database Module:** Feature database is stored in high capacity storage devices such as hard drive or disk. The database size varies according to template size of applied mode and number of enrolled users. This module is very sensitive to attack by the hackers or criminals directly or indirectly from computers or through networks.

**Decision Module:** To find matching score with respect to a pre-specified threshold, test feature vector is compared with the feature vectors stored in the database. The absolute, Euclidean, Hamming or Mahalanobis distance functions are usually applied. Sometimes, the similarity score is expected to be normalized using min/max, median, z-score etc. based techniques [6]. Depending on the score, final decision is made to recognize a person as authentic or unauthorized person.

Biometric systems are developed using a single mode or several modes and categorized as unimodal or multimodal [1] system, respectively. A multimodal system uses certain types of fusion based techniques [6]. Some remarkable benefits over unimodal systems [18] include the following: (i) flexibility and universality (ii) improved matching accuracy (iii) “spoofing” attack minimization (iv) noise effects reduction (v) better reliability etc. Pre-matching fusion is applied at the sensor or feature extraction module and post-matching fusion is applied at the matching and decision module [6]. Fusion is applicable at every level: sensor level (2D and 3D imaging), feature level (similar features for the same trait like hand geometry, dissimilar features for different modes such as hand geometry and palmprint), rank level (some best results are arranged in an order), score level (matching scores are combined or normalized) and decision level (two or more decisions from different sources are combined). Decision

level and score level [11, 28] are two general fusion methods whereas sensor level and feature level [13, 27] are getting more observations. Related contributions on some fusion based methods are described in Section 4.

### 3 Digital Forensics and Biometrics

Digital forensic is derived from computer forensics. It can be defined as the specialized and scientifically proven methodologies, used for the reconstruction of events to identify criminal or unauthorized actions. Forensic Research Workshop (DFRWS) Technical Committee has defined digital forensic science [4] as: “The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.”

The necessary steps of digital investigations are collection, preservation, examination and analysis. Different models of forensics are available and in [7] some well known models (DFRWS (2001), EDIP (2004), CFFPTM (2006) etc.) are discussed. The major challenges of this domain include data encryption, anti-forensics, wireless technology, data size and many others [9].

**Collection:** It is a critical step, involves finding out and collecting the information and evidences (biometric evidences and e-evidences) relevant to the research. Finally, data is stored in digital format.

**Preservation:** Keeping the collected information safely so that no damage can be done during the process.

**Examination:** Scientific and systemic process of research, also known as “in depth systematic search of evidence” related to the event.

**Analysis:** Based on the previous step decisions are taken about the event.

The last two steps are time consuming and depending on the importance of the crime. The entire research process is solely dependent on the group of proper evidences. The Generic Computer Forensic Investigation Model (GCFIM) is described in [7], including two additional phases (pre-process and post-process) along with these general phases. Digital forensic is differentiated into mainly four different categories: computer, database, network and mobile forensics.

**Computer:** It justifies with the present state of a digital system such as computer, browsing history, the storage medium, last accessed and log files etc.

**Database:** Contains the information about database, metadata and log files.

**Network:** Analysis of a networked system (LAN/WAN or internet connectivity) are performed by observing the network traffic, delay and data transfer.

**Mobile:** It handles with the call details, SMS, MMS, browsing history and video clips of the device. Smartphones and 3G enriched with more facilities compared to ordinary cellphones can be used for retrieval of related information. Through the GPS, finding of locations is possible. Crimes through internet and mobile devices have increased unexpectedly, and it surpasses every year.

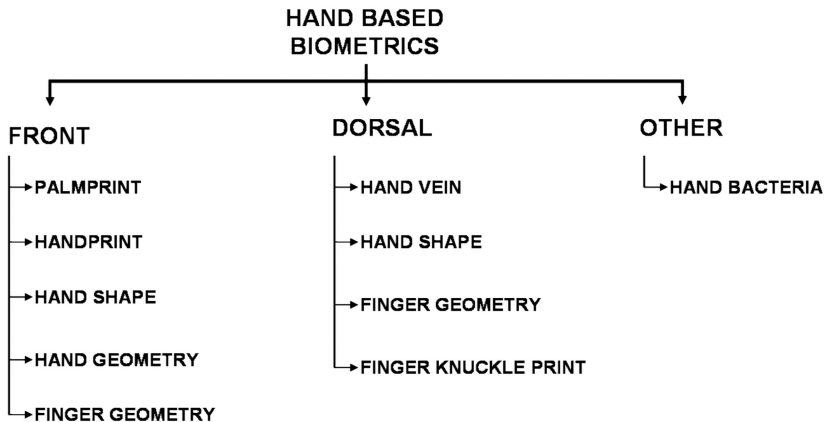
Biometric system works with digital images (2D or 3D), audio and videos. Each of these medium carries their own features and utility. Basically, the strengths of biometrics are employed in creating a strong forensic tool. Forensic is a post-event and biometric is a pre-event phenomenon. In user authentication, one or more particular mode(s) are predefined to access a secured environment. The system is user friendly, and users are cooperative. The reverse situation exists in forensics; the mode(s) is (are) not predefined. It is determined depending on the collectability of evidences associated to the crime. Event reconstruction is the principal challenge for the investigator. Biometric is a real time authentication system and checks liveliness of a user and processing time is low. But, in forensic liveliness is not an important factor, and it is time consuming to reach a final decision.

Biometric is universally implemented in different domains of government, commercial and forensics. For higher security environment retina, facial thermograms, iris, fingerprints are commonly used. Though, the gait is not much exploited, but it is important for real time surveillance systems, where a criminal or suspicious person can be traced by the CCTV footage. Several cases have been solved using this technology. Whereas the other competitor modes: latent fingerprint, palmprint, footprint, voice, DNA profile, and dental radiographs are obtained according to availability related to the event, are utilized for forensic investigations as evidence to identify the criminal. From 1980s, fingerprint bears its responsibility for the FBI [34]. In 1988, FBI introduced DNA analysis for research. After 10 years, in 1998 NDIS (National DNA Index System) was developed as part of CODIS at national level (USA) containing the DNA profiles [34]. Both are dominating other biometric modes since their origin. But, the main problems are the collectability and higher implementation cost because these biometric evidences cannot be collected from all the events.

In biometric applications, the size of the population is either low or medium. Practically, in forensic investigation user search space is millions which is one of the most important factors for any biometric system to provide maximum accuracy in such an extensive search space. From the evolution of biometric to present a situation, unfortunately, a particular method does not exist which can be described as the 'best' in all kind of applications. Not all of these modalities are examined properly. Other than fingerprint and DNA analysis, face and palmprint are already exploited in forensics. Hand geometry, footprint, voice and gait recognition are these exploratory area that need to be more focused. The Biometric Centre of Excellence (BCOE) also wants to improve the potential in the near future, and their next generation approach will be "bigger, faster, and better" than the Integrated Automated Fingerprint Identification System (IAFIS) [34].

## 4 Hand Based Biometrics

The hand is an important interface of human to perform most of the works in our daily life. It is one of the oldest biometric features used for authentication [8]. It is not yet explored in criminal investigation and law enforcement. It may be useful in certain circumstances where other modalities (e.g. DNA, face etc.) are unavailable or unreliable. Other significant benefits it provides are reliability, convenience, non-intrusive, user friendliness and immutable as it is dependent on the intrinsic physical properties directly or behavior of a person. A number of distinct features are measured from different parts of either side of the hand and used as biometric properties. From the front side of the hand, features of palmprint [3, 12], handprint (especially for infants) [29], hand body [13], finger geometry (specific fingers) and hand geometry are measured. From the back side of the hand, vein pattern [2, 17], dorsal hand form [20, 25] and finger knuckle print [10, 16] are commonly considered as biometric characteristics and named according to part of the hand associated with this purpose. All the related hand based modes are shown in Fig.1.



**Fig. 1.** Classification of hand based biometrics

Palmprint and hand geometry are two general and conventional hand based recognition systems. Performance of hand-shape (frontal or dorsal) detection is satisfactory comparatively when fusion based techniques are implemented. Dorsal vein pattern and finger knuckle prints are emerging the field of hand biometric system. But, using these modes no business exists for digital research purposes. Palmprint produces very high accuracy as compared to fingerprint. It is considered as a better substitute of the other because more discriminative features can be extracted as larger surface area of the palm. Hand vein recognition is rendering very significant performance. Another pertinent field is hand bacteria identification. Though it is not regarded as biometric directly, but it resides in hand surface, producing excellent correctness.

### 4.1 Hand Geometry

Hand geometry mainly includes the measurement of many consistent geometric features of fingers (such as length and width at different positions of fingers) and hand

form (such as palm area and palm width). '*Identimat*' was the first successful hand biometric system used for person authentication since 1980s. Afterwards, other devices were developed using this modality in commercial and government applications such as attendance maintenance, Olympics (introduced in 1984), nuclear plants and many others [8]. The number of different geometric features generally lies within the range of 20 to 40. Feature template size is significantly low, requiring only within 10 bytes and having lesser processing time (about 5 sec.). Other important benefits are easier to access, and environmental parameters such as bad weather, dry skin and lighting conditions can't alter system performance heavily. But, lack of high uniqueness of hand features as compared to fingerprint features or DNA profile is a major concern for adaptation in forensics. It is inapplicable for identification because of the larger population search space and thus it suffers from the scalability problem. In the verification, it's a suitable alternative. Most of the unimodal system allows either the left or right hand. Whereas some fusion based systems are developed using both hands [8].

In early days of hand biometrics, a CCD camera or scanner was used for image acquisition, and the image quality was very low (less than 100 ppi). The pose of hand placement was fixed by using 'peg'. A rigid pose minimizes the inter-class and intra-class pose variations, finger alignments and maintains uniform spacing between fingers. This imaging system reduces misclassification rate. The major problem is larger device dimension and impossible to embed in smaller devices (laptops) and may causes hygienic issues for personal. Recently, high resolution digital camera and infrared (IR) camera [19, 26] are applied that facilitates without any pose restrictions, providing more user flexibility. These imaging systems are lesser error prone and lesser noise sensitive, but cost of the device is greater. Modern research interest on hand geometry is paying attention to 3D images and its fusion with 2D images as well [15]. Data fusion is performed with the other hand related modes such as palmprint, finger knuckle print or hand vein model for robust and reliable solution [21, 27]. Multi-biometrics employing fingerprint and/or palmprint along with hand geometry is cost effective due to single working sensor [32]. Other than these modes, human face is considered as other important fusion mode [14]. The fusion techniques are applied at various levels. Brief studies of some general works are given in Table 1.

Some major complexity arises in hand geometry are:

- i) The most challenging issue is to create a standard orientation of all images at preprocessing stage, before feature calculations. Freeness of hand placement causes angle variations between fingers and major axis. Incorrect finger alignment can't locate finger tip and valley points accurately.
- ii) Any hand gadget; ornament or bracelet can affect wrong hand contour or form and determination and feature calculation. Stylish fingernail especially for women can play an important role for the same.
- iii) Any damage or injuries in hand can prevent the usage of this mode.
- iv) Hand shape changes over time and age. Silhouette at childhood is changed at different ages of a lifespan.

It is appropriate for low or medium security based applications with moderate population size. Fusion based hand biometrics is more interesting than single mode. All the related systems support the real time authentication.

**Table 1.** Some state-of-the-arts methods of Hand Geometry

Author	Mode	Classification technique	Image quality	Database size	Accuracy
[5]	2D	i) Hausdorff distance of hand contours, ii) independent component features (ICA1 and ICA2) of hand silhouette images.	HP Scanjet 5300c 383×526 pixels at 45 dpi.	1374 right hand images from 458 subjects.	Verification: i) Hausdorff: 97.36%. ii) ICA1: 97.2%. ICA2: 98.2%
[8]	2D	Independent Component Analysis (ICA) on global hand appearance of both hands.	Flatbed scanners, at 150 dpi.	918 subjects, 3 images per left & right hand per user.	Verification: Left: 1% EER. Right: 1.16% EER.
[11]	Eigen palm and Eigen finger.	Feature extraction by Karhunen-Loeve (K-L) transform. Final decision by the modified k-NN.	Low-cost scanner at 180dpi.	1,820 hand images of 237 people.	Identification: 0.58% EER.
[13]	Hand shape and palm texture.	Correlation-based feature selection (CFS) algorithm. KNN Classifier by minimum Euclidean distance.	Digital camera 300×300 pixel.	1000 images of 100 subjects, 10 images per subject.	Recognition: 97.8%
[15]	2D & 3D hand geometry and palmprint.	3D palmprint, represented by SurfaceCode. Matching score level fusion.	3D digitizer. 640×480 pixels.	3540 right hand images from 177 subjects.	ERR: 2.3%. AUC (area under the ROC curve) : 0.9888
[19]	2D thermal images.	Different geometric features of a hand are extracted. Recognized by Extension theory.	Infrared camera.	300 images, 30 persons, 10 images per user.	Accuracy: 92%
[20]	2D Dorsal hand geometry + finger-print.	Min-max scores normalization. Distance based matching with respect to the threshold. Feature level (same mode) and score level (multimode).	NIR camera. 240×320 pixels, at 72 dpi. Veridicom sensor for finger-print, at 500 dpi 300×300 pixels.	100 users, 5 images for each mode of left and right hand. Total 30 images per users.	EER: 0.0034%

#### 4.1.1 System Model

A simple plan, which is discussed in Fig.2, is to find out the suitability of hand biometrics in the forensic domain. In traditional hand biometric system, hand type (left/right) and number of sample images (for enrolment and testing) are predefined which is one of the most unavoidable limitations. Although, there may some accident or fracture on the hand that has been used for enrolment phase cannot be altered at some later time. But, in case of forensic the type of hand can not be pre-specified. It is defined according to the collected evidences. So, in such a situation, this present scheme is suggested which can determine the hand type automatically, and features can be extracted accordingly. A robust technique is needed for calculating accurate features. So, the most significant step is hand image normalization. It consists of several sub-steps which are environment elimination, rotation, irregularities removal at wrist region, determination of hand type and finger tips and valleys localization. Many features are calculated from the normalized hand images. Finally, users are classified as genuine or imposter according to the classification algorithm.

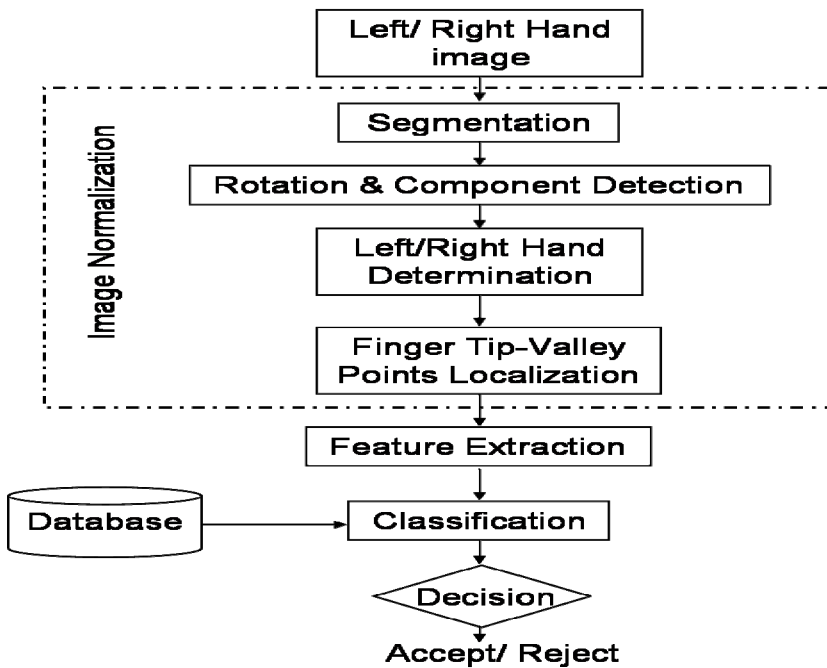


Fig. 2. Hand geometry based forensic system model

#### Hand Image Segmentation

Collected raw images consist of the actual hand texture with or without any stuff as foreground and a dark background. The first step is removal of undesirable background and noise associated with images. Conversion from an original color image into a grayscale image (Fig.3a) is performed using thresholding by the *Otsu's* method,



and median filter is applied for noise removal. Images are then converted into binary (Fig.3b) from the grayscale images and hand contour is detected using 'Sobel's edge detection method, as shown in Fig.3c.

### *Rotation and Component Detection*

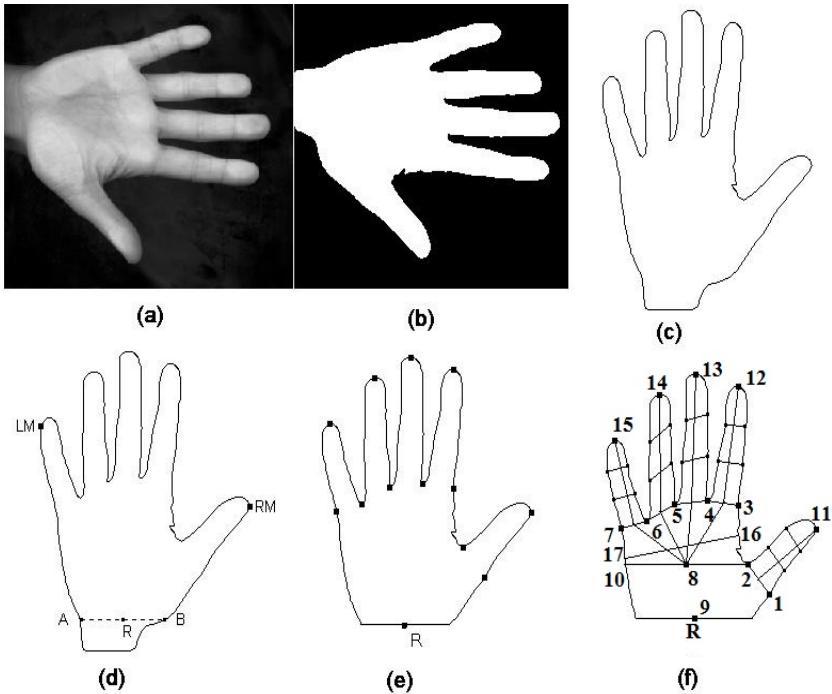
The hand images are acquired at different posture, different angle with their hand gadgets; only with a little attention that no two fingers should be attached or overlapped. So, to define a standard uniform template, a particular orientation for all the images is employed by rotating at various angles (90,180 degrees) as required to make them perpendicular with the major X-axis. Due to any hand ornaments or wristwatch used by a person during image acquisition, the binary image may contain several components. The largest part is detected, and remaining small components are ignored. A reference line (AB) at certain distance, above from the bottom of that part is considered by scanning the pixels from the left to the right direction. The leftmost and rightmost nonzero pixels are the two end points of the reference line. The lower portion of line AB is neglected, and its midpoint (R) is defined as the reference point (Fig.3d).

### *Left or Right Hand Determination*

To decide whether the given hand is the left hand or the right hand the given input is needed to locate the tip region of the thumb. Generally, for left or right hand thumb tip region, the leftmost or the rightmost nonzero pixel above reference line is traced, respectively. Then, the reverse extreme pixel in the opposite direction for either hand is considered. To check whether the image is of left or right hand, examine these two extreme pixels. Say, the leftmost pixel is LM, and the rightmost pixel is RM. If the leftmost pixel LM is below the rightmost pixel RM with respect to the R, then the finger is thumb( related to LM), and other is little finger( related to RM ) and that hand is considered as a left hand. Similarly, reverse reason is followed for the opposite hand. This identification of hand type is helpful for finding out the positions of finger tips and valleys because the space and flexibility between thumb and index finger is not same as between ring and little finger.

### *Locating Finger Tips and Valleys*

A general algorithm has been applied to both the hands to locate tip and valley points and based on which features are extracted. Once the thumb tip region is located, little finger tip region can also be found out in a similar manner. From those tip regions, exact tip features can be traced on hand contour by scanning the pixels. Other finger-tip positions are located using the maximum Euclidean distance from the reference point. First of all, middle finger tip is identified, and then tips of index and ring finger are placed. The valley points between any two fingers of either hand are determined by scanning the hand contour. Other valley points of the thumb, index finger and little finger are defined using equal Euclidean distance from the tip to the other valley point, placed at the opposite side of that particular finger. All the key points are marked (by point 1, 2, 3 etc.) in Fig. 3f.



**Fig. 3.** (a) Grayscale hand image, (b) binary image, (c) rotation and hand contour, (d) determination of the right hand and guillotining, (e) locating landmarks points, (f) feature extraction

### *Feature Extraction*

After locating the landmark points, some unique features from any normalized hand image are extracted by measuring lengths and widths at one-third and two-third position of the finger length of individual fingers. The widths of palm at two different locations and distance from the midpoint of palm width line to the middle of every finger baseline, except the thumb are computed. Total 26 hand features are measured from each normalized image. Most of the features are widely accepted in many previous works, and some new features are incorporated.

The feature set is defined as follows: 5 finger length (1 per finger); 10 finger width (at 1/3 and 2/3 position of every finger); 5 finger baseline width (1 per finger); 2 palm width (at different position of palm ;depicted by line 2-10 and 16-17) and 4 length from the middle point of palm line (illustrated in Fig. 3.f as midpoint 8 of line formed between point 2 and 10) to mid of finger base-lines excluding the thumb. All features are graphically shown in Fig.3f, and the number of features can be increased.

### *Classification*

Persons are recognized by the minimum distance classification algorithm. At first, the Euclidean distance between the test feature vectors and the stored templates are

calculated. Based on the minimum distant, summation of the features from a particular subject is considered as a recognized class. Consider,  $A_{m \times n}$  is an enrolled feature matrix; where, row  $m$  represents extracted features for all subjects and column  $n$  is the feature vector for every enrolled image.  $B_{k \times n}$  is the test feature matrix; where,  $k$  is used for testing for every subject and  $n$  is same as defined above. Matrix  $C_{m \times n}$  is used to store the Euclidean distances between every test vector with all rows of 'A' matrix. Calculate the Sum (S) for every row of C matrix. If multiple samples are used during the enrolment phase then, average of Sum is computed to assign final class label. Formally, the algorithm is given below.

**Input:** Enrolled feature matrix(A), test feature matrix(B).

**Output:** Recognized Class label.

$$1. C_{i,j} = \sqrt{(A_{i,j} - B_{x,j})^2}$$

$$2. S_i = \sum_{j=1}^n C_{i,j}$$

$$3. Class = \min(Avg(S_i))$$

4. End.

Where,  $1 \leq i \leq m$ ,  $1 \leq x \leq k$  and  $1 \leq j \leq n$ ; A, B, C and S are defined above.

### Experimental Results

The database is collected from E. Yörük et al. [5]. In this work, the database contains images of 253 users, 3 images for each hand of a person. 157 left hand user and 96 right hand users are considered. First image with 383×526 pixels and finally at feature extraction step the images are resized to 200×300 pixels. The population is a blend of left and right hand subjects; the result is calculated with the distance threshold value of 3.2. For enrolment (size = K), one and then two images per subject are applied, and only one image is used for testing and the results are given in Table 2.

**Table 2.** Performance evaluation

Enrolment Size (K)=1		
Hand Type	Minimum Threshold(t)	Recognition Rate(%)
Left	5.6	94.9
Right	5.8	96.88
Combined	5.8	95.65
Enrolment Size(K)=2		
Hand Type	Minimum Threshold(t)	Recognition Rate(%)
Left	2.8	98.09
Right	2.4	100
Combined	2.8	98.81

With different population size, the recognition rate is also computed for  $K=1$  and 2, shown in Table 3. The performance is not up to a satisfactory level for  $K=1$ , which is 3.16% lower than  $K=2$ . The experimental result shown in Table 3 is carried out by varying the group size from 50 to 253, at five steps.

**Table 3.** Performance with different population size

Population Size	50	100	150	200	253
Recognition rate ( $K=1$ )	100	99	97.3	96.5	95.65
Recognition rate ( $K=2$ )	100	100	99.3	99.5	98.81

The experimental results mean that, the system provides acceptable performance for the medium security applications with improved 98.8% accuracy for 253 combined subjects.

#### 4.1.2 Applications and Scope in Forensics

Hand geometry recognition systems are not widely used for authentication purpose. Approximately, 10% of systems are used for user verification in academic institutions and industrial purposes. Examples: Sacramento County, California, was using hand geometry. The INSPASS (The United States Immigration and Naturalization Service (INS) Passenger Accelerated Service) System used hand geometry for verification. Wells Fargo Bank uses hand geometry to prevent from unauthorized access to the bank's data centres. An elementary school in New Mexico has also presented hand geometry system. The University of Georgia has used a hand geometry system since 1972 to identify students on the unlimited meal plan, thus preventing them from lending their cards to others. In a lot of other domains, this trait is being used since the last century.

Images and videos of children engaged in sexual activities are increasing the number of criminal cases involving computers. An investigation by the RCFL (Regional Computer Forensic Laboratories) implied maximum percentage of child pornography. Hand biometrics is used in forensic cases such as child abuse, sex exploitation, kidnappings and missing infant identification [23]. Studies imply that in India more than 7,200 children, including infants, are raped every year, and still many cases go unreported [33]. About 53.22% children faced different forms of sexual abuse and in 83% of the cases parents were involved positively. Structure of hand carries the useful information of determining gender and age. So, if the hand images from such activities are collectable, then decision can be taken whether a child is involved or not in those crimes. The Forensic Audio, Video and Image Analysis Unit (FAVIAU) is also interested in this section [23].

It can be prevalent in certain situations from where any strong evidence is not accessible. Most of the criminals use facial masks and hand gloves and most of the time the event is pre planned. The situation can illustrate as criminal used mask and gloves very often during the crime from where it is very complicated to collect evidences and identify the person by face, fingerprint or DNA. Even if the gloves are damaged too.

Form an unidentified body remains due to some accident or blast, the initial research can be started with. Same reason is also legitimate for deformed body identification. Another situation may take place due to environmental factors. Hand with dirt, oil, water, blood or any kind of liquid, from where DNA can't be retrieved, and reliability of fingerprint is minimized. So, hand geometry can be the best option to be utilized in such circumstances.

## 4.2 Palmprint

Palmprint is a well known biometric technology, introduced in India by Sir William Herschel in 1858. In 1994 palmprint supported system were developed by a Hungarian company. It carries similarities with fingerprint. It consists of the friction ridge information such as ridge flow, ridge structure, major palm lines etc. Palmprint is believed as a stronger mode with certain properties like, uniqueness, universality, stability and collectability for authentication. Palmprint offers many advantages over fingerprint and hand geometry [12]. Palmprint area is larger than the fingerprint area and thus carries more robust information than a fingerprint about the person. Palmprint of twins are different, and the palm lines also ensure genetic disorder of people. High resolution (at least 500 dpi) palmprint is suitable for forensic [3] and low resolution (at most 400 dpi) palmprint is useful for authentication. The features include the principal lines, ridges, wrinkles, minutiae and delta points of the palm. Some common subspace based methods used for this mode are Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Statistical and transformed based techniques are also developed. The system should support partial-to-full matching scheme [12]. The accuracy of palmprint is comparable with the fingerprint and DNA. Surveys from law enforcement resources imply about 30% evidences collected from a crime scene contains palmprint [3]. But, main difficulty is collectability. Images of palmprint from different objects are collected with noise or overlapping which create complications in feature extraction. The most difficult job is correctness improvement of partial-to-full matching. Rotation of palmprint at any random angles is another key factor. Minutiae-based latent-to-full matching by the *MinutiaCode* is described in [12], and the performance is evaluated with live-scan partial palmprint and latent palmprint against background samples of full palmprint. Matching is performed both locally and globally using improved feature selection methods. Rank level fusion of latent palmprint is experimented using OR rule to test whether two latent palmprint represents the same or a different person. The radial triangulation based approach is proposed by [22] with full palmprint and latent palmprint. The details of experiments are given in Table 4.

Handprint is another method that is closely similar to the palmprint. It includes all the printable area of the hand, including the fingers whereas palmprint only includes the palm area. Handprint is mainly suitable for child identification as because their palm features are not different, reliable and invasive. On average, the hand area of infants is 2.5 to 3 times smaller than the adults. It means that the feature extraction is difficult because baby hand features are very fragile and change over time. The image quality should be very high for better ridge extraction using this mode. According to

[29], it should be 1500 ppi and in their contribution images are collected at 1000 ppi. From the handprints, gender and height can be guessed and from the skeletal bone structure of the hand, age of can be estimated also. But, baby handprint can't be collected sometimes due to hygienic problems as their skins are very sensitive and oily. It has also been observed baby hands are closed most of the time. So, in such conditions footprints [24] are dependable.

Below 16-24 ages people, are requiring better protection both in the real and cyber world. Newborn babies are being stolen from the hospital in most of the cities. Such a case happened in Mumbai, which led the Mumbai High Court to advise that all newborn babies should have their feet and handprint captured within two hours of birth. In developing nations like ours or Brazil, rate of child missing or swapping from hospitals is also high [29]. Face recognition technology could not be applied reliably below the age of twelve, and similarly voice recognition can't be used due to the undifferentiated nature of voice between boys and girls. So, newborn baby identification is not explored properly, and it needs a robust, low cost and faster solution in real time applications.

### 4.3 Hand Vein

Hand vein pattern recognition is one of the new research area that find it's suitability in criminal identification since the killing of the US journal reporter D. Pearl by the mastermind of 9/11 attack, Khalid Mohammed. This field is not matured enough through worldwide in terms of application domains. In 1997, Hitachi developed first vein pattern matching device and used in Japan and Korea for verification vastly. In Japan, approximately 80% banks use contact free finger vein biometrics. It is highly unique even for twins. It is more reliable and invasive than the other hand biometric systems. Study implies its performance in a constrained environment is very high (99.99%) comparable with DNA matching. Vein pattern includes the inner blood vessels, visible from the outer skin surface. It can't be manipulated or replicated externally, and it's very difficult for 'spoofing' attack, as well. No weather condition can hamper its performance. It is stable over time for adults. But for a child and older people the pattern changes over time. The vascular pattern is complex and different enough for pattern recognition. An infrared (IR) camera (mainly, Near IR) is employed for image acquisition (wavelength 800-1000 nm) from the back side of the palm. During preprocessing, noise is removed, and image quality is improved before feature definition. As it contains complex vascular pattern, some difficulty arises for background removal. Selecting the Region of Interest (ROI) is most significant challenge for feature extraction.

Finally, the feature vectors are matched against the enrolled database. Some general matching algorithms are Minutiae-based; Hausdorff or Euclidean distance based, correlation based etc. vein pattern matching is relevant in higher security based and forensics applications. Liveliness of a person can be verified in a contact free nature, makes it attractive. But, still no strong database present of this mode and no forensic use have yet been developed.

**Table 4.** Study of some related hand based modes for forensics

Author	Mode	Classification	Image quality	Database size	Accuracy
[22]	Palmprint	Latent-to-full matching using radial triangulation method.	Full palmprint: 2304 x 2304. Latent print: 500 ppi.	22 latent print, 8680 full palmprint form 4340 subjects of left and right hand.	Identification: (Rank-1):62%.
[12]		Minutiae-based latent-to-full palmprint.	Full palmprint: 1000 ppi. Latent print: 500 ppi.	i) 150 live-scan partial print ii) 100 latent palmprint, with 10200 full palmprint from Noblis and Michigan State Police (MSP) and Michigan State University(MSU).	Identification: (Rank1) i)78.9% , ii) 69%.
[3]		Fourier Transform of images and combines Modified Phase-Only Correlation with Fourier-Mellin Transform.	THUPALMLA -B and PV-TESTPARTIAL: both at 500ppi.	i) THUPALMLAB: 152 palms (1216 full and 1216 partial palmprint). ii)PV-TEST-PARTIAL: 10 users, 80 full and 40 partial palmprint.	EER: i) 27.1%, ii) 23.8%. With 101x101 inside lobe.
[29]	Handprint	Combined two stage approaches of Simulated Annealing and Oriented Texture Field (Finger Code-FC).	CrossMatch LSCAN 1000P sensor at 1000 ppi. (4964 x5120 pixels).	Original Database (NB_ID): 1221 palmprint from 250 newborn at the University Hospital (Universidade Federal do Parana). 60 images form 20 newborn are experimented.	SA: At 0% FAR, GAR is 78%. FC: Rank 5.
[2]	Hand Vein	Multimodal Fat Distance (FD) + Max-Min Distance (MMD) SVM.	NIR camera set- up able to acquire an image of 320x240 pixels.	342 sample fingers of 114 users, 3 samples per finger.	Identification: GAR: 94%, FAR: 0.15.
[17]		Euclidean Distance based matching method.	Digital SLR camera with infrared and night vision lamp (940nm).	Database: (IITK) 1750 sample images of 341users. Absorption based method for image acquisition.	Verification: 99.26% at 0.03%.FRR.

#### 4.4 Hand Bacteria

A new direction in forensic investigation has been found by using the hand bacteria that live on the human hand surface which is 70-90% accurate, and its performance will be enhanced over time [30, 31]. The precision of this technology will be same as DNA or fingerprint. When blood, hair, fingerprint, saliva, palmprint or any strong evidence is not available from a complex and unfavorable conditions to determine the identity of the criminal, then this method will be most appropriate in the near future. Bacterial diversity of woman is higher than man [31]. Hand bacteria can be collected from the touched surfaces of the keyboards, mice or any object used for daily purposes and applied for the investigations. The bacterial DNA structure of the owner is more perfect and stable than any person even for twins. Bacterial communities on a finger, palm, fingertips or keyboard are distinct and very closely related to a person than others. About 150 bacteria species can be found, and they persist as long as 2 weeks at room temperature out of which only 13% of the species are shared between any two persons. The stability is very high, and the bacterial communities can be recovered even after washing out hands after hours. So, it would be simpler to accumulate evidences from the touched objects or surfaces using the bacterial DNA rather than DNA of that person.

Though the hand bacteria are not directly related to the traditional hand biometrics, but it could be a fantastic alternative tool for the next generation forensic identification. Researches in this excellent field are going on. A specialized and benchmark forensic tool can be developed in the coming years.

### 5 Conclusion

Hand based biometrics is serving many commercial and government domains throughout many decades. Apart from conventional user authentication, it can play a significant role in forensics. Different hand based modes carry some potential benefits which have been enlightened in this domain. Palmprint and dorsal vein structure are comparable to the DNA and fingerprint analysis in terms of accuracy, time and cost. Palmprint and handprint have already established their positions in forensics. According to our exploration, no work persists on hand geometry, finger knuckle print in forensics. Hand vein pattern is trying to find out its relevance in a criminal investigation. Hand bacteria identification belongs to the same category with enormous future possibility. Applications of all related traits are illustrated, and their scopes are depicted. The prime objective of related studies on these traits is to focus on their forensic aspects. Several multimodal fusion based approach are developed to perform well. A straightforward hand geometry recognition method is suggested for the same purpose. In next generation forensic, these hand modes can be utilized by the standard organizations, like FBI. So, we could expect from the forensic professional a strong forensic framework and tool using hand biometrics will be contributed. Finally, some research attention needed in this field in the hope that the applicability of hand based biometrics in forensic investigation will be improved in the near future.



## References

1. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer (2006)
2. Vehils, D., Miguel, J.: Design and Implementation of Finger Vein Identification System. Ph.D Thesis (2011)
3. Singh, S.: Partial Palmprint Matching for Forensic Applications. M.Tech Thesis (2012)
4. Reith, M., Carr, C., Gunsch, G.: An Examination of Digital Forensic Models. *International Journal of Digital Evidence* 1(3), 1–12 (2002)
5. Yörük, E., Konukoğlu, E., Sankur, B., Darbon, J.: Shape-Based Hand Recognition. *IEEE Trans. on Image Processing* 15(7), 1803–1815 (2006)
6. Jain, A.K., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. *Pattern Recognition* 38, 2270–2285 (2005)
7. Yusoff, Y., Ismail, R., Hassan, Z.: Common Phases of Computer Forensics Investigation Models. *International Journal of Computer Science & Information Technology (IJCSIT)* 3(3), 17–31 (2011)
8. Dutağacı, H., Sankur, B., Yörük, E.: Comparative analysis of global hand appearance-based person recognition. *Journal of Electronic Imaging* 17(1), 011018 (2008)
9. Agarwal, A., Gupta, M., Gupta, S., Chandra Gupta, S.: Systematic Digital Forensic Investigation Model. *International Journal of Computer Science and Security (IJCSS)* 5(1), 118–131 (2011)
10. Kumar, A., Ravikanth, C.: Personal Authentication Using Finger Knuckle Surface. *IEEE Trans. on Information Forensics and Security* 4(1), 98–110 (2009)
11. Ribaric, S., Fratric, I.: A Biometric Identification System Based on Eigenpalm and Eigenfinger Features. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 27(11), 1698–1709 (2005)
12. Jain, A.K., Feng, J.: Latent Palmprint Matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(6), 1032–1047 (2009)
13. Kumar, A., Zhang, D.: Personal Recognition Using Hand Shape and Texture. *IEEE Trans. on Image Processing* 15(8), 2454–2461 (2006)
14. Tsalakanidou, F., Malassiotis, S., Strintzis, M.G.: A 3D face and hand biometric system for robust user-friendly authentication. *Pattern Recognition Letters* 28, 2238–2249 (2007)
15. Kanhangad, V., Kumar, A., Zhang, D.: A Unified Framework for Contactless Hand Verification. *IEEE Trans. on Info. Forensics and Security* 6(3), 1014–1027 (2011)
16. Zhang, L., Li, H.: Encoding local image patterns using Riesz transforms: With applications to palmprint and finger-knuckle-print recognition. *Image and Vision Computing* 30, 1043–1051 (2012)
17. Soni, M., Gupta, S., Rao, M.S., Gupta, P.: A New Vein Pattern-based Verification System. *Intl. Journal of Computer Science and Information Security* 8(1), 58–63 (2010)
18. Ross, A., Jain, A.K.: Information fusion in biometrics. *Pattern Recognition Letters* 24, 2115–2125 (2003)
19. Wang, M.H., Chung, Y.K.: Applications of thermal image and extension theory to biometric personal recognition. *Expert Systems with Applications* 39, 7132–7137 (2012)
20. Shahin, M.K., Badawi, A.M., Rasmy, M.E.M.: Multimodal Biometric System Based on Near-Infra-Red Dorsal Hand Geometry and Fingerprints for Single and Whole Hands. *World Academy of Science, Engineering and Technology* 56, 1107–1122 (2011)
21. Choras´, R.S., Choras´, M.: Hand Shape Geometry and Palmprint Features for the Personal Identification. In: *IEEE Proc. of 6th Intl. Conf. on Intelligent Systems Design and Applications*, pp. 1085–1090 (2006)

22. Wang, R., Ramos, D., Fierrez, J.: Latent-to-full palmprint comparison based on radial triangulation under forensic conditions. In: IEEE Proc. of International Joint Conference on Biometrics, pp. 1–6 (2011)
23. Spaun, N.A.: Forensic Biometrics from Images and Video at the Federal Bureau of Investigation. In: IEEE Proc. BTAS, pp. 1–3 (2007)
24. Weingaertner, D., Bellon, O.R.P., Silva, L., Cat, M.N.L.: Newborn's Biometric Identification: Can It Be Done? In: Proc. VISAPP, vol. (1), pp. 200–205 (2008)
25. Shahin, M.K., Badawi, A.M., Rasmy, M.E.: A Multimodal Hand Vein, Hand Geometry and Fingerprint Prototype Design for High Security Biometrics. In: IEEE Proceedings of Biomedical Engineering Conference (CIBEC), pp. 1–6 (2008)
26. Guo, J.M., Liu, Y.F., Chu, M.H., Wu, C.C., Le, T.N.: Contact-Free Hand Geometry Identification System. In: 18th IEEE Intl. Conf. on Image Processing, pp. 3185–3188 (2011)
27. Ross, A., Govindarajan, R.: Feature Level Fusion Using Hand and Face Biometrics. In: Proc. of SPIE Conf. on Biometric Technology for Human Identification II, Orlando, USA, vol. 5779, pp. 196–204 (2005)
28. Hanmandlu, M., Grover, J., Madasu, V.K., Vasirkala, S.: Score Level Fusion of Hand Based Biometrics Using T-Norms. In: IEEE Conf. on Technologies for Homeland Security (HST), pp. 70–76 (2010)
29. Lemes, R.P., Bellon, O.R.P., Silva, L., Jain, A.K.: Biometric Recognition of Newborns: Identification using Palmprints. In: IEEE Intl. Joint Conf. on Biometrics, pp. 1–6 (2011)
30. Fierer, N., Lauber, C.L., Zhou, N., McDonald, D., Costello, E.K., Knight, R.: Forensic identification using skin bacterial communities. PNAS 107(14), 6477–6481 (2010)
31. Fierer, N., Hamady, M., Lauber, C.L., Knight, R.: The influence of sex, handedness, and washing on the diversity of hand surface bacteria. PNAS 105(46), 17994–17999 (2008)
32. Yang, F., Ma, B., Wang, Q.X., Yao, D., Fang, C., Zhao, S., Zhou, X.: Information Fusion of Biometrics Based on Fingerprint, Hand geometry and Palmprint. In: IEEE Workshop on Automatic Identification Advanced Technologies, pp. 247–252 (2007)
33. Kacker, L., Varadan, S., Kumar, P.: Study on Child Abuse: INDIA 2007. Ministry of Women and Child Development, Government of India (2007)
34. FBI website, <http://www.fbi.gov>

# A Review on Age Identification Techniques for Non-human in Forensic Anthropology

Nur A. Sahadun<sup>1</sup>, Mohammed R.A. Kadir<sup>2</sup>, and Habibollah Haron<sup>1</sup>

<sup>1</sup>Faculty of Computing,

<sup>2</sup>Faculty of BioScience and Medical Engineering,

Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia  
afiqah\_sahadun@yahoo.com, {rafiq, habib}@utm.my

**Abstract.** Forensic Anthropology is an application of anthropology techniques to modern human and non-human remains for law enforcement. In general, the forensic anthropologist provides a basic biological profile of the decedent to aid in identification. This biological profile usually includes age, sex, height, ancestry, and postmortem interval. Identifications of the age for un-known birth date non-human provide useful information for variety of circumstances. A comparison of different techniques for age determination on non-human for certain species with regard to their accuracy based on published original data can be performed only with severe limitations. Once the age is known, it can be used for further information on determination gender, ancestry, stature, knowing the time and cause of death of corpse. This paper presents a review on techniques in identifying age for non-human cases regardless the specimen of data. Focusing only on age determination, it covers all range of techniques from physical maturity measurement on growth, development of dentition, development of lenses protein, mathematical computation and soft computing and artificial intelligence approach. There are five aspects in age determination will be discussed namely issues and problems, parameter in measurement, specimen, techniques used, and methodology. The paper is ended with conclusion that leads to a proposal on new trend and techniques to determine age for non-human in forensic anthropology.

**Keywords:** Forensic Anthropology, Age Determination, Skeletal Remains, non-human, dentition, lenses protein, Behavior, Soft-Computing Techniques.

## 1 Introduction

Age determination has not been widely adopted and fully explored particularly in forensic cases. In recent online search engine of the Scopus digital library, back from 2005 until 2013, we obtained about 27 619 entries with the keyword *age determination*. However, with the keyword *age determination forensic*, we found about 2 215 entries and obtain 466 entries keyword *age determination forensic anthropology* (only makes up 1.7 % of total age determination analysis entries). Each researcher has a different definition and standard for age determination. Age determination at death is exhibiting its growing potential also with the issues of identifying living individuals,

forensic anthropology, medico-legal, forensic medicine, physical anthropology and anatomy. In many cases that occur, forensic anthropologists will be tasked to analysis distinguishes human material or nonhuman, determination age at death, sex, ancestry, living stature and evidence of foul play and any other details that may help identification. Thus, major goal of forensic anthropologist is twofold: to contribute to positive identification and to assess what happened the human or nonhuman, including trace evidence of foul play. Forensic anthropologist work must address the search and finding the right bone and with issues such as identification and detection of signs of trauma that could lead to establishing the cause and manner of death. Age determination is one of the main tasks of a forensic practitioner. Determination of the age of nonhuman may be a priority in some cases. For forensic investigations, it is important to create a better method for determination the age of the various elements are more likely to survive and be restored. Accurate assignment of age at time of nonhuman death is important for life history in population research and clinical practice. Many skeletal traits in nonhuman have been investigated in studies of age determination of nonhuman skeletons. There some techniques that have been use in age determination such as laboratory base (Bagi et al., 2011), Physical maturity measurement on growth, Development of bones and dentition (Bolanos et al., 2000) (Andrews, 1982). However, many factors, such as genetic factors, endocrine factors, growth, health and lifestyle, nutrition, activity, ultimately affecting these indicators erratic manner. The results of age determination are affected by subjective factors, which cause considerable divergences between observers. Current age determination research focuses on diagnosis time; related to how long corpse will determinate in age group. While most existing forensic anthropology analysis aims at improving the classification rate by extracting the useful knowledge from forensic data, either through existing techniques or through development of new techniques.

The issues that occur in current age determination in forensic research include conflicting and definitions of types of age determination. Therefore, this chapter aim is to provide a comparative study on the current state of the art in age determination approaches together with its tools. In order to achieve this aim, five main aspects focused in this chapter are:

i. Age determination Issues and problems

There are various issues and problems regarding the age determination. This chapter attempts to unify existing issues and problems.

ii. Age determination parameter in measurement

Various parameters in age determination measured have been applied in previous study. This chapter also looks into parameter in measured for age determination purpose.

iii. Age determination specimen

There are many type specimen has been widely adopted in determination. These comparisons show the evolution of specimen in age determination for forensic anthropology field.

iv. Evaluation of existing techniques

Many techniques have been proposed and implemented as age determination tools in order to determine age. This chapter aims to find out the best techniques that can be used for a comparative study. These techniques are compared and evaluated based on their strengths and weaknesses.

v. Current methodology for age determination

This chapter looks into the methodology age determination. The results are recorded and analyzed based on age determination performance.

## 2 Background

In the area of forensic anthropology represents the application role of knowledge with techniques of physical anthropology to solving medical legal enforcement. The biological details such as age, sex, race or ethnicity, and stature are often the first pieces of data that help focus the investigation on specific group characteristics. The success forensic performance can be achieve when correct match determination with documentation report details and help solve remains problem, especially with regard to the evidence of foul play (Imaizumi et al., 2002). Biological characteristic such as growth layer groups in teeth (Boy et al., 2011), growth pattern of the infant mandible (McGrory et al., 2012) and tooth wear (Cuozzo and Sauter, 2006) on Peale's dolphin, Rhesus Monkey (Gavan and Hutchinson, 1973), Marine mammals (Bowen et al., 1983) by harp seals and wild life (Gee et al., 2012) by White-Tailed Deer focus the search within specific age, sex, ecology, behavior of species, and study life history. Age determination for nonhuman such as domestic and wild population is an important technique in a variety of settings, particularly where age is a pre-condition for access to infectious diseases like Cysticercosis, Hydatidosis, Tuberculosis and anthrax. Beside that age determination allow to understand the ecology and behavior of the species and studies of population dynamics different generations need to be recognized. Being able to determine accurately age of nonhuman is essential to the study of population nonhuman. Determination of age faced by many osteologists and forensic researchers are often called upon to determine the age at death of skeletal remains. Whichever of these may be of immediate concern, one problem is essentially the same. How accurate is the age determination. We see the evolution of forensic anthropology happened in many years, showing the rapid growth of forensic field. Numerous investigators have addressed themselves to the problem of age determination. Many tables, equations and graphs plot using a variety of criteria are now available for a variety of species based on (Gandal, 1954) review. Often samples are used to connect these criteria for chronological age was just a sample of the standard can be checked for accuracy. when checks are made on independent samples subsequently disappointing results (Gavan and Hutchinson, 1973). This causes difficulties to determination age and increase accuracy rate. The problem of age determination is:

- i. Lack of data collection  
(Ramsthaler et al., 2010)

Forensic anthropologists often bemoan the lack of modern spinal series

known subjects with gender, age, ethnicity, demographics and quality of bone is unknown.

ii. Lack of equipment

Since forensic anthropology growth fast, demanding the forensic modern techniques. They demonstrate the accuracy of the technique for age determination but cautioned that the most accurate technique is not necessarily the one that should be used if time is limited or if the need specialized equipment is lacking.

iii. Sample size

Since computational apply in forensic, there are varieties of computer models have been developed in the area of machine learning and statistics that can be used for classification forensic identification task outcomes. The statistical technique still lack in term of small sample

iv. A complex procedure

The gaps appear in Forensic Anthropology is how to obtain quick and easy of process diagnosis identification besides maximum percentage accurate result parameters. The total examination procedure is complex.

v. Time consuming

In order forensic practitioner to make improvement to existing traditional age determination method, additional time and attention are needed in understanding the existing age determination technique and concerns that need to be implemented.

### 3 Age Determination

#### 3.1 Scenario Issues and Problems

Determination of skeletal age is based on standards and methods developed from collections of skeletons of skeletons with complete documented biological data, such as age, sex and species. There are two methods for age determination; known age and unknown age. (Kohn et al., 1997) noted estimates of the chronological age for non-human of unknown age provides useful information for medical, demographic, and evolutionary studies. The similar situation to researcher (Ramsthaler et al., 2010), the unknown skeletal identity should as a basic framework for further development of modern methods of osteological. (Kohn et al., 1997) study on three calitrichid species namely: *Saguinus fuscicollis*, *Saguinus Oedipus*, and *Callithrix jacchus* from New World family of primates monkeys, including marmosets and tamarins. The study using 22 epiphyseal sites with known age range 0 year to 3 years as subject of age determination. (Barlow and Boveng, 1991) believe with limited information expected patterns of mammalian survivorship to define a general technique of age specific mortality rates. The study using fur seal age 18 years, Old World monkeys age 34 years and human females ages 81 years as age indicators provide expected result in

termination of age determination. In USA, researcher (Lubinski, 2001) point to Pronghorn antelope (*Antilocapra Americana* Ord) to age determination of wildlife nonhuman species with unknown age. (Boy et al., 2011) Peale's from dolphin species off southernmost South America as subject experiment and was carried out without knowing the real life age-term species, however, the age distribution identified through a process of comparison with the maximum age of the similar species. When samples of non-human taken from a population, it usually is impossible to assign an actual age of any specimen unless the birth process has been observed and unique identification marked for later identification (Schmitt, 2004). An absolute age can often be determined by the incremental growth line structure (Leyssac and Madsen, 2001). Ordinarily, a relative age can be given to specimen based on a comparison with other specimen in the sample (Kohn et al., 1997). Standard procedure for result absolute and relative age shall be standardized by study of nonhuman known age. Moreover, age determination of non-human is distinguished based on habitation population and sex. (Castillo and Ruiz, 2011) Claim, men are usually stronger than women in the morphology of changes that occurred during development of skeletal.

Age determination requires sophisticated equipment and often a long determination time. (Augusteyn et al. 2003) study on determining kangaroo age from lens protein content. In Australia around year 2003, using protocols for data collection at both sites were the same as at *Hattah-Kulkynne* for calculation lens mass and protein content for older age. While body weight, foot length, leg length as age indicator for pouch young. The demonstrate accuracy of this technique for aging kangaroo but cautioned that the most accurate technique is not necessarily the one that should be used if time is limited or if the need specialized equipment is lacking.

Sample size determine as a common task for organizational researchers. Inappropriate, inadequate, or excessive sample size directly influences the quality and accuracy of research. Modeling skeletal across different ages amounts to building computational models for skeletal aging that account for a multitude of factors such as; age group, gender, species, weight loss, etc. With growing technology needs in computational method involvement in forensic field, there are varieties of computer models have been developed in this machine learning and statistics area that apply in age determination outcomes. Several limitations on small sample size may be drawn from the study by (Stauber and Müller, 2006), the number of samples was too small for showed an age-related linear trend. Some of them have been done involving structural deterioration of monkey bone tissue (Havill, 2003). The author used a small number of monkey age 26 years, it is difficult to comment on bone tissue pattern values by t-tests. (Hans et al., 1995) applied multiple regression statistical model due limited sample size did not give enough power to the study to reach statistically significant age related bone loss correlations. Study from (du Jardin et al., 2009) agreed since the sample size is relatively small which 76 femurs, the goal study was not to design forecasting models that could be used in practice, but to estimate the accuracy of neural network, discriminant analysis and logistic regression. Support by (Gibson, 1993) from United Kingdom gain result with set of Neural Network give initial result based on tooth attrition image and Gaussian highlight as filtered for age determination for using a larger set in term of increase success rate result. Other study; see (Buk et al., 2012) by Radial Basis Functions (RBF) Neural Networks give guaranteed result with

large and wide range group age. Researcher (Stauber and Müller, 2006) claim the results must be interpreted carefully since the sample size was relatively small.

The work evolution of forensic anthropology starting conducted in laboratory based traditionally by the anthropologist. The anthropologist's United in Kingdom (Stander, 1997), Japan (Wada et al., 1978), USA (Lubinski, 2001) and Norway (Kranioti and Paine, 2011) contribution traditional forensic scenario method deals for anthropologist's (Cattaneo,task 2007) in define autopsy traditional anthropology is not enough to meet forensic requirements context. Identification is a complex process (Grabherr et al., 2009) in cases where corpse remains are rendered unrecognizable (Zeybek et al., 2008) by advanced decomposition (Bruning-Fann et al., 2001), or are completely skeletonized (Sakuma et al., 2010), or there is fragmentation (Tocheri and Molto, 2002) of the body. In year 2004, starting research by (Craig et al., 2004) (Grabherr et al., 2009) consider deceased persons for determination age and gender in 3D virtual skeleton by multidetector computed tomography have abilities in easy storing, handle structure sample and future study. In the mid-twentieth century, researchers began to look at the architecture of trabecular (Macdonald et al., 2011) or cancellous bone, to measure degree of bone loss and to diagnose osteoporosis. Based on this research, anthropologists started to ask whether or not trabecular bone loss was consistent enough to be used as an indicator of age at death. (Craig et al., 2004) Found thirty 3 dimension distal femoral and proximal tibia growth plates have complex anatomy for show relationship results between growth plates change with age. An enhanced version was proposed to overcome the forensic anthropology issue is how to obtain quick and easy of process diagnosis identification besides maximum percentage accurate result parameters. The total examination procedure is complex. Because of the increasing lack of recent bone collections, ethical issues concerning maceration procedures, and progress in radiological imaging techniques, computed tomography (CT) scans offer an alternative to traditional anthropological bone collection (Ramsthaler et al., 2010).

Time consuming been subject for issue in year 1997, (Stander, 1997) study on age determination by nineteen teeth of leopards were measured using vernier callipers, and body measurements were taken with a tape measure give for ecology and behavior knowledge of a leopards species. Researcher admits the data collection characteristics are lacking which requires the expertise. In addition, tape measure methods are time consuming, to complete study take as long as 5 years (1991 until 1995). In term of time consuming same feeling to other researcher (Rösing et al., 2007) by recommendation skeletons methods for examining in forensic practitioner, (Tassani et al., 2012) by tips for time-consuming issues to applicable in limit scale analysis, (Augusteyn et al., 2003) by suggest eye lens as an alternative method for age determination from use the mandibles methods which is a time-consuming process. (Quimby and Gaab, 1957) study, 168 known age and 527 assigned age mandibular dentition of rocky mountain elk calf. Researcher argued that the study failed to select the best characters results. (Pietka et al., 1991) applying computerized approach as solution for age determination problem where are misclassification by atlas method based (Bull et al., 1999). The study argued positive identification achieve depends match an atlas pattern and give less robust for unmatched pattern. Claim by (Pietka et al., 1991) Tanner and Whitehouse (TW2) method result more robust but high complexity process.



## 4 Evaluation of Existing Techniques

The focus reviews on age determination for nonhuman evolution since year 1982 with highlight to abilities and limitation each technique for development of lenses protein, physical maturity measurement on growth and development of dentition.

### i. Development of Lenses Protein

In Australia (Augusteyn et al., 2004) around year 2003, using protocols for data collection same as at *Hattah-Kulkyne* in 1992 (n = 209) and 1999 (n = 245) and from pouch young (n = 64), obtained in the 1999 cull. Total 556 lenses from 40 red (*Macropus rufus*), 476 western grey (*M. fuliginosus*) and 57 eastern grey (*M. giganteus*) kangaroos, ranging in age from 3 days to 20 years for lenses protein sample. While sex, head length and foot length as physical scale measurement for pouch young parameters were recorded. Development of lenses protein in term of age related involve several procedures, start with collect eyes by professional shooters and Parks Victoria staff by shot in the kangaroos head and label with numbered ear tag. The eyes were removed by dissection and stored at ambient temperature.

Then, all of a severed head from the neck placed in nylon mesh bags individually and were buried in sandy soil, extracted several months later and cleaned for measurement of MI. Lenses from 44 eastern grey kangaroos were obtained from a cull conducted by consultant biologists at Government House, Canberra, in 1993. The kangaroos were captured using tranquilliser darts then euthanased by lethal injection. Lenses from 13 eastern grey kangaroos culled at Portland Aluminium Smelter, Victoria, in 2001, conducted by smelter staff were also included in the study. The same protocols also apply for calculation lens mass and protein content for older age. Meanwhile, ages of pouch-young were estimated from head and foot length, using sex-specific growth equations calculated by Poole Method (1982) for western grey kangaroos, and read off graphs presented by Sharman Method for red kangaroos. Ages of older animals were calculated from MI, using Kirkpatrick's (1965) equation for eastern grey kangaroo for both grey kangaroo species, and Kirkpatrick's (1970) equation for red kangaroos. Lenses were removed through an incision made in the limbal region of the eye, freed from adhering vitreous and pigmented tissues, gently blotted dry and weighed. Obviously damaged lenses were discarded while most of the intact lenses were placed in 5.0 mL phosphate-buffered saline (PBS) and homogenised by sonication. Protein contents of the extracts were then determined by the Lowry method. The result of the study found no dissimilarity is observed in the masses of the 725 red and western grey kangaroo. Classless appear in gender to have size lens at the same age with regardless of body size. Addition, researcher detects variability in situation lenses from older animals and with lenses that had been stored that lens size increases asymptotically with age. While, the protein content of the lens is also less variability parameter and also increases asymptotically with age, suggesting that there may be a finite limit to the amount of protein that can be accommodated in the adult lens. Furthermore, there is no dissimilarity of

protein contents by differ storage 1 year with those from freshly assayed lenses. However, 3/10 of the protein had been lost when doing examination of lenses removed from 44 eastern grey kangaroo heads that had been stored for 8 years and repeatedly thawed and frozen indicated. No different when all data were combined on experiment on curve fitting of lens protein data from 177 western grey kangaroos collected in the 1999 cull (256 lenses) was undertaken to determine the relationship between lens protein content and age. Final result shows the relationship between total protein contents (in milligrams) and age (in years estimated from molar index) with a single equation for all three species over the whole age range from newborn to adult. Result found all body parameters were found to increase with age with achieve best correlation over the whole age range with  $R=0.9963$  for western grey kangaroo, while the red kangaroo data could also be adequately described with same equation for western grey kangaroo and the same values for the constants. The 26 eastern grey kangaroo lenses obtained  $R=0.9983$ .

## ii. Behaviour

Researcher from African (Evans and Harris, 2008) was focus on their study on adolescence in male sex African elephants, *Loxodonta Africana*. The study shows significant knowledge within female elephant social by using Mann Whitney test. The researcher collected data from a population of free-ranging African elephants from February 2002 to February 2005 in Wildlife Management Area NG26 with an estimated population of 1576 elephants within the study area; of these 333 males were individually recognized. There were five main habitat types: open grassland/floodplain, mopane (*Colophospermum mopane*) woodland, mixed woodland (*Acacia* spp., *Combretum* spp.), *Terminalia* (*Terminalia sericea*) woodland and island vegetation (dominant plants were mainly *Hyphaene petersiana* or *Phoenix reclinata*). The researcher was collected focal and data in the Okavango Delta Botswana in order to assess behavior and social interactions during adolescence. Using binoculars from a minimum distance of around 50 meter and 500 meter observe the elephants. There are four age in adolescence group namely group 1 (10-15 years of age), group 2 (ages 16-20), group 3(ages 21-25) and group 4 ( $\geq 36$  ages). Technique use for age calculation is using footprint measurement to calculate shoulder height and a combination of visual clues such as length of tusks, tusk girth and head shape. The parameter use is analyzed the effects of season and age on the frequency of social behaviors', greeting, sparring/ playing, vocalization and distance to nearest neighbour separately using nonparametric statistics.

The study summary, closer to older elephants gave many experiences and high rate social. Season affects the rate of social interactions in many mammals and, whereas greeting in male elephants was affected by season, sparring was not, highlighting the importance of this activity and the importance for male elephants of asserting their dominance at every opportunity. The rate of vocalization per half-hour focal was not affected by season. Season did not affect the distance to nearest neighbor, but age did, with median distance increasing. Addition, researcher suggest study that mature males are reservoir of knowledge in society bull.

### iii. Physical Maturity Measurement on Growth

Physical maturity measurement on growth approach often used based on observation that have done (Archie et al., 2006). Experiment have done in Kenya, cover issues of size and logistic that difficulties associated by capture of very large mammals species; African elephants, *Loxodonta africana*, and developed ways to estimate ages. With this technique, five age groups 10 to 15 years, 16 to 20 years, 21 to 25 years, 26 to 35 years. The footprint length measurement method with shoulder height data that had been used previously Laws 1969 (Archie et al., 2006) to determine the ages of elephants. Process developed from adult female gender of known or estimated age. Research project in 1972, all individually recognized elephants recorded the dates of birth in month and year. Others born after 1972 were positively known, and dates of birth for individuals born before the study began were estimated by multiple ageing methods include footprint length, back length, shoulder height or visual estimates based in head morphology. C.J.M was use to estimate ages of 374 females and infants in Tarangire. Twelve elephant body dimensions were checked against tooth eruption and wear. Social interactions between adult females were recorded in two ways, first all-occurrence records during 20-min focal animal samples and second way ad libitum sampling during observation sessions on family groups. The sample collected during separate time periods. The study complete with result derived from 478h and 488h of both sampling. There are two measurement in describe female rank namely: (1) transitivity of relationships across multiple dyads, measured as the number of circular dominance relationships within families and (2) the degree of symmetry within dyadic relationships across two or more agonistic interactions, measured with the 'directional consistency index' (DC index).

(Rozenblut and Ogielska, 2005) The research from Poland development and growth of long bones in three species European juvenile frogs and adult female frogs water frogs. According to Polish legal regulations concerning wild species protection. Femora, metatarsals, and phalanges of the hindlimb digits were dissected. Methodology research modified from (Castanet et al., 1993). The thin sample bone preparation need to clean from soft tissues then fixation. Finish, bones were decalcified by liquid chemical. The researcher obtained best results after 4 h for phalanges and metatarsals of adults, and 2–3 h for juvenile bones. Decalcification of femora lasted 30 min for tadpoles and 5–6 h for adults. By using microscopy, the samples were scan slice into 12- $\mu$ m thick sections. A complete series with their epiphyses, were collected. The sections were stored at room temperature up to 12 months. Then, image analysis process and measurement to the bone diameter. The results were calculated as equivalent circle diameters (ECDs). The lengths of diaphyses in total bones were measured with digital caliper to the nearest 0.01 mm. Addition, similar patterns of differentiation on femora, metatarsals and phalanges of hindlimbs. Researcher found trabeculae formed within the boundary zone in 4- and 5-year- old frogs of the same and closely related species, *R. lessonae* and *R. ridibunda*, but not in younger individuals. In summary, both longitudinal and radial growth of lissamphibian long bones is the result of periosteal activity at the epiphyseal edges and around

the diaphysis, respectively. Metaphyseal cartilage has no role in the elongation of the bone, and the only probable effect of its activity is in the enlargement of the marrow cavity and in growth of the cartilage itself.

#### iv. Development of Dentition

In year 1997 (Stander, 1997) study on age determination in United Kingdom have been developed by nineteen teeth of leopards were measured using vernier callipers, and body measurements were taken with a tape measure give for ecology and behavior knowledge of a leopards species. All data on tooth eruption and wear came from live leopards that had been immobilized with a combination of Zoletil and xylazine hydrochloride. Measurement done with using vernier callipers, and body measurements were taken with a tape measure and body length was measured from the tip of the nose to the tip of the tail. Process on leopards run with the Mann-Whitney U with two tailed test used to testing in this study experiment. Result study was representing that male leopards were larger and heavier than females. Both maxillary and mandibular canines in males were longer than in females. The study defines after the age of four years the extent of tooth deterioration between the sexes. Male tooth wear were more prone to flaking of enamel layers and to broke canines. At an estimated age of four years, 71% of seven males had one broken canine and by the estimated age of five to eight years, 57% of seven males had at least one broken canine. Addition, females appeared less prone to tooth breakage as only 17% of six females between the estimated ages of two to seven years had a broken canine tip. Researcher argued tooth eruption and growth vary significant between the sexes, age and population. Researcher admits the data collection characteristics are lacking which requires the expertise. In addition, tape measure methods are time consuming, to complete study take as long as 5 years (1991 until 1995). Research claim that ageing criterion is an important tool in wildlife studies and management, but the level of error in most systems for estimating age needs to be considered.

Mandibular dentition age indicator in Montana (Quimby and Gaab, 1957) study, total 168 known age and 527 assigned age mandibular dentition of rocky mountain elk calf. Animals were trapped and tagged during their first winter at ages ranging from about 6 to 8 months (elk of this age are readily recognized at close range by experienced workers). The study using type of teeth, number of teeth, and wear on permanent teeth as age indicator features. The development phase started with data preparation by selecting quality sample. There are three samples are removed. Then followed aging techniques are done in three ways namely: direct comparison with known-age jaws, comparison with assigned-age jaws and comparison with good drawings or photographs of known age jaws. The combination characters II, III, V appeared to be of more value than any of the others. Researcher argued that the study failed to select the best characters results but characters used are indicative of wear of the 3-, 4-, 5-, 6- and 7 year olds respectively are positive for combination II, III, V of 4 year old characters. (Wada et. al., 1978) noted that age determination with grade of wear teeth of right maxilla and mandible can be done

on Japanese monkeys with known age. From experiment working at Japan, upper and lower molars from 14 skull specimens of Japanese monkeys as point of study. All subject comprised ten individuals of *Macaca fuscata fuscata* and four individuals of *M. f. yakui*. Dental tissues would be far more accurate than such a macroscopic one, but it has its own difficulties and limitation of effectiveness. The study consist two sections are; grinding a tooth on a whetstone by hand then stained with hematoxylin were observed in transmitted light under a microscope, second section is it was decalcified and embedded in polyester resin sectioned at a thickness of about 20 microns with the hard tissue microtome. The sections were decalcified with Planck-Rychlo's solution, and stained with Delafield's hematoxylin. Then counting the number of annual growth layers in tissues of the teeth, in the mandible and in other structures. Researcher Wada claim that method of age estimation by means of histological features in dental tissues would be far more accurate than such a macroscopic one. Researcher suggest in preparation phase of specimens, the hard tissue microtome is better to use, while to use ground sections without staining is a simple and practical method for preliminary estimation of age.

#### v. Bone Loss

Craig (Craig et. al., 2004) studied age related changes using three cross sections of distal femoral and proximal tibia growth plates have complex anatomy for show relationship results between growth plates change with age. The collection data have done with seven distal femurs, eight proximal tibias in eleven patients with 3D models were created using an offline independent 3D workstation. Methodology for study is including process data preparation, pre-size sample into block form, three dimensional modeling and yield technique. Parameter focus for age related trabecular growths are femoral growth plate area and tibial growth plate area. Research by (Craig et. al., 2004) found thirty 3 Dimension distal femoral and proximal tibia growth plates have complex anatomy for show relationship results between growth plates change with age. The collection data have done with seven distal femurs, eight proximal tibias in eleven patients with 3D models were created using an offline independent 3D workstation. This study was focus on growth plate criteria; growth plate area and volume at the knee.

Age determination (Lubinski, 2001) paper in brief, study new methods of scoring tooth eruption and wear have been developed and have been applied to a sample of over 500 pronghorn mandibles to obtain improved eruption and wear schedules. Based on new stage tooth cusp eruption modified by researcher were determined by measuring the height of the tooth crown above the alveolar surface, and then scaling the erupting tooth to the adjacent fully erupted tooth. The measured erupted height was record from the alveolar surface to the top of each crown along its lingual side in a direction approximately perpendicular to the alveolar surface of mandible. When cusp reached the height of adjacent teeth or exhibited exposed dentine, full eruption for a cusp was recorded. The measure were nearest 0.1 mm, with sliding Vernier calipers were taken. Then, recorded tooth wear pattern as appropriate, for each mandibular cheek tooth in four categories; 1) number of fossettes or infundibula 2) tooth wear stage 3) tooth wear

diagram and code and 4) a new tooth wear scoring system based in principle, on the scoring system of Brown and Chapman (1990). The criteria selected of each sample and recorded. The study method more easily, beside greatly improves the sample size of eruption and wear criteria for age determination of pronghorn. The study suggests that zooarchaeologists will have to work more closely with wild-life biologists and managers to obtain additional mandibles from known-age animals. Addition, crown height measurements may profitably be applied to archaeological pronghorn studies in the future.

## 5 Other Techniques

### 5.1 Mathematical Computation and Soft Computing

Age determination using mathematical computing, approach for nonhuman was used along 1973 like Trust-Region nonlinear least squares by (Steele and Weaver, 2012), study from Unites States find Rocky Mountain elk mandible and Montana ungulates mandible contribute for age determination by using Trust-Region nonlinear least squares as mathematical approach add on to classical technique before. Researcher investigates wear stages to determine elk age-at-death (tooth crown height and age) and investigate the use of cementum annuli. Sample of mandibles from 226 Rocky Mountain elk collect by two sources; Quimby and Gaab (1957) with Hamlin (2000). The crown height measurements were recorded on all specimens. MATLAB's curve as accuracy index for show correlation between age indicator with features.

The linear regression, previous researchers (Ding et al., 1999) have done study in age related bone growth. The study sample on 40 human proximal tibiae age from 16 to 85 years was taken from a Caucasian donor. There were ten women and 30 men and all had died suddenly either from trauma or acute disease. The sample stored at -20°C in sealed plastic cylindrical size. All samples keep them moist during scanning and testing. The sample were scanned with high resolution micro-CT system, after scanning the micro-CT images were segmented using individual optimal thresholds to obtain the accurate 3D datasets. In calculation of 3D microstructural to obtain Structure model index (SMI) parameters. Other parameter use in this study are: Degree of anisotropy, trabeculae per volume in mm<sup>-3</sup>, 3D volume-weighted trabecular thickness in mm, 3D trabecular spacing in mm, bone surface density in mm<sup>-1</sup> (bone surface area per total volume of specimen), and bone surface-to-volume ratio in mm<sup>-1</sup> (bone surface area per bone volume) were calculated as were the mean trabecular volume in μm<sup>3</sup> and the mean marrow space volume in μm<sup>3</sup>. Mechanical testing compressions have done to obtain strain rate and Young's modulus. Then sample need to measure it density, collagen and volume fraction. These data were further divided into three age groups: young (16 to 39 years), middle (40 to 59 years) and old age (60 to 85 years) to assess the multiple associations among properties in different age groups. Pearson's correlation coefficient show result among various properties of bone samples.

Experiment have done in United States of America (Havill, 2003) collected on ba-boons (*Papio hamadryas*) involve 466 females and 210 males with age range between 5.5 to 30 years. The measurement of skeletal mass obtained by dual energy X-ray

absorptiometry (DXA), which provides an estimate of the amount of bone mineral (g/cm<sup>2</sup>) in a region of interest. All DXA measurements were recorded by manufacturer's software. Each sample were scanned used Auto Width with setting speed set depend on the thickness of sample. The result study is produced in six measurements for axial skeleton. Then the sample will through the process styloid. The study evaluated effect of age on bone at each site was examined by t-test, scatter plots and regression.

Since year 1973, researcher (Gavan and Hutchinson, 1973) determine problem of age by using Rhesus Monkeys population. The study was using Schultz (1933) method. This method based on chronological age is known and that the most reasonable estimate of the measurement is desired. The parameters on this study focus on weight, sitting height and number of teeth. The ability of this method obtained from evaluated mean and standard error calculated. Study using t-test to obtained mean value.

Multiple regression by (Thomas et al., 2000) applied statistical technique for age determination. Total data 50 men and 46 women from adult samples of bone of known ages 21-92 years. Physical maturity parameter in this study included supine length height (cm), weight (kg), sex, and cause of death. In the sample preparation phase the femoral sample pre size in block form. Macroscopic measurements, moments of area, microscopic measurements and histological parameters are recorded for all sample. Four multiple regression were applied for age determination and obtained satisfactory result.

## 5.2 Artificial Intelligence

(Corsini et al., 2005) study apply artificial intelligence for identification task by the result Neural Network better in youngest and oldest group than intermediate age class. (Gibson, 1993) from United Kingdom gain result with set of Neural Network give initial result based on tooth attrition image and Gaussian highlight as filtered for age determination. Gibson study had done by analysis sample image of 2 Dimension for tooth attrition of ungulates. The parameter working for this study namely: degree of tooth wear, tooth descriptor and its position within the image. The result tested show an 85% success accuracy.

Other study (Buk et al., 2012) by Radial Basis Functions (RBF) Neural Networks give guaranteed result with wide range group age. Phase sample preparation on 10 samples of adult pelvic bone which total of 955 individuals, ranging age from 19 to 100 year. Parameter consider for the study is Pubic symphysis, Posterior plate, Ventral plate, Dorsal lip, Surface sacro-pelvic, Transverse organization, Texture and porosity, Apical activity and Retroauricular activity. The evaluation performance data view by self-organising maps (SOM), confusion matrix and ROC curves. The summary result show accuracy achieve higher rate. Table 1 shows the comparison of existing age determination techniques.

**Table 1.** Comparison of existing age determination techniques

<b>County</b>	<b>Approach/Model</b>	<b>Strength</b>	<b>Weakness</b>
Australia (Augusteyn et al., 2004)	Development of lenses protein	-kangaroo lens could provide a much more reliable estimate as long as the samples have been properly stored -body parameters were found to increase with age -less variable for protein content parameters measured. -Success accurate	-physical scale measurement difficult stored -limited extrapolation of ~465 mg for eastern grey -time is limited with if the need specialized equipment is lacking.
Africa (Evans and Harris, 2008)	Behavior	- The pattern of association was not limited to adolescents -Find adolescent males (ages 16-20) showed a tendency for higher social levels. -Identify parameter season as affects the rate of social interactions in many mammals.	- Data collected in a limited period
Kenya (Archie et. al., 2006)	Physical maturity measurement	-Data of two separate studies encompassed the same behavior with scored in the same way outcome.	-Elephant range widely and unpredictably need opportunistic behavioral sampling scheme.



**Table 1.** (continued)

Poland (Rozenblut and Ogielska, 2005)	Physical maturity measurement	- Researcher found trabeculae formed within the boundary zone in 4- and 5-year- old frogs of the same and closely related species, <i>R. lessonae</i> and <i>R. ridibunda</i> , but not in younger individuals.	-Certain sample present only in the midlength of the diaphysis and effects model performance - The sample not long enough to reach the bone extremities
United Kingdom (Stander, 1997)	Development of dentition	- Tooth eruption and growth vary significant between the sexes, age and population.	- Time consuming in data collection phase
Montana (Quimby and Gaab, 1957)	Development of dentition	- Characters used are indicative of wear of the 3-, 4-, 5-, 6- and 7 year olds respectively are positive for combination II, III, V of 4 year old characters.	- failed to select the best characters results.
Japan (Wada et. al., 1978)	Development of dentition	- Histological features in dental tissues would be far more accurate than such a macroscopic - The period of formation of the first dark layer is related to the function or the kind of tooth.	- Time consuming in data collection phase
United States of America (Craig et. al., 2004)	Bone loss	- Expansion of trabecular bone growth plates occurs in a linear fashion with age	- Time consuming in data collection phase

**Table 1.** (continued)

United States of America (Lubinski, 2001)	Bone loss	- This study provides a much larger comparative sample	- The lack of adequate sample sizes for some age classes (e.g. 0.1–0.3) - The sample that need to be filled for more accurate estimates of age.
United States of America (Steele and Weaver, 2012)	Mathematical computation and soft computing	-The measurement more easy	-Strong correlation with age
Denmark (Ding et al., 1999)	Mathematical computation and soft computing	-Obtain significant p value. -The study found type of structure played an important role in determining the mechanical properties of cancellous bone. -The results must be interpreted carefully since the sample size in the young-age group was relatively small.	-Result of the correlation between Young's modulus and density was at the lower end but still within the normal expected range
United States of America (Havill, 2003)	Mathematical computation and soft computing	-The first study to characterize normal variation in a bone density in baboons.	- the diaphyseal radius and ulna failed in term to show correlate bone mineral density with age

**Table 1.** (continued)

Columbia (Gavan and Hutchinson, 1973)	Mathematical computation and soft computing	-Weight measure is easier than sitting height measure. - Accuracy by sitting height higher than weight.	-The figure difficult to determine which a good method is. -The study found that it is difficult to determine which variable is most likely to give the best -The lack of using number of teeth in term of time.
Australia (Thomas et al., 2000)	Mathematical computation and soft computing	-Obtained satisfactory result.	- The lack in term of small sample size
France (Corsini et al., 2005)	Artificial Intelligence	-Produce good rate classification	-Varying the parameter set learning and momentum rates, exploring other architectures (the number of hidden units, the connections) and including new skeletal samples
United Kingdom (Gibson, 1993)	Artificial Intelligence	- The experiment fails to highlight teeth boundaries.	- Small sample set of mandible
France (Buk et al., 2012)	Artificial Intelligence	-The result show accuracy higher rate -The study found variable "sex" is not important in age classification.	-Found biological indicators does not allow for a more precise age determination than three classes

## 6 Discussion

This section digests the derived results from the comparative of the literature search. Data collection using unknown age commonly for wild nonhuman, that wild born and imported or accessioned from the wild for museum collections, the date of birth is unknown. And otherwise species nonhuman such as domestic animal the details series

already known with gender, age, ethnicity, demographics and quality of bone is known because the species life one of part our environment in other word directly the basic knowledge such as gain from experience and easy to find with low cost with parallel to the time available to solve the issues. In short, characteristic target subject of the age determination is contributing in the process of age determination task for forensic identification as practitioner. In the area of forensic anthropology represents the application role of knowledge with techniques of physical anthropology to solving medical legal enforcement. The success forensic performance can be achieve when correct match identification with documentation report details and help solve remains problem, especially with regard to the evidence of foul play. Bone microarchitecture study is the study of the bone architecture based on width, number and separation of trabecular as well as on their spatial organization. Basically, there are common micro-architecture skeletal parts in wide use like researcher from United States of America and France was do identification investigation especially for age determination base distal femoral, proximal tibia and pelvic bone. From a clinical point of view, micro-architecture is an interesting aspect to study and define patterns of bone alterations with aging and pathology. Many skeletal traits in human and non-human have been investigated in studies of age determination. The work evolution of forensic anthropology starting conducted in laboratory based traditionally by the anthropologist. The anthropologist's

Kenya (Archie et. al., 2006), Poland (Rozenblut and Ogielska, 2005), United in Kingdom (Stander, 1997), Montana (Quimby and Gaab, 1957) and Japan (Wada et al.,1978) contribution traditional forensic scenario method deals for anthropologist's in define autopsy traditional anthropology process is not enough to meet forensic requirements. Researcher (Cattaneo, 2007) define traditional anthropology is not enough to meet forensic requirements context.

## 7 Future Research Directions

Age determination is a fast expanding research area in the research domain of forensic anthropology. It could be seen that almost forensic anthropology are focusing age determination on diagnosis time by introducing of new or enhanced age determination techniques or approaches that have been implemented. The aim of age determination is to discuss the concept and limitation of classification of age determination and the abilities of artificial intelligence to solve the classification problems. In addition, age determination of trabecular bone morphology properties can give new insight in field of forensic anthropology as we know human subject usually used for this purpose.

## 8 Conclusion

This chapter has digested the potential all techniques in focus domain forensic anthropology in term of correlation between age indicator and aging. This chapter has discussed the traditional and current forensic anthropology with techniques in age determination. Justification all concept, application, advantages and disadvantages of age determination contained in this reviews. Age determination from measurements

of skeleton using artificial intelligence, is a legitimate alternative in cases of badly fragmented, when other classic measurements are not available. Artificial intelligence can improve the age determination rate, with equally balanced results in both males and females when compared to linear classifier like such as linear regression.

## References

- Bagi, C.M., Berryman, E., Moalli, M.R.: Comparative bone anatomy of commonly used laboratory animals: implications for drug discovery. *Comparative Medicine* 61(1), 76 (2011)
- Bolanos, M.V., Manrique, M.C., Bolanos, M.J., Bri-ones, M.T.: Approaches to chronological age assessment based on dental calcification. *Forensic Science International* 110(2), 97–106 (2000)
- Andrews, A.H.: The relationship of bovine mandibular cheek tooth development to age determined by post-mortem radiographic examination of cattle aged between 12 and 24 months. *Journal of Agricultural Science* 98, 109–117 (1982)
- Imaizumi, K., Saitoh, K., Sekiguchi, K., Yoshino, M.: Identification of fragmented bones based on anthropological and DNA analyses: case report. *Legal Medicine (Tokyo, Japan)* 4(4), 251–256 (2002)
- Boy, C.C., Dellabianca, N., Goodall, R.N.P., Schiavini, A.C.M.: Age and growth in Peale's dolphin (*Lagenorhynchus australis*) in subantarctic waters off southern South America. *Mammalian Biology - Zeitschrift für Säugetierkunde* 76(5), 634–639 (2011)
- McGrory, S., Svensson, E.M., Götherström, A., Mulville, J., Powell, A.J., Collins, M.J., O'Connor, T.P.: A novel method for integrated age and sex determination from archaeological cattle mandibles. *Journal of Archaeological Science* 39(10), 3324–3330 (2012)
- Cuozzo, F.P., Sauter, M.L.: Severe wear and tooth loss in wild ring-tailed lemurs (*Lemur catta*): a function of feeding ecology, dental structure, and individual life history. *Journal of Human Evolution* 51(5), 490–505 (2006)
- Gavan, J.A., Hutchinson, T.C.: The problem of age estimation: a study using rhesus monkeys (*Macaca mulatta*). *American Journal of Physical Anthropology* 38(1), 69–81 (1973)
- Bowen, W.D., Sergeant, D.E., Øritsland, T.: Validation of age estimation in the harp seal, *Phoca groenlandica*, using dentinal annuli. *Canadian Journal of Fisheries and Aquatic Sciences* 40(9), 1430–1441 (1983)
- Gee, K.L., Holman, J.H., Causey, M.K., Rossi, A.N., Armstrong, J.B.: Aging white-tailed deer by tooth replacement and wear: a critical evaluation of a time-honored technique. *Wildlife Society Bulletin*, 387–393 (2002)
- Gandal, C.P.: Age Determination In Mammals. *Transactions of the New York Academy of Sciences* 16(6 series II), 312–314 (1954)
- Ramsthaler, F., Kreutz, K., Verhoff, M.A.: Accuracy of metric sex analysis of skeletal remains using Fordisc® based on a recent skull collection. *International Journal of Legal Medicine* 121(6), 477–482 (2007)
- Augusteyn, R.C., Coulson, G., Landman, K.A.: Determining kangaroo age from lens protein content. *Australian Journal of Zoology* 51(5), 485–494 (2004)
- Kohn, L.A.P., Olson, P., Cheverud, J.M.: Age of epiphyseal closure in tamarins and marmosets. *American Journal of Primatology* 41(2), 129–139 (1997)
- Barlow, J., Boveng, P.: Modeling age-specific mortality for marine mammal populations. *Marine Mammal Science* 7(1), 50–65 (1991)

- Lubinski, P.M.: Estimating age and season of death of pronghorn antelope (*Antilocapra americana* Ord) by means of tooth eruption and wear. *International Journal of Osteoarchaeology* 11(3) (2001)
- Schmitt, A.: Age-at-death assessment using the os pubis and the auricular surface of the ilium: a test on an identified Asian sample. *International Journal of Osteoarchaeology* 14(1), 1–6 (2004)
- Archie, E.A., Morrison, T.A., Foley, C.A.H., Moss, C.J., Alberts, S.C.: Dominance rank relationships among wild female African elephants, *Loxodonta africana*. *Animal Behaviour* 71(1), 117–127 (2006)
- Buk, Z., Kordik, P., Bruzek, J., Schmitt, A., Snorek, M.: The age at death assessment in a multi-ethnic sample of pelvic bones using nature-inspired data mining methods. *Forensic Science International* 220(1-3), 294.e1–294.e9 (2012)
- Bull, R.K., Edwards, P.D., Kemp, P.M., Fry, S., Hughes, I.A.: Bone age assessment: a large scale comparison of the Greulich and Pyle, and Tanner and Whitehouse (TW2) methods. *Archives of Disease in Childhood* 81(2), 172–173 (1999)
- Gibson, P.M.: The application of hybrid neural network models to estimate age of domestic ungulates. *International Journal of Osteoarchaeology* 3(1), 45–48 (1993)
- Hans, D., Arlot, M.E., Schott, A.M., Roux, J.P., Kotzki, P.O., Meunier, P.J.: Do ultrasound measurements on the os calcis reflect more the bone microarchitecture than the bone mass?: a two-dimensional histomorphometric study. *Bone* 16(3), 295–300 (1995)
- Havill, L.: Bone mineral density reference standards in adult baboons (*Papio hama-dryas*) by sex and age. *Bone* 33(6), 877–888 (2003)
- Rozenblut, B., Ogielska, M.: Development and growth of long bones in European water frogs (*Amphibia: Anura: Ranidae*), with remarks on age determination. *Journal of Morphology* 265(3), 304–317 (2005)
- Stander, P.E.: Field age determination of leopards by tooth wear. *African Journal of Ecology* 35(2), 156–161 (1997)
- Stauber, M., Müller, R.: Age-related changes in trabecular bone microstructures: global and local morphometry. *Osteoporosis International: a Journal Established as Result of Cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA* 17(4), 616–626 (2006)
- du Jardin, P., Ponsaillé, J., Alunni-Perret, V., Quatrehomme, G.: A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population. *Forensic Science International* 192(1-3), 127.e1–127.e6 (2009)
- Castillo, R.F., Ruiz, M.D.C.L.: Assessment of age and sex by means of DXA bone densitometry: application in forensic anthropology. *Forensic Science International* 209(1-3), 53–58 (2011)
- Wada, K., Ohtaishi, N., Hachiya, N.: Determination of Age in the Japanese Monkey from Growth Layers in the Dental Cementum 19, 775–784 (1978)
- Kranioti, E., Paine, R.: Forensic anthropology in Europe: an assessment of current status and application. *J. Anthropol. Sci.* 89, 71–92 (2011)
- Cattaneo, C.: Forensic anthropology in the new millennium 165, 185–193 (2007)
- Grabherr, S., Cooper, C., Ulrich-Bochsler, S., Uldin, T., Ross, S., Oesterhelweg, L., Thali, M.J.: Estimation of sex and age of “virtual skeletons”—a feasibility study. *European Radiology* 19(2), 419–429 (2009)
- Zeybek, G., Ergur, I., Demiroglu, Z.: Stature and gender estimation using foot measurements. *Forensic Science International* 181(1), 54-e1 (2008)

- Bruning-Fann, C.S., Schmitt, S.M., Fitzgerald, S.D., Fierke, J.S., Friedrich, P.D., Kaneene, J.B., Muzo, D.P.: Bovine tuberculosis in free-ranging carnivores from Michigan. *Journal of Wildlife Diseases* 37(1), 58–64 (2001)
- Sakuma, A., Ishii, M., Yamamoto, S., Shimofusa, R., Kobayashi, K., Motani, H., Iwase, H.: Application of Postmortem 3D-CT Facial Reconstruction for Personal Identification\*. *Journal of Forensic Sciences* 55(6), 1624–1629 (2010)
- Tocheri, M.W., Molto, J.E.: Aging fetal and juvenile skeletons from Roman Period Egypt using basiocciput osteometrics. *International Journal of Osteoarchaeology* 12(5), 356–363 (2002)
- Craig, J.G., Cody, D.D., Van Holsbeeck, M.: The distal femoral and proximal tibial growth plates: MR imaging, three-dimensional modeling and estimation of area and volume. *Skeletal Radiology* 33(6), 337–344 (2004)
- Macdonald, H.M., Nishiyama, K.K., Kang, J., Hanley, D.A., Boyd, S.K.: Age-related patterns of trabecular and cortical bone loss differ between sexes and skeletal sites: A population-based HR-pQCT study. *Journal of Bone and Mineral Research* 26(1), 50–62 (2011)
- Rösing, F.W., Graw, M., Marré, B., Ritz-Timme, S., Rothschild, M.A., Rötzscher, K., Gericke, G.: Recommendations for the forensic diagnosis of sex and age from skeletons. *HOMO-Journal of Comparative Human Biology* 58(1), 75–89 (2007)
- Tassani, S., Matsopoulos, G.K., Baruffaldi, F.: 3D identification of trabecular bone fracture zone using an automatic image registration scheme: A validation study. *Journal of Biomechanics* 45(11), 2035–2040 (2012)
- Quimby, D.C., Gaab, J.E.: Mandibular dentition as an age indicator in Rocky Mountain elk. *The Journal of Wildlife Management* 21(4), 435–451 (1957)
- Pietka, E., McNitt-Gray, M.F., Kuo, M.L., Huang, H.K.: Computer-assisted phalangeal analysis in skeletal age assessment. *IEEE Transactions on Medical Imaging* 10(4), 616–620 (1991)
- Poole, W.E., Carpenter, S.M., Wood, J.T.: Growth of grey kangaroos and the reliability of age determination from body measurements. I. The eastern grey kangaroo, *Macropus giganteus*. *Australian Wildlife Research* 9, 9–20 (1982)
- Kirkpatrick, T.H.: Studies of the Macropodidae in Queensland. 2. Age estimation in the grey kangaroo, the red kangaroo, the eastern wallaroo and the red-necked wallaby, with notes on dental abnormalities. *Queensland Journal of Agricultural and Animal Sciences* 22, 301–317 (1965)
- Kirkpatrick, T.H.: Studies of the Macropodidae in Queensland. 8. Age estimation in the red kangaroo (*Megaleia rufa* (Desmarest)). *Queensland Journal of Agricultural and Animal Sciences* 27, 461–462 (1970)
- Evans, K.E., Harris, S.: Adolescence in male African elephants, *Loxodonta africana*, and the importance of sociality. *Animal Behaviour* 76(3), 779–787 (2008)
- Castanet, J., Francillon-Vieillot, H., Meunier, F.J., De Ricqlès, A.: Bone and Individual Aging. *Bone: Bone Growth-B* 7, 245 (1993)
- Steele, T.E., Weaver, T.D.: Refining the Quadratic Crown Height Method of age estimation: do elk teeth wear quadratically with age? *Journal of Archaeological Science* 39(7), 2329–2334 (2012)
- Hamlin, K.L., Pac, D.F., Sime, C.A., DeSimone, R.M., Dusek, G.L.: Evaluating the accuracy of ages obtained by two methods for Montana ungulates. *Journal of Wildlife Management* 64, 441e449 (2000)
- Ding, M., Odgaard, A., Hvid, I.: Accuracy of cancellous bone volume fraction measured by micro-CT scanning. *Journal of Biomechanics* 32(3), 323–326 (1999)

Thomas, C.D.L., Stein, M.S., Feik, S.A., Wark, J.D., Clement, J.G.: Determination of age at death using combined morphology and histology of the femur. *Journal of Anatomy* 196(3), 463–471 (2000)

Corsini, M.-M., Schmitt, A., Bruzek, J.: Aging process variability on the human skeleton: artificial network as an appropriate tool for age at death assessment. *Forensic Science International* 148(2-3), 163–167 (2005)

## **Key Terms and Definitions**

Positive identification- Matching identification can be done through dental records, finger print identification, DNA analysis, and through medical implants.

Forensic Anthropology- the application of the science of anthropology in a legal setting most often physical anthropology and human biology are used in criminal cases where the victim's remains are in the advanced stages of decomposition.



# Integrating Computational Methods for Forensic Identification of Drugs by TLC, GC and UV Techniques

Francisco José Silva Mata, Dania Porro Muñoz, Diana Porro Muñoz,  
Noslen Hernández, Isneri Talavera Bustamante,  
Yoanna Martínez-Díaz, and Lázaro Bustio Martínez

Advanced Technologies Application Center (CENATAV),  
7ma A #21406 e/ 214 y 216, Rpto. Siboney, Playa. C.P. 12200. La Habana, Cuba  
{fjsilva, dpmunoz, dporro, nhernadez,  
italavera, ymartinez}@cenatav.co.cu

**Abstract.** The combination of computational methods and the techniques of Thin Layer Chromatography (TLC), Ultraviolet (UV) and Gas Chromatography (GC), brings significant improvements in the speed and accuracy of the analysis and identification of drugs of abuse. In the case of the TLC technique, the processing is fully automatic, through a sequence of algorithms that are run on the images of the resulting plates. By means of this process, the spots corresponding to each substance are characterized with the determination of the respective value of the retardation factor ( $R_f$ ), and the descriptions of their shape and color. The identification of a sample is made by calculating its dissimilarity with respect to the previously stored patterns. During this process, the quality of the plate is also evaluated. For the analysis of Ultraviolet data, the computation of dissimilarities is performed by taking into account the shape of the spectra. The analyzed sample will take the class of the closest pattern. Spectra are previously standardized in order to eliminate the variation of the slope of the curves caused by the dispersion and the variation in the particle size. With this purpose, the standard normal variate pre-processing method is applied. When using the GC technique, the identification process is based on the detection of peaks that belong to each drug according to their retention times. The drugs are identified by comparing both, the absolute and the relative retention times (with respect to two internal standards) of the detected peaks, with those of the patterns stored in the system.

**Keywords:** Drug identification, Thin Layer Chromatography, Ultraviolet, Gas Chromatography.

## 1 Introduction

The identification of drugs is usually performed by applying different analytical techniques, where independently, each of them provides a criterion in the definitive identification of drugs. The UV spectrum analysis, for example, is rarely used as the unique identification technique of drugs, but rather as a confirmation of the results of other techniques [15]. At the meantime, the TLC method must be combined with at least one spectroscopic method [21]. For instance in [13] both techniques TLC and GC were combined to make an exhaustive drug screening in urine. The TLC offers the advantage of the chromatographic separation techniques and in addition it is the most economic

for obtaining low costs when the number of samples increases [15], but its result is almost never enough to decide the identity of a drug as we explained above. The three mentioned analytical techniques can be used separately in order to discover the presence of drugs and their results can be integrated to achieve a more secure identification of the analysis.

For the interpretation of the analytical techniques results, the intervention of the analysts is always required, who must measure and evaluate manually those results with different instruments, in different moments, and compare them with appropriated patterns. Human errors can occur in chemical labs during these activities and cause false positives and false negatives for numerous reasons. In TLC specifically, the measurements of the analysts include the determination of retardation factors by measuring the distance traveled by the substance spot from the baseline and the assessment of their shapes and colors. In UV technique, the shapes of the spectra of the substances must be compared visually with the spectra patterns. In GC analysis, the peaks of the chromatogram must be scanned and then should be visually compared their retention times with the previously known.

Automating the process of measurement and calculation, usually performed manually by the analysts, can contribute significantly to the elimination of some of the mentioned errors, and speed up its execution. The introduction of computer vision techniques allows us to work over the digital images of the TLC plates, rather than work directly over them. Similarly, the output digital data of the GC and UV equipments can be subjected to analysis by using signal processing methods. The application of computational methods, particularly pattern recognition techniques can provide a significant increase in the efficiency and efficacy of these analytical processes, which is one of the goals of the forensic analysis of drugs. Integrating own computational methods for interpreting the results of different techniques in a single system will provide to the analysts the possibility to work with new tools that would help significantly in decisions, regarding the identity of substances, particularly in the case of drugs of abuse.

In TLC, the proposed processing is done automatically through a sequence of image processing algorithms that are applied over the captured images of the plates. This process basically consists of obtaining the characteristics of each spot through the determination of its  $R_f$  value, its shape, and its color, that correspond to each separated substance. The identification of an unknown sample is performed by calculating a dissimilarity value with respect to previously stored patterns. This dissimilarity value combines properly the differences between the sample and the pattern with respect to color, shape and  $R_f$  value with a proper weight respectively.

The automatic identification of drugs by using UV data is performed by the comparison of the shape of the spectrum from the analyzed sample with the shape of the spectral patterns previously stored. In order to standardize the spectra before to be compared, the standard normal variate (SNV) method [3, 5] must be applied. The drug name (class of the sample) is assigned according to the minimum value of dissimilarity obtained.

The computational analysis of the chromatograms is based on the automatic detection of peaks that belong to each drug, according to their retention time. In order to discriminate the smaller peaks that are normally considered as noise, a threshold must be applied. The introduction of two internal standards whose peaks are perfectly visible

and whose retention times do not match with none of the possible drugs to identify, allow to determine relative retention times instead of absolute.

The chapter is organized as follows. In section 2 are described in detailed the materials and methods used in each of the three techniques. In sections 3, 4 and 5 are explained the computational processing of the TLC, UV and GC techniques, respectively. Experiments and results analysis are reported in section 6. Conclusion and future research directions are given in section 7.

## 2 Materials and Methods

### 2.1 Thin Layer Chromatography (TLC)

**Supplies and equipment:** All separations and reactions were carried out on aluminum-layer silica-gel plates 60 F254, thickness 0,25 mm (Merck KGaA), activated during 5 min to 100<sup>0</sup>C. The plates were developed in a Twin Trough Chamber for 20x20 cm plates with glass lid (no.022.5255 CAMAG Scientific, Wilmington, NC).

The spray bottles, paper saturation pads, dipping containers, and other glassware included pipettes, capillary pipettes for sample application, conical plastic disposable evaporation cups were supplied by Alltech.PA. The UV viewing chamber (Spectroline CM-10) with 254-365nm and the lamps for UV revision were supplied by Alltech. The digital photographic camera Panasonic Lumix DMC-FZ7 was used to capture the images of the TLC plates.

**Reagents and substances for experimental:** For chromatographic runs were used four solvent systems as mobile phases *a*) Methanol: strong ammonia solution (99:1), *b*) Methanol-ethyl acetate: strong ammonia solution ((85:10:5), *c*) Chloroform: Acetone (40:10), *d*) Chloroform-n-butanol-ammonia (35:20:15). Forty-three drug patterns were used for the training set, supplied by the Central Criminalistic Laboratory divided into three groups:

1. Basic substances: Amitriptyline, Atropine, Carbamazepine, Chlordiazepoxide, Clonazepam, Chlorpromazine, Cocaine, Codeine, Desipramine, Dextropropoxifen, Diazepam, Dibucaine, Ephedrine, Fluphenazine, Homatropine, Imipramine, Ketamine, Levomeprazine, Lidocaine, Mephenesin, Meprobamate, Methylphenidate, Morphine, Nitrazepam, Oxazepam, Papaverine, Procaine, Strychnine, Scopolamine, Trihexyphenidyl, Temazepam, Tetracaine, Thioridazine, Tramadol, Trifluoperazine.
2. Amphetamine Derivatives: Metamfetamine MBDB, MDA, MDEA, MDMA.
3. Barbiturics: Amobarbital, Barbitol, Phenobarbital and Secobarbital.

The reagents used are from MERCK and SIGMA-ALDRICH companies. The visualization reagents used are: Dragendorff reagent, diethylamine, diphenylcarbazone, fluo-rescamine, furfural, iodoplatinate spray reagent, Mercury(I) sulfate, Ninhydrin spray reagent and vanillin.

The Chlorpromazine was selected as internal standard. The TLC detection/ visualization reagents were prepared using the formulations given in Clarke's guidebook [15] : *a*) Dragendorff, *b*) Iodoplatinate spray, *c*) Fluorescamine, *d*) Ninhydrin spray,

*e*) Diphenylcarbazone, *f*) Mercury, *g*) sulfate, *h*) 254 nm (UV) light projected onto a developed TLC plate (must be fluorescent indicator plate). The reagents used were supplied from MERCK and SIGMA-ALDRICH companies.

**Methods:** Sample pattern solutions were prepared obeying the following sequence: For solid samples: a) Triturating and homogenizing the samples if necessary, weighing 10 mg of the powder and diluting in 10 mL of methanol, to help the dilution applying ultrasonic bathroom for 15 minutes, filtering if necessary. For liquid samples: Filtering if necessary.

Aliquots of 30  $\mu\text{L}$  of pattern solutions (1mg/ml in methanol) were applied on the plate base line, leaving a space of 1cm between substances and 2cm from the edges of the TLC plates. The selected internal standard (chlorpromazine) was applied always on the lane 2 of each plate.

The plate was prepared according to methodology referenced in [22], where is textually mentioned that “a paper pad was placed on one side of a two-sided TLC chamber, and 6 mL of the selected developing solvent system was run (pipetted) down the pad, the plate was placed on the dry side of the chamber for 10 min before developing to full plate height by the addition of 6 to 8 mL of solvent. Chromatography per se took 10.5 min. and ran to 10cm”. The solvent systems a) and b), mentioned above, were used as mobile phases for the substances of the groups 1 and 2, and the solvent systems c) and d) for the substances of the group 3.

The plates were dried during 5 min. at 60 °C and observed in the viewing UV light chamber at 254nm and made note of any fluorescent spots. After that, the plates were treated with the detection/visualization reagents according to the type of substances. The plates of Group1 (basic substances) with Dragendorff reagent or Iodoplatinate spray reagent. The sequence of reactions for the Group 2 (amphetamine derivatives) was as follows: first Fluorescamine reagent, next Ninhydrin spray reagent and finally Dragendorffs reagent. For Group 3 (barbiturates) the sequence was first Diphenylcarbazone and finally Mercury (I) sulfate. The TLC process was repeated 3 times for each substance.

Once the spots appeared, the plates were photographed under a fluorescent lamp of white light using as background a paper of black color and with a no reflecting surface, avoiding involuntary shadows during the registration of the image and without using the flash. The commercial digital camera Panasonic Lumix DMC-FZ7, at a distance of 30 cm attached to a tripod was used for the capture of the images. The images were registered in color JPEG format 2048x1360, with resolution of  $x=72$  and  $y=72$  ppi, in automatic mode without flash.

## 2.2 Ultraviolet-Visible Spectroscopy (UV-VIS)

**Supplies and equipment:** Two UV-VIS spectrophotometers were used, Jasco model V-530 and GBC model CINTRA101 in order to take in account the reproducibility of the computational identification results using different spectrometers. The chemical glassware including Quartz cuvettes of 3 mL for the spectrometers was supplied by Alltech.

**Reagents and substances for experimental:** The forty-three drug patterns used for the TLC training set, supplied by the Central Criminalistic Laboratory, divided into three

groups were used too for the UV-VIS analysis. Methanol PA, HPLC degree and Sulfuric Acid (0.1N) were used in samples preparation. The reagents were supplied from MERCK and SIGMA-ALDRICH companies.

**Methods:** The sample pattern solutions (1mg/mL in methanol) were prepared obeying the same sequence used for TLC analysis. For the UV-VIS analysis was added to each cuvette 2mL of Sulfuric acid 0.1N completing their contents with 10 $\mu$ L of the pattern solution. The spectrum measurement parameters for the two spectrophotometer are shown in table 1.

**Table 1.** Spectrum measurement parameters

Parameters	Jasco spectrometer	Cintra spectrometer
Mode	Absorbance	Absorbance
Response	Fast	–
Scanning speed	400 nm/min	450 nm/min
Upper wave length	350 nm	350 nm
Lower wave length	200 nm	200 nm
Data pitch / step size	0.2 nm	0.2 nm

The analysis process was repeated three times for each substance. The resulting data were transformed to the .txt format for the JASCO V-530 Model and to .csv format for the GBC Cintra 101.

### 2.3 Gas Chromatography (GC)

**Supplies and equipment:** A Gas Chromatograph AGILENT 7890A with flame ionization detector (GC-FID), column “AGILENT, DB-ALC2” and Nitrogen as carrier gas was used for the analysis.

**Reagents and substances for experimental:** The same forty-three drug patterns used for the TLC and UV-VIS training set, supplied by the Central Criminalistic Laboratory, divided into three groups were used for the GC analysis. Methanol PA, HPLC degree was used in samples preparation. Methyl dodecanoate and Squalene (20mg/mL in methanol) were used as internal standards. The reagents were supplied from MERCK and SIGMA-ALDRICH companies.

**Method:** Setup the chromatograph work parameters as follows:

- \* Fix initial oven temperature at 100<sup>0</sup>C with isotherm during 1 minute, then increasing the temperature at the rate of 15<sup>0</sup>C/min to 280<sup>0</sup>C and maintaining for 7 minutes.
- \* Fix injector and detector temperature at 250<sup>0</sup>C and 290<sup>0</sup>C, respectively.
- \* Column flow of 1, 3 mL/min and splitless injection mode.
- \* Column HP-5 350<sup>0</sup>C: 30 m x 320  $\mu$ m x 0.25  $\mu$ m.

Sample pattern solutions (1mg/mL in methanol) were prepared obeying the same sequence used for TLC analysis. For the GC analysis 30 $\mu$ L of each internal standard was added to 300 $\mu$ L of the samples pattern solution, then homogenizing and injecting 1L of the final solution in the column. The process was repeated three times.

### 3 Computational Methods of Analysis for TLC, UV and GC

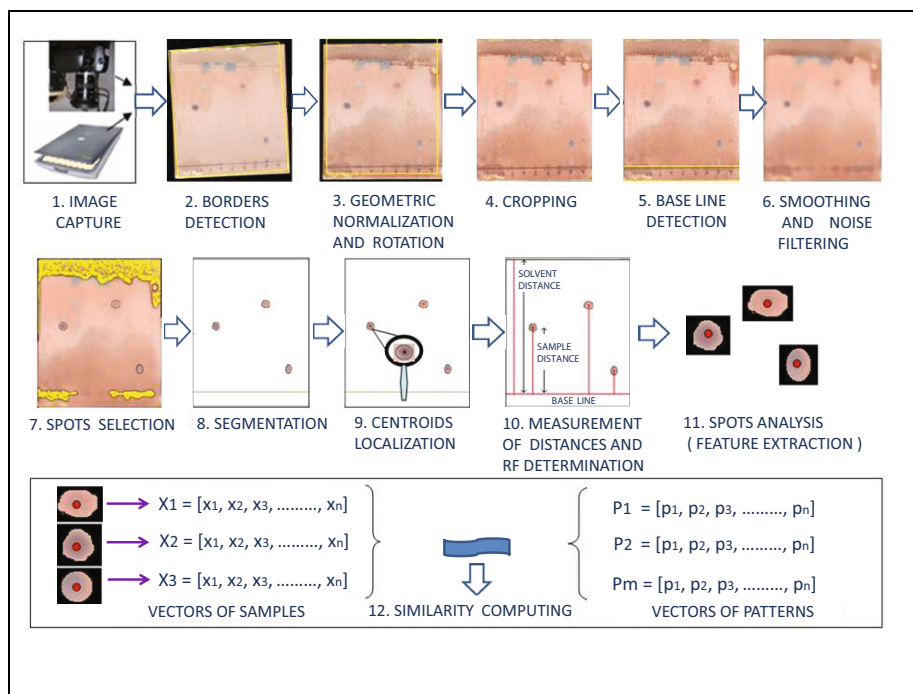
#### 3.1 Analysis of the TLC Plate Images

The most important parameter for substance identification by TLC is the  $R_f$  value. However, the  $R_f$  determination varies considerably due to the variation of experimental conditions. Often, the  $R_f$  values given as reference in specialized manuals, cannot be used directly as standards of comparison, because the real difficulty for their repeatability caused by the ambient and experimental conditions. Although the  $R_f$  value can offer definitive information for the identification of diverse substances, frequently the analysts perform the manual identification of the substance of interest on the basis of the shape and the color of each spot, besides the  $R_f$  value [8]. Nevertheless, some factors can influence changing typical  $R_f$  values, shape and/or color resulting of the TLC process such as: the absorption of moisture in the blank plates, the particles size, temperature, impregnation, layer thickness, the solvent parameters, pH., volume and height in the chamber, time gradients and the spotting techniques [21]. In addition to these problems, those typical errors of any image processing must be avoided in this case where the analysis of the TLC plate is made digitally. Such errors are those produced by the illumination, the optical aberrations of the image because of the lenses and the variations in the geometry of the forms by parallelism errors. Other causes that affect the repeatability of these processes are the imprecision in the process of spotting which provokes a variability on the size of the migrant spots, the non-uniformity of the atomization process and the plate's edge effect that can alter the direction of the rising spots and even its own shape.

The technique of TLC has analytical tasks perfectly separated such as: preparation, application, development and evaluation [21]. The automatic processing of the plates must be invariant to most of the problems. A common strategy to deal with these difficulties and to obtain patterns that allow automatic comparison, is the inclusion of one internal standard substance in the TLC plate. This internal standard is subjected to the same experimental conditions and variations as the samples of the substances submitted to the analysis. Thus, the measured  $R_f$  value is divided by the internal standard's  $R_f$  value trying to compensate the variations in experimental conditions. The chlorpromazine was selected like internal standard, taking into account that the spots corresponding to this drug were always high contrasted against the background, its size was always an average size clearly visible, with a highly stable shape and color, and note that the absolute  $R_f$  value corresponding to the spots of this drug does not match with other possible drugs.

The main objective of this module is to create a computational process capable of identifying the drugs screened using TLC with a high degree of confidence, even with the disadvantages mentioned above and with a high speed of the response.

A general description of how to perform drugs identification by TLC automatic process can be seen in the scheme presented in Fig.1. This scheme has been divided in different blocks for its best understanding. It shows the complete procedure with each image of TLC and the detailed description of each one of the blocks.



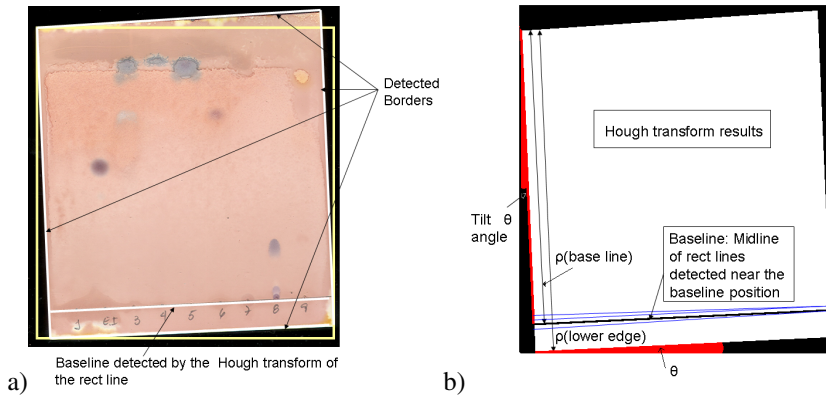
**Fig. 1.** Summary diagram of the analysis of the TLC images

**Image Capture.** The image capture can be done using a digital camera or scanner, but the experiments were performed with a Panasonic DMZ-FZ7. The requirements for a good capture of the image are the following: it is recommended that the background where the plate is placed to get the image should be of a uniform dark color. The black color is ideal since none plate is black. This prevents some errors in detecting the plate edges automatically. The camera must be set in a stable place at a safe distance from the plate, to avoid mistakes of parallel and radial distortion. One element to take into account when using a camera, is not to use the flash light when the surfaces of the plates are wet. This cause sometimes some reflections of light creating light saturated image areas. The light must be uniform and should ensure the necessary visibility of the results. The camera or any capture device must be calibrated to ensure the good behavior with the colors. For the calibration was used the methodology documented in the site [1] and the OpenCv calibration method for the camera calibration [4].

The minimal resolution of capture must guarantee approximately a height of 600 pixels that correspond to the height of 11 cm of the plate, that means 55 pixels/cm approximately like the minimal value.

**Borders Detection.** The process of automatically detecting the actual edges of the plate of TLC, is based on combining the Canny filter [7] and the Hough transform of the line [25]. Before applying the Canny filter, the image is smoothed by applying a Gaussian

filter to avoid the detection of false edges. By this way, the straight lines constituting the edges of the plate are detected and the tilt angle  $\Theta$  thereof is obtained directly from the output of the Hough transform itself. This angle  $\Theta$  is used later to rotate and align the image automatically. By checking the value of the tilt angle  $\Theta$  and radio  $\rho$  obtained in each case, it is possible to determine where to cut automatically the image too (See Fig. 2a and 2b). Using the proposed method it was properly detected the 100 % of the borders of the processed TLC plates.



**Fig. 2.** Detection of the borders and the baseline by Hough transform and the tilt angle needed for the automatic rotation of the image

**Geometric Normalization and Rotation.** The angle  $\Theta$  obtained from the Hough transform in the previous step is used to rotate the image automatically. In order to obtain a single height in pixels each image should be resized using a cubic interpolation method or a sampling in order to obtain a geometrically normalized image with a height value of 600 pixels. Thus, the width that varies according to the number of lanes is also resized proportionally for maintaining the aspect ratio of the original image.

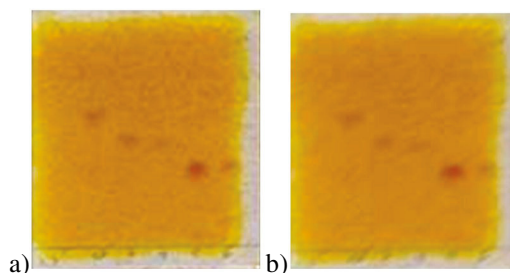
**Cropping.** According to the previous steps, from the original image just the interest region of the image (ROI) should be preserved. Sometimes there remain some thin rests of black background at the edges of the image, that can be considered as noise in it. These are eliminated through a classical process of thresholding and cropping [7] to avoid posterior errors.

**Baseline Detection.** The zone of the plate where the samples or patterns are spotted for a further analysis is called baseline. This line is drawn by a pencil and is located at 1 cm from the bottom of the silica plate. This baseline is detected in the plate, by applying the Hough transform of the line. Commonly, as a result of this method are obtained more than a straight line, according to the discrete nature of the image. One alternative



to compensate this variation is to take the midline that can be obtained directly from the values of the radio ( $\rho$ ) from the Hough transform with origin coordinates (0,0). The baseline value must be validated with enough accuracy, to avoid any error in distance measurements, which is necessary for calculating retardation factors (Rf) (See Fig. 2a and Fig. 2b).

**Smoothing and Noise Filtering.** Although this process has been placed on the diagram of Fig. 1 with the number 6, the filtering process can be part of other steps of the complete processing sequence. The Gaussian smoothing [7] is also applied to attenuate high frequency noise. An example of the result of this filtering process can be seen in Fig. 3. As a result of this filtering, some of the isolated small artifacts are removed. Principally those noises of a small size are reduced by this filtering action.



**Fig. 3.** Results of the application of the gaussian smoothing filter where a) is the original image and b) is the filtered image

**Segmentation.** The spots segmentation is done primarily through the following steps:

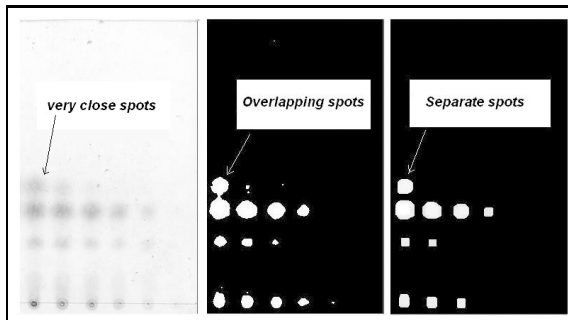
1. Converting the color image to a grey scales image.
2. Determining the threshold for the image binarization.
3. Detecting the edges of each spot.
4. Marking the edges and storing the sequence of their respective pixels.

The problem of obtaining the best threshold has been addressed by different methods discussed in the literature [9], but often it depends on the context of or user interests. The Otsu's method [16] is commonly used to automatically perform the thresholding of the image based on the histogram shape. A new fast variant of the Otsu algorithm was used to do the segmentation [26]. This method has allowed the successful segmentation of the 95% of the spots tested, in the case of images with good quality.

The remaining 5 % of the spots, which were not detected because of its low contrast were segmented in a second pass of the algorithm dedicated to detect and to segment these spots considered weak. For this stage a pyramid segmentation method was applied [14]. The parameters of this method were estimated according to the proposed in the paper by Kosir [11].

In the case of mixtures, the substances were segmented without any difficulty, when there was not connectivity between the pixels of the of neighboring spots of the different substances.

When there are overlapping spots, an optional step of morphological analysis is applied. The detection of this type of overlapped spots is not a simple task, but this situation is solved when the two nearly spots are connected by a narrow connection like is shown in Fig. 4.



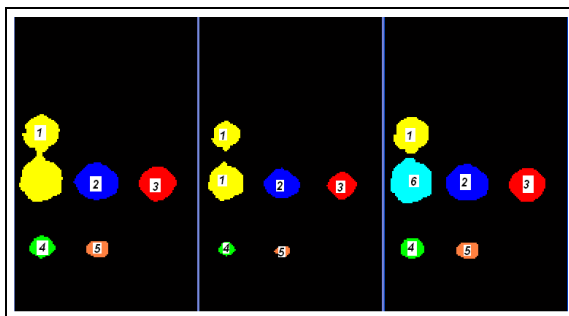
**Fig. 4.** Results of erosion-dilation to eliminate overlapping

The complete procedure to solve this situation is as follows:

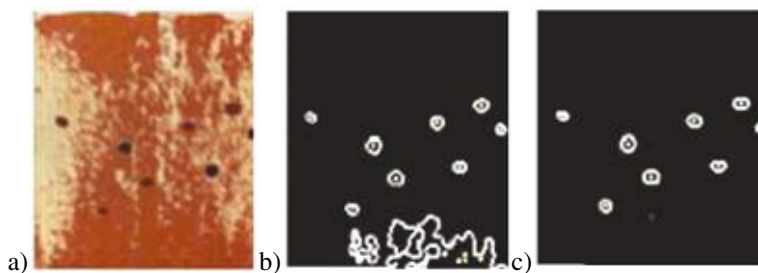
1. Load the TLC image.
2. Find and label all the spot regions.
3. Erode the TLC with the goal of removing the narrow connections between near spots.
4. After the operation of erosion must be checked in the image whether there are two regions with the same label, which is a sign that a spot was splitted in two spots, which occurs commonly when the union between two spots is narrow.
5. A new label is assigned to one of these spots, which can be done even if more than one spots appear with the same label.
6. Then is executed an operation of dilation, so the spots are kept separate and with new labels. Further, this opening operation (defined as an erosion followed by a dilation), with a proper structuring element in order to separate near regions is made only if it is needed.

In Fig. 5 are shown the steps of the solution for overlapping the situation in the lane 1 of the TLC plate.

**Spots Selection.** A selection process is performed to remove some spots with characteristics of size, position and shape that do not fulfill certain requirements. This process is performed by successive application of morphological operations (erosion- dilation) [7] and a heuristic based on the above mentioned characteristics. In order to select the



**Fig. 5.** Steps to detect and eliminate the overlapping. The numbers are the labels, the regions were colored to understand better the process.



**Fig. 6.** Results of the application of the combined process of b) thresholding and erosion-dilation and c) spot selection process over a) noisy image

spots by their size, two thresholds are applied, one upper for the too large spots and one lower for the too small spots, determined empirically. In Fig. 6 is shown an example of a noisy image with false spots, the result of the removing action and the image resulting.

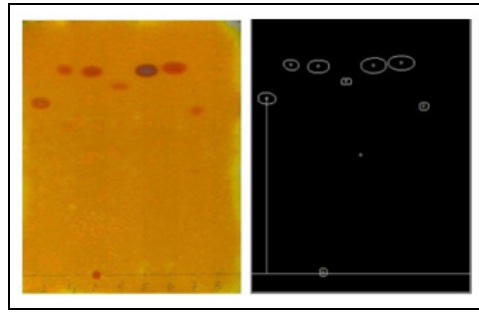
**Localization of the Centroids.** The centroid of each spot is calculated according to Eq. 1 [7].

$$m_{pq} = \sum_x \sum_y x^p y^q f(x, y) \tag{1}$$

where the coordinates of the centroid are:

$$(m_{10}/m_{00}, m_{01}/m_{00}) \tag{2}$$

The result of centroid localization process is shown in Fig.7



**Fig. 7.** Results of the localization process of the centroids. In the first lane it is shown the correspondent distance to measure the absolute Rf value of the sample and in the second lane the correspondent to the pattern (internal standard).

**Measurements of Distances and Determination of the Rf.** The fact of performing this step fully automatic avoids possible errors that can be committed when the Rf measurements are done manually. The Rf value measured for each substance is divided between the Rf value of the substance used as internal standard (Chlorpromazine). This final Rf is known as Rx (relative Rf) [20]. In Fig. 7 is shown an example of a plate's image with the respective segmentation and the determination of the Rf.

**Spots Analysis (Feature Extraction).** Once the spots have been segmented, the next step is the extraction of the corresponding shape descriptors: dimensions, area, perimeter, perimeter-area ratio, complementary area, principal angle of inclination, Hu moments and Legendre moments. The vector is formed with the information that describes the shape of each spot. Particularly the Legendre moments [6] have been incorporated as shape descriptors [7] due to the success of its use in our own laboratory. The values of these features are normalized to avoid scale errors. The color of the spots on thin layer chromatography provides valuable information for the final identification of the analysis. The color histograms of each spot (R (red), G (green) and B (blue)) are also extracted. Each color band is processed to compute its corresponding minimum and maximum values and the value of each pixel is adjusted by Equations 3, 4 and 5; with the purpose of reduce the effect of the lighting variations. Finally, the information describing shape and color is associated with each spot and stored in a file. This information will be used in the following identification step. Table 2 shows the vector of information extracted from each substance spot.

$$R = (R - Rmin)/(Rmax - Rmin) * 255.0 \quad (3)$$

$$G = ((G - Gmin)/(Gmax - Gmin)) * 255.0 \quad (4)$$

$$B = ((B - Bmin)/(Bmax - Bmin)) * 255.0 \quad (5)$$

**Table 2.** Spot descriptors

Component	Name	Description
1	Maximal Height	Value of the maximal height of the spot
2	Maximal width	Value of the maximal height of the spot
3	Area	Number of pixels of the spot
4	Perimeter	Number of pixels of the contour of the spot
5	Area-perimeter relation	Quotient between area and perimeter
6	Complementary area	Area of the bounding box - area of the spot
7	$\Theta$	Tilt angle of the principal axis of the spot
8	Bounding box area	Number of pixels of the bounding box
9-15	Hu1 - Hu7	Hu $\acute{e}$ seven moments invariants
16-77	Legendre Moments	Legendre Moments
78-85	Histogram R	Histogram of 8 bins of the red component of the spot
86-93	Histogram G	Histogram of 8 bins of the green component of the spot
94-101	Histogram B	Histogram of 8 bins of the blue component of the spot
102-103	Centroid	Coordinates of the spot centroid

**Similarity Computing.** The identification of the chemical compound associated with each spot is done by comparing the Rf values, the shape descriptors and the color descriptors with respect to the values of pattern substances previously stored. It is difficult to know which is an exact quantification of the contribution of each of these three entities of information respect to the decision about the identity of a drug. In the literature recognized Rf values are registered for each substance in each of the systems. However, significant variations may occur in experimental results as it was mentioned before. Although importance of the Rf value is crucial for the identification of the substance, also the shape and color of the spots are commonly taken into account by the analysts. In our proposal, the similarity calculation between the values of the vectors that describe the analyzed substances and the pattern substance is performed by Eq. 6.

$$\hat{p} = \underset{p \in (1, \dots, n), n \in N}{\operatorname{argmin}} \quad w_{1(j)} \left( 1 - \frac{\min(Rx_s, Rx_p)}{\max(Rx_s, Rx_p)} \right) + w_{2(j)} D(f_s, f_p) + \dots \quad (6)$$

$$w_{3(j)} D(l_s, l_p) + w_{4(j)} D(h_s, h_p)$$

- $p, j, s$ : Index of the pattern substance, index of the system visualization reagents, index of the spot respectively
- $Rx$ : Quotient between the Rf value of the analyzed spot and the spot respective to the internal standard on the same plate
- $w_1, w_2, w_3, w_4$ : Learned weights, for each solvent system and visualization reagent,
- $f$ : Vector of shape descriptors,
- $l$ : Vector of shape descriptors based on Legendre polynomials,

- $h$ : Vector of color histograms,
- $D$ : dissimilarity between pattern and analyzed spot vector (Euclidean distance for shape descriptors and distance of Bhattacharyya for the histograms). See Eq. 7 [4].

$$D(H_s, H_p) = \sqrt{1 - \sum_i \frac{\sqrt{H_s(i) * H_p(i)}}{\sqrt{\sum_I H_s(i) * \sum_I H_p(i)}}} \quad (7)$$

### 3.2 Automatic Evaluation of the Quality According to the Homogeneity of the Plates

To evaluate the quality of the plates related with its homogeneity, a process based on a statistical analysis of its histogram has been performed. These quality measurements have the purpose of registering basically the presence of certain deficiencies produced by chemical or physical process, those characteristic of capturing images under poor lighting conditions, mainly by the light saturation in some regions, specular behavior of the surface of the plate, or shadows produced by external objects. The Kurtosis and Skewness values [23] are obtained from the histogram of grays scales by Eq. 8 and Eq. 9. These statistics allow estimating an index of the images quality through a learning process. The resultant value is expressed in percent and the process was cross-validated. The following equations allow the calculating of the Kurtosis (Eq. 8) and the Skewness (Eq. 9).

$$kurtosis = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N - 1)s^4} \quad (8)$$

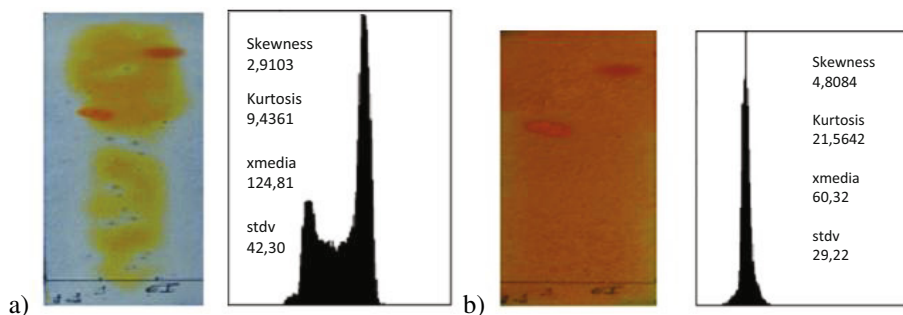
$$skewness = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3}{(N - 1)s^3} \quad (9)$$

where  $\bar{Y}$  is the mean,  $s$  is the standard deviation, and  $N$  is the number of data points. Three quality categories were defined to describe verbally the result of the quality evaluation: bad, regular and good. These quality categories were decided after a statistical study of the behavior of the results obtained in different experiments, using different kind of images. Fig. 8 shows two examples of TLC plate images and their corresponding histograms.

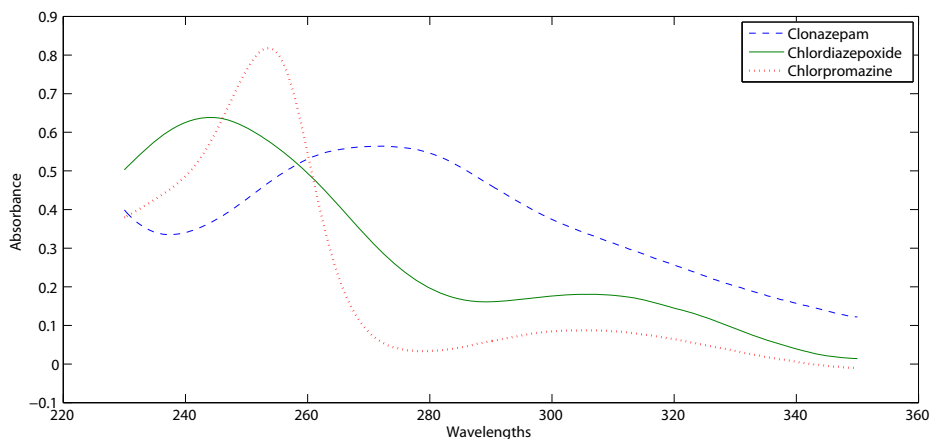
In Fig. 8a it can be seen a noise presence which affecting the image quality. This noise is a result of occurrent deficiencies in the complete TLC process. They are reflected in relatively low values of the Kurtosis and Skewness of the image histogram. Fig. 8b shows the high values of these measurements and the corresponding histogram of a high image quality. It should be noted that both images correspond to the same drug testing case.

## 4 Analysis of Ultraviolet Spectrum

Spectral information (depending on the applied instrumental technique) can be seen as a distinctive signature of substances. In the case of an UV spectrum, for a certain



**Fig. 8.** Examples of the results of quality assessment by the Kurtosis and Skewness of the histograms of the images, where a) was classified as bad and b) as good



**Fig. 9.** Ultraviolet spectra of three different substances

substance, the relation between ultraviolet light absorbance and light wavelength is represented. Therefore, an UV spectrum is often unique for each pure substance.

This property is the reason why specialists have frequently used UVS for comparing/matching substances of abuse, to perform tasks like database searches, detection or identification of an unknown sample, to decide if two materials come from a common source, the structure elucidation and classification of spectra, among others [15, 10, 24].

The criteria for comparing spectra can be divided into direct and indirect. Direct matching uses the spectral data directly, and the indirect matching uses derived information from spectra. The latter relies on identification of selected peaks and the extraction of information from them. Multivariate data analysis techniques and the distance/angle methods are usually direct methods which treat digitized spectra directly without any prior identification of peaks. Therefore the success of a matching criterium rests on its ability to discriminate subtle differences in samples that are inherently similar [12].

In our case, the comparison of spectral data is carried out by using a proximity (similarity/dissimilarity) measure that defines how much (dis)similar are two spectra.

A (dis)similarity can be seen as a function that assigns high (low) values to alike objects and low (high) values to objects that have distinct characteristics. Therefore, a large similarity and a small dissimilarity mean both the same thing with respect to comparison of objects, i.e., similar objects. For substance identification, the goal of a good similarity measure for UV spectrums is to achieve higher values for spectra belonging to the same substance, and lower values for those belonging to different substances.

There are many ways to compare objects, and it eventually depends on what we consider that makes objects (in this case, spectra) similar. This implies that the suitability of a proximity measure depends on the problem at hand; its definition should be based on the knowledge one has about the data. However, defining good measures for substances identification is still a challenge.

In the case of UV spectra, they have a shape as a function of wavelengths. They are often continuous functions (they do not jump) and it is in fact their shape what is unique to a considerable amount of pure substances of abuse (usually have a characteristic peak that identifies them), hence its discriminative power for differentiating among them. However, for different families of chemical compounds, there are certain similarities in the curves of their corresponding UV spectra. Therefore, it is extremely necessary that the (dis)similarity measure to be applied for the comparison (for identification purposes), considers as much discriminative contextual information as possible.

For the above explained, there have been studies of several (dis)similarity measures to compare spectra [24, 17, 10], but a few actually take into account the continuity information and / or shape of this type of data. In this work, we applied the dissimilarity measure called Shape Measure [17]. In [17], authors propose to compute the Manhattan distance on the first Gaussian derivatives of the curves. This way, the sum of absolute differences between derivatives of the curves, allows considering the shape information that can be obtained from their derivatives:

$$d(x_S, x_P) = \sum_{j=1}^m |x_{Sj}^{\sigma} - x_{Pj}^{\sigma}|, \quad x^{\sigma} = \frac{d}{d_j} G(j, \sigma) * x \quad (10)$$

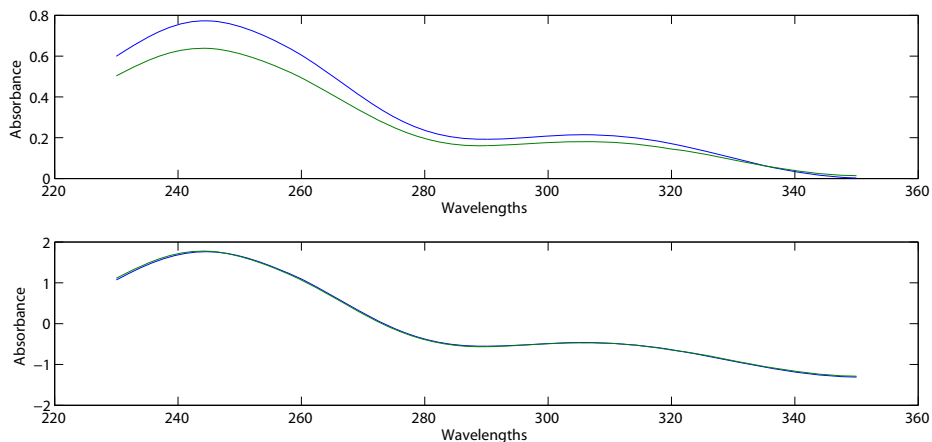
where  $x_S$  is the spectrum of the sample and  $x_P$  is the spectra of the pattern. The expression of  $x^{\sigma}$  corresponds to the computation of the first Gaussian (that is what G stands for) derivatives of the spectra. A smoothing (blurring) is done by a convolution process (\*) with a gaussian filter and  $\sigma$  stands for the smoothing parameter. In ours and other studies it has been proven that this measure is very efficient for comparing chemical spectral data [17, 18, 19].

One factor that may affect the performance of the selected measure is the existence of spectra of the same substance with different concentrations (See Fig. 10(top)). Although this measure is based primarily on the curve shape changes, when the spectra are similar for substances of the same family, the differences in concentration for several samples of the same substance may create confusion between the spectra of different substances, but similar. Hence, it is necessary to normalize the spectra prior to their comparison.

In this case, we suggest to apply the pre-processing method Standard Normal Variate (SNV) [3, 5], such that each spectrum is pre-processed independently. This method is based on transforming the spectrum  $x = (x_1, x_2, \dots, x_n)$  measured at  $n$  wavelengths, into the spectrum  $z = (z_1, z_2, \dots, z_n)$  where  $z_j = \frac{(x_j - m)}{s}$ , with  $m$  and  $s$  the mean and



standard deviation of  $x$ , respectively. In this way, the variation of the slope of the spectra caused by the dispersion and the variation in the particle size (change in concentration) are eliminated (See Fig. 10(bottom)). After applying the pre-processing, pure spectra can be compared by the selected dissimilarity measure.



**Fig. 10.** Ultraviolet spectra of the same substance at different concentrations: (top) before SNV, (bottom) after SNV

## 5 Analysis of the Gas Chromatogram

Gas chromatography (GC) is a common type of chromatography used in analytical chemistry for separating and analyzing compounds. With GC, a compound can be identified from a mixture of chemicals by its retention time, if the method conditions are constant. In our case, GC is used to identify a predefined set of drugs of abuse. The identification algorithm consists of three basic steps: peak detection, internal standard identification and peak identification.

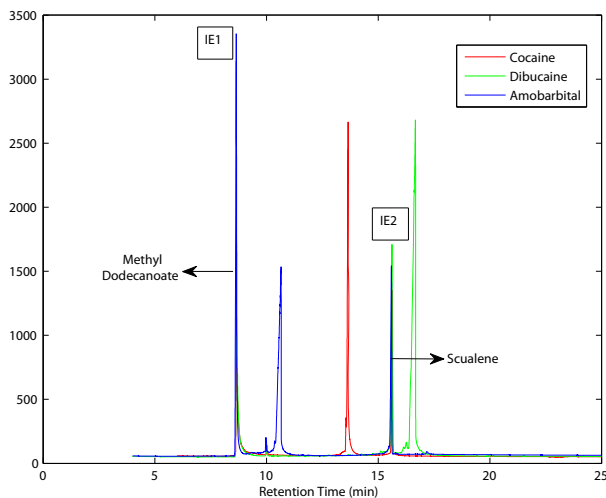
### 5.1 Peak Detection

In this step, peaks in the chromatogram are automatically detected using the algorithm *peakdet* proposed by Billauer [2]. This algorithm implements a simple idea, where the trick is to realize that a peak is the highest point between valleys. What makes a peak is the fact that there are lower points around it. In this way, the algorithm looks for the highest point, around which there are points lower by a “peak threshold” on both sides.

This peak threshold can be tuned by the user in order to determine at what level of resolution the peaks are going to be detected. Therefore, very small peaks which arise due to instrumental noise will not be detected.

## 5.2 Internal Standard Identification

As it was mentioned in Section 2, two internal standards (Methyl Dodecanoate and Squalene) were used with the purpose of working with their relative retention times. The peaks corresponding to these internal standards are visible in the chromatogram and their retention times do not match with any of the compounds of interest. These peaks are identified using their absolute retention time. Fig. 11 shows three different chromatograms and the peaks corresponding to the internal standards.



**Fig. 11.** Chromatograms of three different substances

## 5.3 Peak Identification

Once the peaks are detected and the internal standards are identified, the next step is to identify every detected peak. In this case three criteria are used to identify a peak based on its retention time. The first criterion compares the absolute retention time of the peak with the ones of the patterns stored in the system. The other two criteria compute the relative retention times with respect to the two internal standard and these values are also compared to those of the patterns. Thus, if the chromatogram is displaced and as a consequence the absolute retention time does not match the ones of the patterns, the peaks can be identified by its relative retention times. The comparison of both, absolute and relative retention times are performed by calculating confidence intervals.

Finally, to identify a peak with some pattern, two of the three criteria have to be fulfilled. If a peak does not match any pattern, not any identification is given to it.

## 6 Experimental Results

### 6.1 Experiments with TLC Technique

The experiments performed to validate the results of the automatic analysis of TLC plates were directed to check separately major stages. A first experiment was designed to test the results of the segmentation and selection of the spots of interest. A second experiment was intended to evaluate the automatic quality assessment of the plates and a third experiment was designed for the identification of the substance spots corresponding to the patterns of each drug.

**Segmentation and Selection of the Spots.** In this experiment we evaluate the result of the segmentation and selection of the substance spots in plates with three different levels of quality: Good, Regular and Bad. These qualifiers were given by the specialists according to the noise, homogeneity and contrast between spots and the background. The performance of the algorithm was visually evaluated by the specialists by checking the resulting segmented spots signaled by their enhanced edges. Table 3 shows the results of this experiment in terms of True Positive percentage (TP) and True Negative percentage (TN). TP measures the spots correctly classified as real samples of the drugs and TN indicates how many spots resulting from the process do not correspond to any drug.

**Table 3.** Performance of spot segmentation in TLC plates with different levels of quality

TLC plates quality	% TP	% TN
Good	98	99
Regular	95	94
Bad	90	89

**Evaluation of the Quality of the TLC Images.** The automatic evaluation of the quality yielded the results shown in the Table 4. This experiment was performed by taking as reference the quality evaluation of the TLC plates, which were visually made by the specialists. This quality was divided into three categories: good, regular and bad for a total of 105 plates. The 95 % of the categories proposed by the specialists were coincident with those proposed by the algorithm.

**Table 4.** Performance of the automatic quality assessment process

Automatic evaluation of the TLC plates quality	% Good	% Regular or bad
Good quality	95	5
Regular or bad quality	4	96

**Drug Identification by the Automatic Analysis of Images of TLC.** The experiment of drugs identification was performed with 43 different patterns of drug. The experiment consisted of identifying automatically the spotted samples of drugs. Three drug samples of the same pattern were dotted in each TLC plate. A pattern of chlorpromazine was also dotted, always in the second lane to be used like internal standard. Each TLC plate was repeated three times at different moments for a total of 9 samples for each drug by using the best effective solvent system. Specialists first determined the presence of each spot and measured the respective  $R_f$  values and appreciated shape and color of the spots. In this experiment identification errors were only considered, without taking into account any manual error. The algorithm identified a total of 362 drug spots, for a 93.5% of effectiveness. The spots that were poorly identified, were always found among the five most similar substances, according to the value of similarity calculated. Several reasons must be mentioned for the identification. Namely, the insufficient quality of some patterns stored, the presence of noise caused by defects in silica plates, the inhomogeneous distribution of some visualization reagents caused by a poor atomization, and the stretching of some spots due to the called “tail” effect. These types of errors that occurs, even when the specialists try to avoid it, but where they get a good result in the manual identification, sometimes could not be resolved by the automatic system. The summary of the performance of the identification is shown in Table 5.

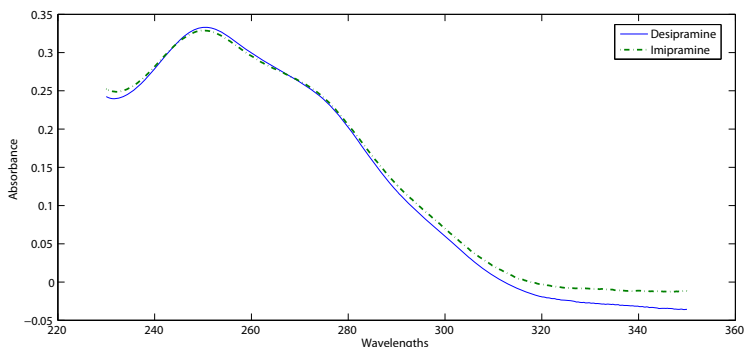
**Table 5.** Performance of the automatic identification of drugs by TLC

Drugs identification results by TLC	% correctly classified	% bad classified
Overall rate	93.5	6.5

## 6.2 Experiments with UV Technique

A total of 43 different substances of abuse were tested according to the defined method for the automatic analysis of UV spectra. We have the patterns for each substance and an independent test set. A value of  $\sigma = 2$  was used in the dissimilarity measure. There are several manners to switch from dissimilarities to similarities, thereby studies can be based on one of them solely. As the purpose of identification is to find the corresponding substance (closest one), although we used a dissimilarity measure, the dissimilarities were converted into similarities for a better understanding. When matching the test samples with the defined patterns, we got that for most test samples, the corresponding pattern was always the most similar one (higher similarity value).

However, as we mentioned before, there are families of substances for which the UV spectra are not differentiable. It is a limitation of the instrumental technique itself, therefore the need of analyzing the substances with different analytical techniques to identify an unknown sample. Only for those cases, the procedure did not work as expected, as the corresponding pattern might not be the most similar one to the test sample. Such is the case of e.g. Desipramine and Imipramine, as their UV spectra are so similar (See Fig. 12), a Desipramine test sample can be more similar to the Imipramine pattern than that of the Desipramine and viceversa. Nevertheless, even for those substances, the corresponding pattern always appears among the first three most similar substances to the



**Fig. 12.** Ultraviolet spectra of Desipramine and Imipramine

test sample. In any case, it is always the analyst responsibility to decide what the correct identification is.

A total of 41 drugs were correctly identified by the proposed automatic method of analysis of the UV spectra.

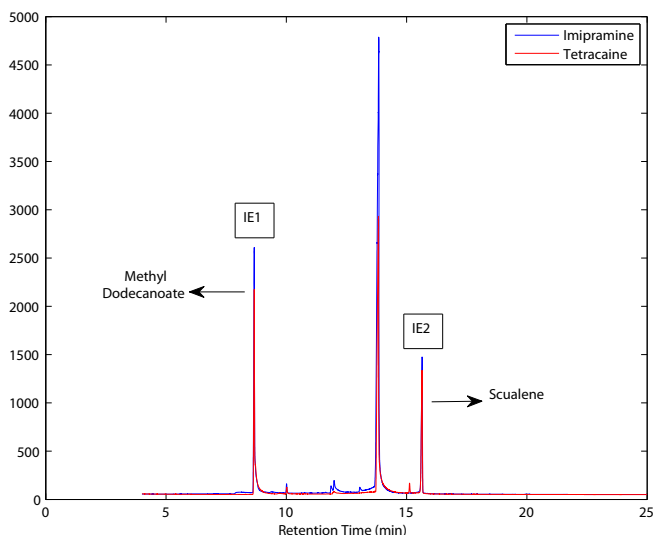
### 6.3 Experiments with GC Technique

The identification procedure by GC was tested for 43 different substances of abuse. We have a lookup table with the data of patterns necessary for the comparisons and an independent test set in which the algorithm was verified.

In all the cases, the peak detection algorithm detected all the peaks corresponding to the internal standards and compound of interest. It means that, all of them were processed in the steps 2 and 3 of the algorithm. Of course, some others peaks were found during the peak detection step, but most of them were rejected during step 2 and 3. Just in the 0.09 % of the analyzed cases, besides the internal standards and the compound of interest, others peaks in the chromatogram were identified with some of the patterns stored in the system. An example of this phenomenon was found in a chromatogram of trifluoperazine, in which was correctly identified the trifluoperazine, and there was also other peak which was matched with Diazepam.

Regarding the identification of the internal standards, in the 95.34 % of the cases both internal standards were correctly identified. It is important to note that in the other cases at least one of the internal standards was identified. In this way, we did not have cases in which we affront the identification process without at least one internal standard. The identification of the internal standards is done taking into account only their absolute retention time. This process is affected by the chromatogram displacements, so those results reached they were expected.

With the above results, after going to step 3, the algorithm identified correctly the drug of abuse of interest in the 98.9 % of the cases. One of the problems found is that there are substances which have very similar retention times; as a consequence



**Fig. 13.** Corresponding chromatograms of Imipramine and Tetracaine which are very similar

their confidence intervals used for identification, even relative to the internal standards, have great intersection. This is the case of Imipramine and Tetracaine, which can not be differentiating by the algorithm. Fig. 13 shows the chromatograms of these two substances and it can be seen that their corresponding peaks almost match.

## 7 Conclusions and Future Works

In this work a set of automatic methods for the identification of drugs by using thin layer chromatography, ultraviolet spectroscopy and gas chromatography techniques are provided.

The main contributions of this work is that it offers the possibility to obtain parallel responses of the identification of drugs from different points of view on a same integrating framework. The proposal helps the analysts to reach more accurate and faster responses in their tasks.

The experimental results demonstrated the feasibility of the proposed methods, obtaining more than 91% of accuracy in the identification of the analyzed drugs for each of the three analytical techniques. It was also corroborated that in almost all cases in which the response of the three techniques match, the drug was correctly identified. Thus, the coincidence of the three techniques gives a high confidence to the identification.

As future work we pretend to development an algorithm for obtaining an automatic final decision in the identification by fusing the responses gives by each analytical technique, incorporating their confidence in the decision. Additionally it is necessary to endow this proposal with the possibility of recognizing mixture of drugs.

## References

- [1] Adobe camera raw y adobe lightroom, <http://www.copiasxl.com/calacr.html>
- [2] Peak detection using matlab, <http://www.billauer.co.il/peakdet.html>
- [3] Barnes, R., Dhanoa, M., Lister, S.: Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy* 43(5), 772–777 (1989)
- [4] Bradski, G., Kaehler, A.: *Learning OpenCV: Computer vision with the OpenCV library*. O'reilly (2008)
- [5] Fearn, T., Riccioli, C., Garrido-Varo, A., Guerrero-Ginel, J.: On the geometry of snv and msc. *Chemometrics and Intelligent Laboratory Systems* 96(1), 22–26 (2009)
- [6] Flusser, J., Zitova, B., Suk, T.: *Moments and Moment Invariants in Pattern Recognition*. Wiley Publishing (2009)
- [7] Gonzalez, R.C., Wood, R.E.: *Digital Image Processing*, 2nd edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2001) ISBN:0201180758
- [8] Hahn-Deinstrop, E.: *Applied thin-layer chromatography: best practice and avoidance of mistakes*. WILEY-VCH Verlag GmbH and Co. KGaA, Weinheim (2007)
- [9] Hess, A.: Digitally enhanced thin-layer chromatography: an inexpensive, new technique for qualitative and quantitative analysis. *Journal of Chemical Education* 84(5), 842–847 (2007)
- [10] Komsta, L., Skibinski, R., Grech-Baran, M., Galaszkiwicz, A.: Multivariate comparison of drugs uv spectra by hierarchical cluster analysis-comparison of different dissimilarity functions. In: *Annales Universitatis Marie Curie-Sklodowska, Polonia*, vol. 20, pp. 2–13 (2007)
- [11] Kosir, A., Tasic, J.: Pyramid segmentation parameters estimation based on image total variation. In: *EUROCON. Computer as a Tool. The IEEE Region 8*, vol. 2, pp. 179–183 (2003)
- [12] Li, J., Brynn-Hibbert, D., Fuller, S., Vaugh, G.: A comparative study of point-to-point algorithms for matching spectra. *Chemometrics and Intelligent Laboratory Systems* 82(1-2), 50–58 (2006)
- [13] Lillsunde, P., Korte, T.: Comprehensive drug screening in urine using solid-phase extraction and combined tlc and gc/ms identification. *Journal of Analytical Toxicology* 15(2), 71–81 (1991)
- [14] Marfil, R., Molina-Tanco, L., Bandera, A., Rodriguez, J., Sandoval, F.: Pyramid segmentation algorithms revisited. *Pattern Recognition* 39(8), 1430–1451 (2006)
- [15] Moffat, A., Osselton, D., Widdop, B., Clarke, E.: *Clarke's analysis of drugs and poisons: in pharmaceuticals, body fluids and postmortem material*. *Clarke's Analysis of Drugs and Poisons*. In: *Clarke's Analysis of Drugs and Poisons: In Pharmaceuticals, Body Fluids and Postmortem Material*, Pharmaceutical Press (2004)
- [16] Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man and Cybernetics* 9(1), 62–66 (1979)
- [17] Paclik, P., Duin, R.P.W.: Dissimilarity-based classification of spectra: computational issues. *Real Time Imaging* 9(4), 237–244 (2003)
- [18] Porro, D., Duin, R.W., Talavera, I., Hdez, N.: The representation of chemical spectral data for classification. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) *CIARP 2009. LNCS*, vol. 5856, pp. 513–520. Springer, Heidelberg (2009)
- [19] Porro-Muñoz, D., Talavera, I., Duin, R., Hernández, N., Orozco-Alzate, M.: Dissimilarity representation on functional spectral data for classification. *Journal of Chemometrics* 25(9), 476–486 (2011)
- [20] Sajewicz, M., Pietka, R., Pienak, A., Kowalska, T.: Application of thin-layer chromatography to investigate oscillatory instability of the selected profen enantiomers in dichloromethane. *Journal of Chromatographic Science* 43(10), 542–548 (2005)

- [21] Sherma, J., Fried, B.: Handbook of Thin-Layer Chromatography. Chromatographic science. Taylor & Francis (2003)
- [22] Siek, T., Stradling, C., McCain, M., Mehary, T.: Computer-aided identifications of thin-layer chromatographic patterns in broad-spectrum drug screening. *Clinical Chemistry* 43(4), 619–626 (1997)
- [23] Spiegel, M.R., Espadas, J.L.G.: Teoría y problemas de estadística. McGraw-Hill, México (1970)
- [24] Varmuza, K., Karlovits, M., Demuth, W.: Spectral similarity versus structural similarity: infrared spectroscopy. *Analytica Chimica Acta* 490(1-2), 313–324 (2003)
- [25] Williams, V.: Detección de curvas generales utilizando la transformada rápida de hough. *Investig. Pensam. Crit.* 2, 03–09 (2004)
- [26] Zhu, N., Wang, G., Yang, G., Dai, W.: A fast 2d otsu thresholding algorithm based on improved histogram. In: Chinese Conference on Pattern Recognition (CCPR), pp. 1–5 (2009)



# Detecting Counterfeit RFID Tags Using Digital Forensic

JingHuey Khor, Widad Ismail, and Mohammad Ghulam Rahman

Auto-ID Laboratory, School of Electrical and Electronic Engineering,  
Universiti Sains Malaysia (USM), 14300 Nibong Tebal, Penang, Malaysia  
khorjinghuey@yahoo.com, {eewidad, eeghulam}@eng.usm.my

**Abstract.** Radio frequency identification (RFID) tag counterfeiting issue is prevalent throughout the world, especially in the access control and information technology sectors. Counterfeit detection is vital in digital forensic practices, especially for counterfeit RFID tag analysis. Hence, electronic fingerprint matching method is proposed to be used as a detection mechanism to detect counterfeit tags. The electronic fingerprint matching method is presented in the digital forensic investigation model that consists of seven phases. The received and backscatter powers of tag are proposed to be used as unique electronic fingerprint in the fingerprint matching method. Two statistical tests, namely, t-test and ANOVA test, are used in the fingerprint matching method. Five fingerprint matching methods are presented and are categorized based on the power response of tag and the statistical test used. Method V which uses three way ANOVA test to analyze backscatter power of tag has the most accurate results. This is because Method V has the highest area under curve (AUC) (0.999) and lowest equal error rate (EER) (0.01) values. Besides that, the false acceptance rate (FAR) and false rejection rate (FRR) obtained are 0.1 % and 1.3 %, respectively. Therefore, the proposed Method V has been proven able to detect counterfeit tags efficiently.

**Keywords:** Counterfeit, EPCglobal Class-1 Generation-2, Fingerprinting Matching Method, RFID.

## 1 Introduction

Radio frequency identification (RFID) technology has been pervasively adopted in many areas, such as supply chain, library, healthcare management, and waste management. RFID that offers contact-less identification and automatic data management enables real-time monitoring of authentic products [1]. Manufacturers have started to utilize RFID technology to fight counterfeiting issues [2-5] due to counterfeiting is prevalent throughout the world, especially in the pharmaceutical and information technology sectors [6]. Counterfeiting issues should be treated seriously because global counterfeit industries generate an estimated \$670 billion annually [7]. Global RFID market is expected to grow at a compound annual growth rate of roughly 17 % to a value of approximately \$9.7 billion in the period 2011 to 2013 [8]. Rapid growth in RFID market is triggered by emerging usage of RFID technology in various applications.

RFID system that consists of tag, reader, and back-end system as well as transmission channels can become the targets of attackers to perform both passive and active attacks [9]. Passive attacks mean monitoring of channels to listen transmitted data without any data modification. On the other hand, active attacks denote modifying transmitted data and alteration of device computation. Information security is benchmarked using three categories, namely, confidentiality, integrity, and availability [10]. Confidentiality can be achieved by using cryptosystems and proper access control to protect data in transmission channels. Integrity can be ensured by proving the data is unique and without any modification [11]. Availability means that the data is always readily available upon request from legitimate users. Hence, all of the categories must be fulfilled to protect RFID system from any security threats.

EPCglobal is an industry-driven reference for RFID standardization that develops standards to describe all the components and architecture of RFID tags, readers, and information systems [12]. EPCglobal Class 1 Generation 2 (Gen 2) standard is the second generation RFID air interface protocol. This protocol is ratified as ISO 18000-6C by International Organization for Standardization [13]. RFID tags that conform to Gen 2 standard are known to be inexpensive and are broadly used in many identification and tracking methods. EPC tags offer higher reliability, greater read range, and enhanced security and privacy protection. Hence, the high performance of the EPC tags is the key contributor to the deployment of RFID technology in various applications. But, the EPC tags are susceptible to cloning attack due to lack of anti-cloning features [14]. EPC tags offer minimal resistance against eavesdropping, which is one of the most serious threats in RFID communication [15]. Communication between a legitimate tag and a reader is often unprotected and can be easily intercepted by adversaries. In addition, an EPC tag is vulnerable to impersonation threat because of its characteristic of releasing data information to any compatible reader [16]. Impersonation occurs when an entity attempts to gain access to resources and information by pretending and adopting the identity of an authorized user [17]. This indicates that EPC tags lack of authentication and encryption, which can enable readers to collect information of the tags they scan. Hence, any adversary can gather required information and can manipulate the collected information to clone counterfeit tags [18]. This information may be used to create counterfeit tags that bear the same information as that of a legitimate tag. Counterfeit tags can be attached to bogus products and disguise these as authentic products in the market [19,20]. The counterfeit tag issue is very serious because it is capable of causing a menace ranging from public privacy and safety issues to loss of industry revenues.

The EPC tag is widely used in electronic passport (E-passport) for identity and document security purpose. For an example, EPC tag has been incorporated with the United States Passport Card in the summer of 2008 [21]. However, the first RFID E-passport in the world is issued by Malaysia in 1998 [22]. The E-passport not only has the identity of the passport holder, but also contains the information of the holder's travel history, including time, date, and place of entries and exits from the country. But, the E-passport is facing tracking, identity theft, and counterfeit attacks because of the vulnerabilities of the EPC tag in preventing passive attacks, such as skimming and eavesdropping attacks. Cloning attack is the most serious issue in the E-passport because it is able to threaten the privacy and security of E-passport holder [21]. Hence, a detection mechanism is proposed as a digital forensic practice to identify the

counterfeit tags. Electronic fingerprint matching method is utilized to detect counterfeit tags using unique electronic fingerprint of tag. The unique electronic fingerprint of RFID tag can be identified based on its physical characteristic [23]. The analysis of the fingerprint can be conducted using various statistical tests. The accuracy of the fingerprint matching method must be verified to determine its accuracy and reliability.

## 2 Preview of RFID Technology

RFID technology was first used in military application, called 'Identify Friend or Foe' during World War II. In recent decade, RFID technology has replaced the barcode system because of its higher reliability, read rate, and read range. The high performance of RFID is a key contributor to the deployment of RFID technology in various applications. RFID is a technology that allows RFID reader to remotely send command to read and store information on RFID tag. An RFID system comprises of three major components, namely, RFID reader, tag, and back-end system. RFID middleware application uses multiple scripting languages, including JavaScript, extension markup language, and hypertext preprocessor [17].

### 2.1 RFID Reader

An RFID reader consists of an antenna, a microprocessor, and an interface device for forwarding data to the back-end system [9]. RFID reader communicates with tag and back-end system by receiving and sending data in the transmission channel. In addition, RFID reader is capable to write data in tag memory, authenticates tag as well as powers up passive RFID tag via electromagnetic field. RFID reader can be categorized into two categories, namely, stationary reader and mobile reader [24]. Stationary reader has a fixed location and network connection. In contrast, mobile or handheld reader can be moved around that offer a more flexible applications.

### 2.2 RFID Tag

An RFID tag consists of antenna, microchip, and encapsulation. RFID tag can be classified based on its functionality, power supply, and operating frequency.

- i. Functionality

RFID tags have memory size from a single bit up to several kilobytes. Tag memory technologies are categorized into non-volatile and volatile storages. The non-volatile storage includes read only, write once read many (WORM), and read/write. The volatile storage is used for performing calculation after tag power up. Tag can be classified into five broad classes based on the tag computation capability [25].

Class 1: Passive read-only tags offer only basic functionality, including a fixed EPC identifier, a tag identifier, kill function, and optional password-protected access control.

Class 2: Passive tags offer same functionality as Class 1 but with read-write memory as well as extended tag identifier, user memory and authenticated access control.

Class 3: Semi-passive tags offer all Class 2 functionality as well as possess sensor and on-tag power source.

Class 4: Active tags offer all Class 3 functionality as well as tag-to-tag communications and ad-hoc networking.

Class 5: Active tags can communicate with all classes of tags.

ii. Power supply

Passive tag obtains power from the electromagnetic field of reader. The tag does not have an internal source of power. Semi passive tag has own power supply for the microchip but communicate by obtaining energy from the electromagnetic field of reader. Active tag uses own source of power (i.e. battery) to support all activities. Hence, active tag has more functionality and able to communicate over a longer distance compared to passive tag. However, passive tag outperformed active tag in terms of cost, size, and lifetime. This is because passive tag offers low-cost, small size, and economic lifetime due to no internal source of power.

iii. Operating frequency

The operating frequency of a tag is categorized into four categories to enable communication between tag and reader. The electromagnetic spectrum consists of low frequency (LF) (125-134 kHz), high frequency (HF) (13.56 MHz), ultra high frequency (UHF) (860- 960 MHz), and microwave (2.54-5.8 GHz) [1]. Different frequencies have different physical properties. HF tags are designed to carry more data and longer read range. LF tags offer better signal penetration of objects and have little absorption through liquid [26].

## 2.3 Back-End System

A back-end system is required to process data obtained from tags. A back-end system consists of two parts, namely, middleware and applications. Middleware plays an important role in back-end system to perform all data grouping and filtering tasks [27]. Middleware is used to provide unified interface and semantics towards various applications. The middleware should be designed with full-features and be able to support different hardware [28]. The middleware acts as a server to connect hardware at one end and support a number of applications at another end. Applications are the software components that act as an end user interface of a complete RFID system. Applications are used to interpret the data obtained from reader and configure the middleware.

An RFID middleware consists of four layers, namely, reader interface, data processor and storage, application interface, and middleware management.

- i. Reader interface  
Reader interface is the lowest layer of middleware that used to handle interaction with the hardware.
- ii. Data processor and storage  
This layer processes all the raw data obtained from the reader by storing, filtering, and grouping the obtained data.
- iii. Application interface  
This layer configures the RFID middleware by providing an application programming interface for the applications. The layer manages the application with the interface of middleware.
- iv. Middleware management  
This layer manages the configuration of middleware by providing information to the processes running in the middleware.

### 3 Related Works

Low-cost RFID system is susceptible to cloning attack. Many cryptographic protocols have been proposed by other researchers to prevent cloning attack. But, the cryptographic protocols are proved unable to protect the RFID tag from being duplicated. Hence, researchers start to use physical characteristic of tag as a unique fingerprint in detecting cloning tag instead of cryptographic protocol. Each RFID tag is unique in terms of their radio frequencies and manufacturing differences. Although adversaries can duplicate the information of tag memory, they cannot develop a tag that has the same unique fingerprint of the tag [29]. The unique physical characteristics of RFID tags enable the creation of electronic fingerprint for the detection of cloning tags. Cloning tags could be detected by comparing the extracted specific fingerprints of tags and the stored fingerprints of legitimate one. When the extracted fingerprints match the ones stored in the system, this indicates the tag is legitimate; otherwise, the tag could be considered as counterfeited.

Electronic binding [29] is a method introduced to store unique fingerprint of product into tag memory. The fingerprint is signed by the manufacturer of the product and can be verified by an authentication device. The fingerprint is acted as a digital signature to identify the authenticity of a product. The authentication device is used to regenerate the product's fingerprint to compare with the value stored in the tag memory. However, this method guarantees the authenticity of the product rather than the authenticity of the tag. The cloning threat is existed due to the information stored in the tag can be copied and duplicated. Hence, unique fingerprint of tag shall be stored in the tag memory instead of the unique fingerprint of the product. Any counterfeit tag can be easily detected due to each tag has its own unique fingerprint. Although adversaries can duplicate the information of tag memory, but they are unable to develop a tag that has the same unique fingerprint as stored in the tag.

Physical-layer identification of passive UHF RFID tags from three different manufacturers is analyzed in [30]. RFID reader that capable to simulate an inventory protocol is built to activate tags. RF signal features are extracted from the preambles of tags' replies. Time domain and spectral features of the collected signals are analyzed. The tags can be classified with an accuracy of 71.4 % from different

locations and distances to the reader based on the time domain features. In addition, UHF RFID tag is proved to be uniquely identified in controlled environment based on the signal spectral features with 0 % of EER. But, the physical-layer identification method is complex and the reader used in conducting the experiment is purposely built. By contrast, the physical characteristics of low-cost tag such as minimum power response as well as received and backscatter powers of tag can be easily obtained with any compatible reader and tag. Hence, unique electronic fingerprint of low-cost tag can be obtained easily by using a spectrum analyzer.

### **3.1 Conventional Electronic Fingerprint Matching Method**

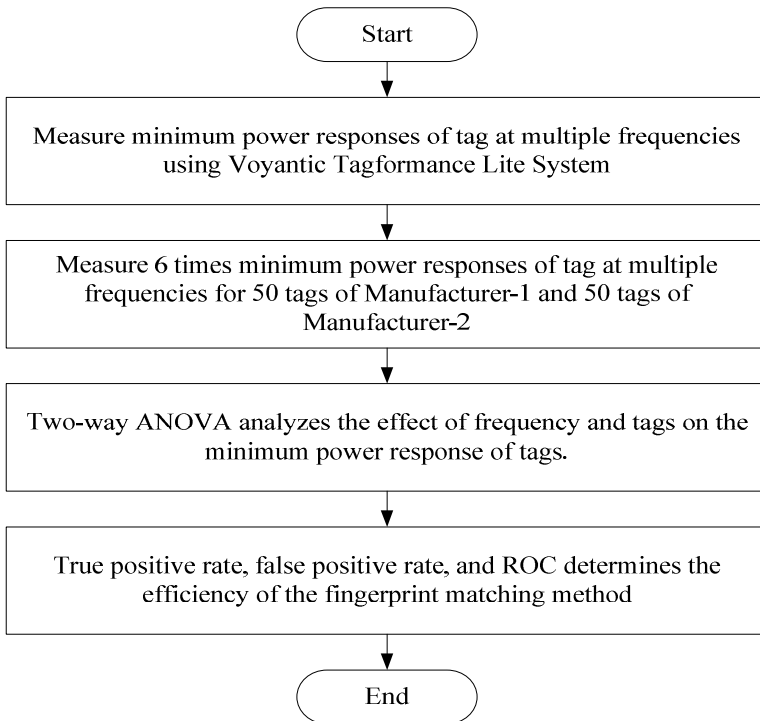
Minimum power response of tag measured at multiple frequencies is used as an unique electronic fingerprint by [31]. The minimum power response is measured using a Voyantic Tagformance Lite System. A bottom-up algorithm is used to send signal repeatedly from a reader to the tag starting from -20 dBm, incrementing it by 0.01 dBm until a response from the tag is detected. The tag is horizontally polarized along with the antenna and is mounted at same distance and position from the reader. The minimum power responses at all the frequencies are measured 6 times for each 100 passive RFID tags to obtain a total of 600 measurements. There are two types of tags taken from two major manufacturers. 50 of 100 tags are labelled as Manufacturer-1, and another 50 tags are labelled as Manufacturer-2.

Two-way ANOVA is applied to test the effect of frequency and tag types on the minimum power response of tags. The null hypothesis is there are no differences in the minimum power responses at different frequencies, there are no differences in the means of the minimum power responses for different tags, and there is no significant interaction between the two factors. The fingerprint matching method is verified using true positive rate, false positive rate, and ROC. The overall fingerprint matching method using minimum power responses at multiple frequencies are shown at Figure 1.

## **4 Digital Forensic Practice in Detecting Counterfeit Tags**

Digital forensic is a branch of forensic science based on scientifically proven methods to collect and analyze digital information as source of evidence in investigations and legal proceedings [32]. Digital forensic is not limited to computer, but it also encompasses any electronic devices that able to store data, such as cell phones, RFID tags, and GPS units. Digital forensic can be used in criminal investigation, civil litigation, intelligence, private sector, and administrative matters [33]. A digital forensic investigation model was presented by [34]. The model consists of seven phases, namely identification, preservation, collection, examination, analysis, presentation, and decision phases.

In this chapter, electronic fingerprint matching method is proposed to distinguish counterfeit tags from legitimate tags. The power responses of tag used can be categorized into received and backscatter powers of tag. The statistical tests used

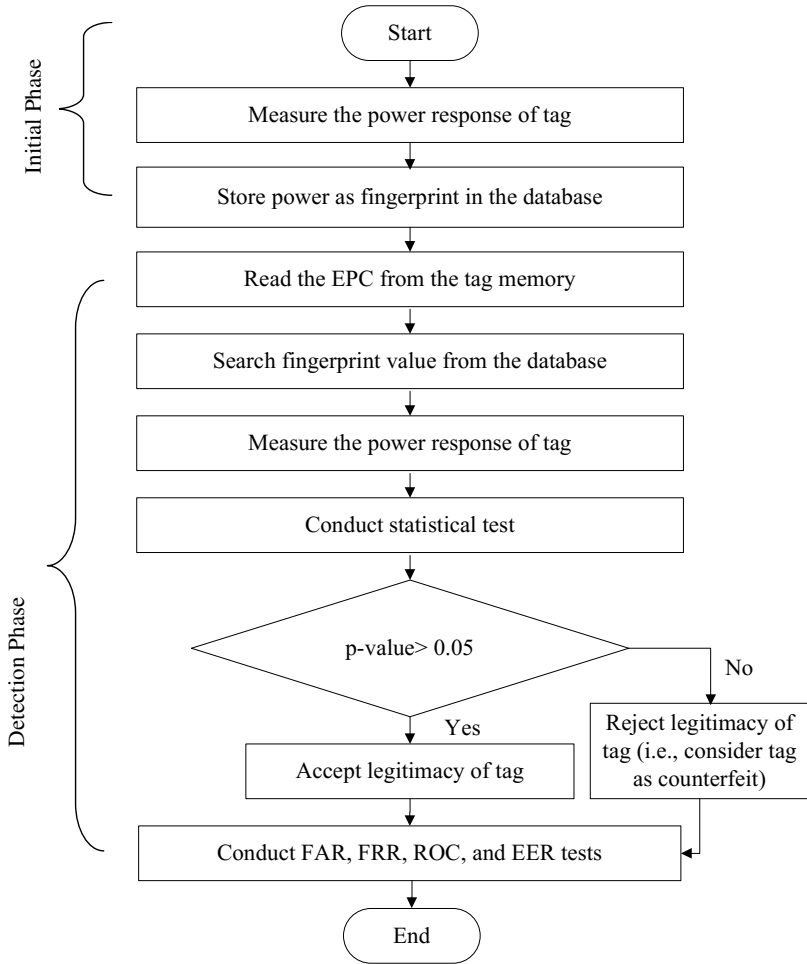


**Fig. 1.** Overall development process of conventional fingerprint matching method

consist of t-test, two-way, and three-way ANOVA tests. The fingerprint matching is categorized into five methods based on the types of power response of tag, which measures at different tag positions and different statistical tests used. The accuracy of the five electronic fingerprint matching methods is verified based on the false acceptance rate (FAR), false rejection rate (FRR), area under curve (AUC), and equal error rate (EER) values. An excellent electronic fingerprint matching method should have high AUC as well as low FAR, FRR, and EER values. The proposed RFID tag fingerprint matching method illustrated in Figure 2 consists of an initial phase and a detection phase.

In the initial phase, the power response of each Gen 2 tag is measured in a controlled environment. The power responses of tag, including received and backscatter powers of tag, are used as unique electronic fingerprint. Both the power responses of tags are measured at the frequency ranges from 919 MHz to 923 MHz. Next, the measured power responses of tags are stored in the database for further reference.

In the detection phase, the EPC of a suspected counterfeit tag is read. The stored fingerprint value is searched in the database according to the tag's EPC that acted as an index. In addition, the power response of the suspicious tag is measured using the same measurement platform. The stored fingerprint and measured fingerprint are then compared using the statistical test algorithm. The t-test and ANOVA test algorithms



**Fig. 2.** Overall process of proposed fingerprint matching method

are used to determine the true nature of the tags (genuineness or otherwise). The suspected counterfeit tag is proven to be a legitimate tag if the p-value for both t-test and ANOVA algorithms is greater than 0.05. Otherwise, the tag is proven as a counterfeit tag. Finally, the accuracy of the fingerprint matching method is verified using the FAR, FRR, ROC, and EER.

The electronic fingerprint matching method is presented in the digital forensic investigation model. The seven phases of electronic fingerprint matching method are described as follows:

i. Identification

The potential of unique electronic fingerprint of tag that can be used to accurately distinguish between legitimate and counterfeit tag are notified.



- ii. **Collection Phase**  
The collection phase of the digital forensic process is when the unique fingerprint of tag are measured and collected.
- iii. **Examination**  
The validity of the power response of tag to be used as unique fingerprint of tag is examined to verify its accuracy and reliability.
- iv. **Preservation Phase**  
The preservation phase of the digital forensic process is the preservation of the unique fingerprint of tag value in a manner that is reliable, complete, accurate, and verifiable.
- v. **Analysis Phase**  
The analysis phase of the digital forensic process analyzes the extraction of the unique electronic fingerprint of tag using various statistical tests.
- vi. **Presentation Phase**  
The presentation phase of the digital forensic process is to accurately present the result of the analysis of the unique electronic fingerprint of tag.
- vii. **Decision Phase**  
Counterfeit tag is verified if the p-value obtained from the statistical test is less than 0.05, which is the significance level that used in most of the RFID system.

#### **4.1 Identification Phase**

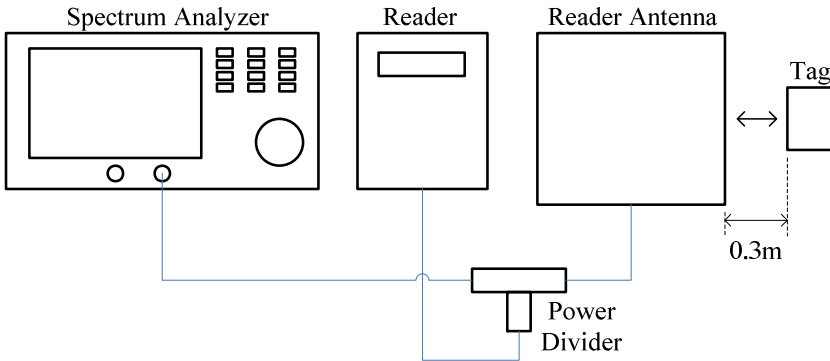
Physical characteristic of low-cost RFID tags can be used as a unique electronic fingerprint to detect cloning tags. The RFID tags are unique because of having divergent radio frequencies and manufacturing differences. Hence, the power responses of tag include received and backscatter powers of tag can be used as unique electronic fingerprint. The reliability of the received and backscatter powers of tag to be use in the fingerprint matching method must be tested and verified.

#### **4.2 Collection Phase**

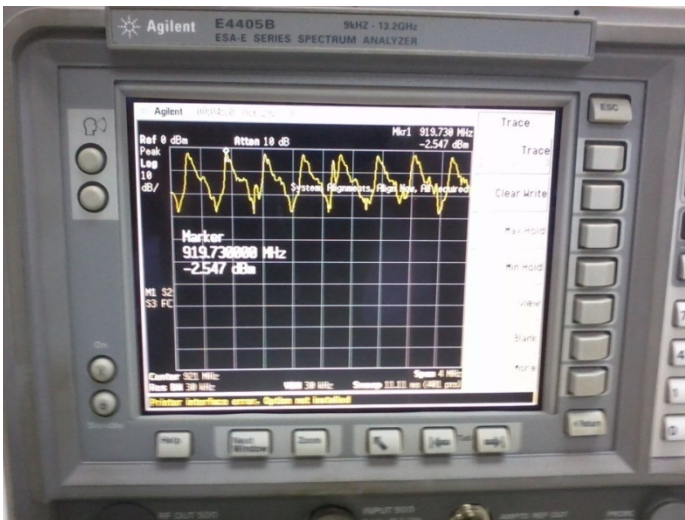
Received and backscatter powers of tag can be used as electronic fingerprint in creating a unique identity for RFID tag. The received power of tag can be obtained by calculating the transmitted power of reader using Friis transmission equation. The readers used in measuring the transmitted power of reader and backscatter power of tag must be able to operate at UHF 919 MHz to 923 MHz and supports the Gen 2 protocol. The antenna and tag must be placed at a fixed position to obtain accurate and reliable results. The transmitted powers of reader and backscatter powers of tag are measured at 8 frequency bands, which are between the ranges from 919.25 MHz to 922.75 MHz, with increments of 0.5 MHz.

**4.2.1 Measurement Platform of Received Power of Tag**

The measurement of the reader transmitted power platform is shown in Figure 3. The setup consists of a passive RFID reader and antenna, a passive Gen 2 tag, and a spectrum analyzer. To determine a precise transmitted power of reader, the cable loss and power loss within the power splitter must be considered [35]. Hence, the power value obtained from the spectrum analyzer is added to the total power loss to obtain an accurate transmitted power of reader. There are a total of 8 peak signals from the range of 919 MHz and 923 MHz as shown in Figure 4. Hence, the frequency bands of 8 peak signal are selected to measure the transmitted power of reader.



**Fig. 3.** Measurement of the reader transmitted power platform



**Fig. 4.** Reader transmitted power measured with spectrum analyzer

The received power of tag is calculated using the Friis transmission equation, as demonstrated in Eq. (1) [36].

$$P_r = P_t G_t G_r \left( \frac{\lambda}{4\pi R} \right)^2 \quad (1)$$

where,  $P_r$  is the power received by the tag antenna and  $P_t$  is the power input to the reader antenna. In addition,  $G_r$  is the antenna gain of the reader antenna,  $G_t$  is the antenna gain of the tag antenna,  $\lambda$  is the wavelength, and  $R$  is the distance between the reader and tag antennas. The Friis transmission equation is only applicable in the Fraunhofer region. Hence, a minimum Fraunhofer region is determined by using Eq. (2) [37].

$$r_{ff} = \frac{2D^2}{\lambda} \quad (2)$$

where,  $r_{ff}$  is the minimum far field distance,  $D$  is the diameter of the transmitting antenna, and  $\lambda$  is the wavelength. The diameter of the transmitting antenna is 0.185 m and the wavelength is approximately 0.33 m for all 8 frequency bands. Hence, the minimum far field distance is 0.21 m. The tag should be placed at a distance greater than 0.21 m such that it is in the Fraunhofer region. In this setup, the distance between the tag and the reader antenna is 0.3 m to satisfy the Fraunhofer region condition. Hence, the received power of tag is analyzed using the t-test and the ANOVA test at a distance of 0.3 m only. The parameters used in the measurement are shown in Table 1.

**Table 1.** Parameters used in Friis Transmission Equation

Parameters	Value
Gain of reader antenna	6 dBi
Gain of tag antenna	2.15 dBi
Gain of reader antenna in power ratio, $G_t$	3.981
Gain of tag antenna in power ratio, $G_r$	1.641
Frequency, $f$	919.25-922.75 MHz
Wavelength, $\lambda$	0.33 m
Distance between reader and tag antennas, $R$	0.3 m

#### 4.2.2 Measurement Platform of Backscatter Power of Tag

The measurement of the backscatter power of tag platform is shown in Figure 5. The setup consists of a passive RFID reader, two antennas, a passive Gen 2 tag, and a spectrum analyzer. The measurement of the backscatter power of tag is made where the distance between the antennas and tag is at 0.1 m, 0.2 m, and 0.3 m. There are a total of 8 peak signals from the range of 919 MHz and 923 MHz as shown in Figure 6. Hence, the frequency bands of 8 peak signal are selected to measure the backscatter power of tag. The antenna connected to the reader is used to transmit power and activate a corresponding tag. On the other hand, the antenna connected to a spectrum analyzer is used to receive the tag backscatter power. The distance between the two

antennas is fixed at 0.132 m which is obtained based on the formula of  $0.4 \lambda$ . The value of  $\lambda$  used is 0.33 m by referring to the Table 1. The ideal distance between two adjacent antennas is  $0.4 \lambda$  according to [38,39]. Hence, the distance of  $0.4 \lambda$  is selected in the proposed measurement to obtain the slightest effect of mutual coupling and spatial correlation.

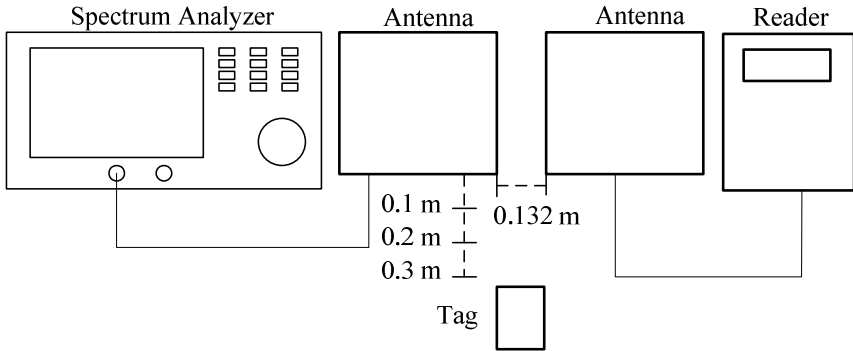


Fig. 5. Measurement of backscatter power of tag platform at 0.1, 0.2, and 0.3 m

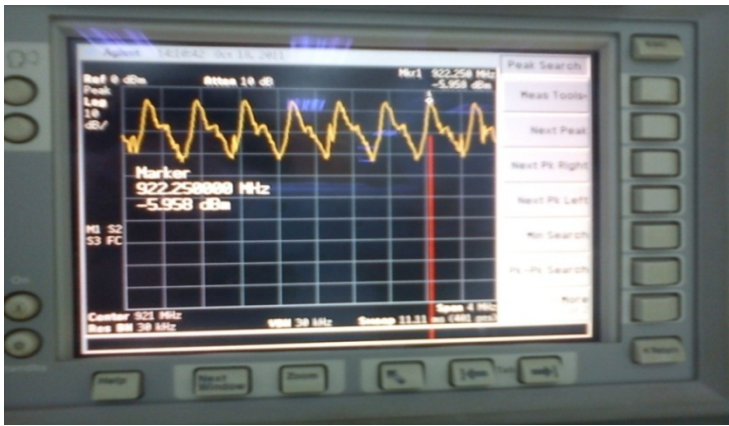


Fig. 6. Backscatter power of tag measured with spectrum analyzer at 0.2 m

### 4.3 Examination Phase

Both received and backscatter powers of tag are measured for 100 passive RFID tags at a fix temperature and in a controlled environment. The reliability of the received and backscatter powers of tag to be used as a unique electronic fingerprint is verified by choosing 5 from the total 100 passive RFID tags which are having the closest mean values.

### 4.3.1 Received Power of Tag

The received powers of 5 tags which having the closest mean values are significantly different as shown in Figure 7. Hence, the received power of tag can be used as a unique fingerprint to identify legitimate and counterfeit tags. The legitimate tag fingerprint template is determined by obtaining the average received power of 50 readings per tag.

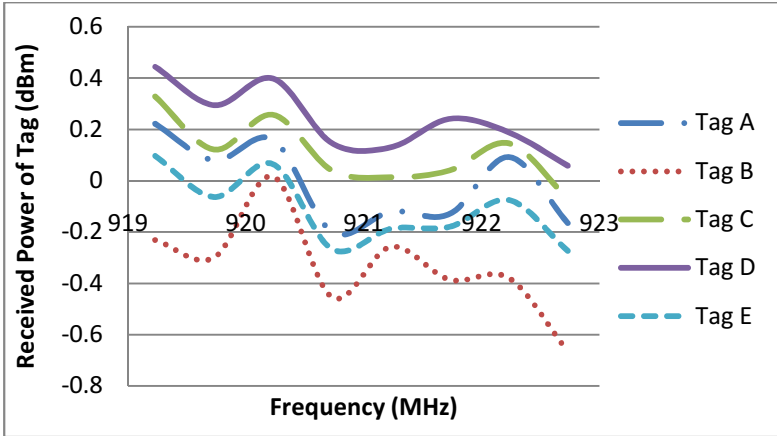


Fig. 7. Received power of tag at 919–923 MHz

### 4.3.2 Backscatter Power of Tag

The 5 tags which having the closest mean values are significantly different from the measurement result as shown in Figure 8. Hence, the backscatter power of tag can be used as a unique fingerprint to identify legitimate and counterfeit tags. The legitimate

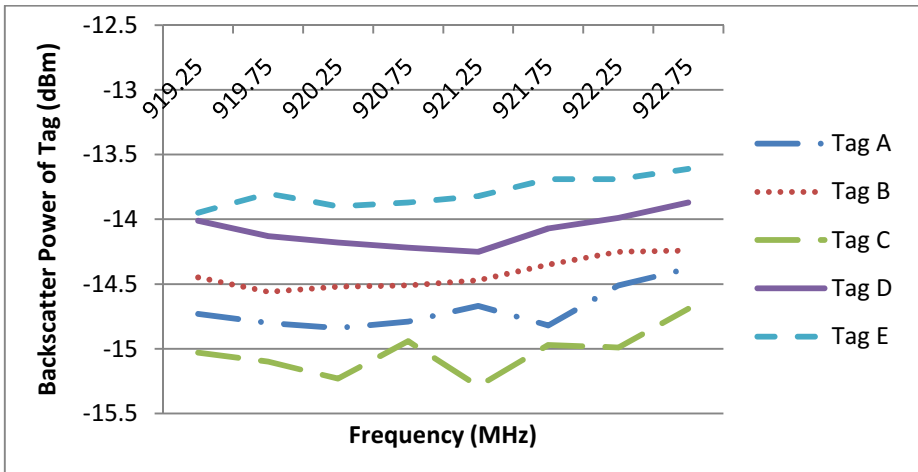


Fig. 8. Backscatter power of tag at 919 MHz to 923 MHz

tag fingerprint template is determined by obtaining the average backscatter power of 50 readings per tag.

#### 4.4 Preservation Phase

The unique electronic fingerprint is only stored in the database to protect the confidentiality of the fingerprint value from being obtained by adversaries. The unique fingerprint value stored in the database can be searched based on the tag's EPC. Hence, the stored fingerprint value in the database and the measured fingerprint value obtained from the experimental measurement can be compared to verify the genuineness of the tag.

#### 4.5 Analysis Phase

A null hypothesis is made where the suspicious tag is considered as a legitimate tag based on the results of the statistical tests, namely, the t-test and the ANOVA test. The t-test is used when there are only two groups (legitimate tag type and suspicious tag type) and one dependent variable (either received or backscatter power of tag) is obtained. In other word, tag type is the primary factor. The frequency band is fixed at 8 different levels, i.e., 919.25 MHz, 919.75 MHz, 920.25 MHz, 920.75 MHz, 921.25 MHz, 921.75 MHz, 922.25 MHz, and 922.75 MHz. The ANOVA test is used when there are more than two groups as well as one dependent variable (either received or backscatter power of tag) is obtained. For the two-way ANOVA, the primary factors are tag type and frequency. For the three-way ANOVA, the primary factors are tag type, frequency, and tag position.

##### 4.5.1 T-Test

The t-test algorithm is a statistical test used to identify differences in the means and variances of two populations. The formula of the t-test algorithm is illustrated in Eq. (3) [40].

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_p^2 \left( \frac{1}{N_1} + \frac{1}{N_2} \right)}}$$

$$S_p^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} \quad (3)$$

where,  $\bar{X}_1$  and  $\bar{X}_2$  are the means of the reference and suspect groups,  $N_1$  and  $N_2$  are the number of samples for the reference and suspect groups, respectively,  $S_1$  and  $S_2$  are the sample variances, and  $S_p^2$  is the pooled variance.

### 4.5.2 ANOVA Test

The ANOVA test is a procedure to test the equality of mean among two or more groups [41]. The two-way ANOVA is conducted when two factors (tag group and frequency) are used and only one dependent variable (received or backscatter power of tag) involved in the data analysis. It is used to test the effect of each two factors and the interaction between the factors [31]. The null hypothesis for this test is that the means of the power response of tag for different frequency bands between the two tag groups is equal. The formula of the two-way ANOVA is shown in Eq. (4) [42].

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

for  $i = 1,2 \dots, 8; j = 1,2; k = 1$  (4)

where,  $y_{ijk}$  is the power response of tag,  $\mu$  is the overall mean value response,  $\alpha_i$  is the effect due to the  $i^{\text{th}}$  level of frequency bands, and  $\beta_j$  is the effect due to the  $j^{\text{th}}$  level of tag groups. In addition,  $\gamma_{ij}$  is the effect due to the interaction between the  $i^{\text{th}}$  level of frequency bands and the  $j^{\text{th}}$  level of tag groups. Besides that,  $\varepsilon_{ijk}$  is the error obtained from the statistical test.

By contrast, a three-way ANOVA is used when there are three factors (tag group, frequency, and position) and only one dependent variable (backscatter power of tag) involved in the analysis. The null hypothesis for this test is that the means of the power response of tag for different frequency bands and tag positions between the two tag groups is equal. Eq. (5) [43] indicates the formula of the three-way ANOVA.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

for  $i = 1,2 \dots, 8; j = 1,2; k = 1,2,3; l = 1$  (5)

where,  $y_{ijkl}$  is the power response of tag,  $\mu$  is the overall mean value response,  $\alpha_i$  is the effect due to the  $i^{\text{th}}$  level of frequency bands,  $\beta_j$  is the effect due to the  $j^{\text{th}}$  level of tag types and  $\gamma_k$  is the effect due to the  $k^{\text{th}}$  level of tag positions.  $(\alpha\beta)_{ij}$  is the effect due to the interaction between the  $i^{\text{th}}$  level of frequency bands and the  $j^{\text{th}}$  level of tag groups. Moreover,  $(\alpha\gamma)_{ik}$  is the effect due to the interaction between the  $i^{\text{th}}$  level of frequency bands and the  $k^{\text{th}}$  level of tag positions. In addition,  $(\beta\gamma)_{jk}$  is the effect due to the  $j^{\text{th}}$  level of tag groups and the  $k^{\text{th}}$  level of tag positions. Besides that,  $(\alpha\beta\gamma)_{ijk}$  is the effect due to the interaction between the  $i^{\text{th}}$  level of frequency, the  $j^{\text{th}}$  level of tag groups and the  $k^{\text{th}}$  level of tag positions. Further,  $\varepsilon_{ijkl}$  is the error obtained from the statistical test.

### 4.6 Presentation Phase

The accuracy of the proposed fingerprint matching methods in distinguishing between legitimate and counterfeit tags is analyzed using FAR, FRR, ROC, and EER. Four outcomes from the data are obtained from the fingerprint values of 1000 reading from a legitimate tag and 1000 reading from a counterfeit tag. A 2 x 2 contingency table is used to verify 4 outcomes from the data obtained as indicated in Table 2.

The outcome is a true acceptance (TA) when the statistical test proves that the measured fingerprint is a genuine value and the fingerprint value match with the value recorded in the database. The outcome is a false acceptance (FA) when the measured fingerprint has a genuine value based on statistical test, but the fingerprint value is proven as a bogus value according to the database record. Conversely, true reject (TR) is obtained when the statistical test verifies that the measured fingerprint is a bogus value, and the fingerprint value is unmatched based on the database record. False reject (FR) is obtained when the tag is proved as a counterfeit tag based on statistical test, but the database record shows that the tag is a legitimate tag.

**Table 2.** Four outcomes from the fingerprint matching method

		<b>Genuineness of measured fingerprint based on database</b>	
		Yes	No
<b>Verification genuineness of measured fingerprint based on Statistical test algorithm</b>	Yes	TA	FA
	No	FR	TR

FAR is the measurement of probability in which a reader falsely identifies counterfeit tags as legitimate. FRR is the measurement of probability in which a reader falsely identifies legitimate tags as counterfeit. The FAR and the FRR are calculated using Eq.(6) and Eq.(7) [44].

$$FAR = \frac{FA}{FA + TR} \tag{6}$$

$$FRR = \frac{FR}{FR + TA} \tag{7}$$

The ROC and EER are used to evaluate the performance of the statistical tests in verifying the measured fingerprint with the stored fingerprint. The ROC curve plots the true acceptance rate (TAR) versus its FAR. EER is the rate at which both the FA error and FR error are equal. The lower the EER, the more accurate the fingerprint matching method would be [45-47].

The AUC is a measurement of the performance of the statistical tests in distinguishing between two fingerprint data sets. The accuracy of the statistical tests is verified using a rough guide for classifying the accuracy of a test as shown in Table 3 [48,49].



**Table 3.** Accuracy of test categorization

<b>AUC Range</b>	<b>Categories</b>
0.50-0.60	Failure
0.60-0.70	Poor
0.70-0.80	Fair
0.80-0.90	Good
0.90-1.00	Excellent

**4.7 Decision Phase**

The smaller the p-values for the individual factors and the joint effects of the factors, the stronger the proof is against the null hypothesis [50]. The p-value is the smallest level of significance that would lead to rejection of the null hypothesis. The most popular significance level used in the RFID research is 0.05 [51]. The tag used can be considered as counterfeit if the p-value obtained is less than the significance level ( $\alpha$ ). Hence, the null hypothesis is rejected if p-value is less than 0.05.

**5 Fingerprint Matching Methods**

The electronic fingerprint matching methods are analyzed using difference power responses of tag and statistical tests. The power responses of tag used can be categorized into received and backscatter powers of tag. The statistical test used consists of t-test, two-way, and three-way ANOVA tests. The fingerprint matching is categorized into 5 methods based on the types of power response of tag, which measures at different tag positions and different statistical tests used. The overview of the fingerprint matching methods is shown in Table 4.

**Table 4.** Fingerprint matching methods

<b>Method</b>	<b>Power response of tag</b>	<b>Distance (m)</b>	<b>Statistical test</b>
I	Received power	0.3	T-test
II	Received power	0.3	Two-way ANOVA test
III	Backscatter power	0.3	T-test
IV	Backscatter power	0.1,0.2,0.3	Two-way ANOVA test
V	Backscatter power	0.1,0.2,0.3	Three-way ANOVA test

The statistical tests were performed using SPSS software version 16.0. Parts of the outputs from the SPSS for the 5 methods are in Appendices A-J. The result of the tests are presented and discussed in the following sections.

## 5.1 Result of Statistical Tests Using Legitimate Tag as Suspicious Tag

### 5.1.1 Method I

The response is received power of tag.

For this method, the hypotheses relating to tag type is tested. Based on the SPSS output in APPENDIX A, the p-value for tag type is 0.319. Since this is greater than  $\alpha = 0.05$ , the suspicious tag can be considered as legitimate tag.

### 5.1.2 Method II

The response is received power of tag.

For this method, the hypotheses relating to tag type and the joint effect of tag type and frequency are tested. Based on the SPSS output in APPENDIX B, the p-value for tag type is 0.738 and the p-value for joint effect of tag type and frequency factors is 0.492. Since both the p-values are greater than  $\alpha = 0.05$ , the suspicious tag can be considered as legitimate tag.

### 5.1.3 Method III

The response is backscatter power of tag.

For this method, the hypotheses relating to tag type is tested. Based on the SPSS output in APPENDIX C, the p-value for tag type is 0.739. Since the p-value is greater than  $\alpha = 0.05$ , the suspicious tag can be considered as legitimate tag.

### 5.1.4 Method IV

The response is backscatter power of tag.

For this method, the hypotheses relating to tag type and the joint effect of tag type and frequency are tested. Based on the SPSS output in APPENDIX D, the p-value for tag type is 0.997 and the p-value for joint effect of tag type and frequency factors is 0.354. Since both the p-values are greater than  $\alpha = 0.05$ , the suspicious tag can be considered as legitimate tag.

### 5.1.5 Method V

The response is backscatter power of tag.

The hypotheses relating to tag type, tag type and tag position, tag type and frequency, as well as tag type, tag position, and frequency, are tested. Based on the SPSS output in APPENDIX E, the p-values for all of the above cases (i.e., 0.706, 0.486, 0.967, and 0.601, respectively) are all greater than  $\alpha = 0.05$ . Hence, the suspicious tag can be considered as legitimate tag.

## 5.2 Result of Statistical Tests Using Counterfeit Tag as Suspicious Tag

### 5.2.1 Method I

The response is received power of tag.

For this method, the hypotheses relating to tag type is tested. Based on the SPSS output in APPENDIX F, the p-value for tag type is 0.000. Since this is less than  $\alpha = 0.05$ , the suspicious tag can be considered as counterfeit tag.

### 5.2.2 Method II

The response is backscatter power of tag.

For this method, the hypotheses relating to tag type and the joint effect of tag type and frequency are tested. Based on the SPSS output in APPENDIX G, the p-value for tag type is 0.000 and the p-value for joint effect of tag type and frequency factors is 0.037. Since both the p-values are less than  $\alpha = 0.05$ , the suspicious tag can be considered as counterfeit tag.

### 5.2.3 Method III

The response is backscatter power of tag.

For this method, the hypotheses relating to tag type is tested. Based on the SPSS output in APPENDIX H, the p-value for tag type is 0.000. Since the p-value is less than  $\alpha = 0.05$ , the suspicious tag can be considered as counterfeit tag.

### 5.2.4 Method IV

The response is backscatter power of tag.

For this method, the hypotheses relating to tag type and the joint effect of tag type and frequency are tested. Based on the SPSS output in APPENDIX I, the p-value for tag type is 0.025 and the p-value for joint effect of tag type and frequency factors is 0.015. Since both the p-values are less than  $\alpha = 0.05$ , the suspicious tag can be considered as counterfeit tag.

### 5.2.5 Method V

The response is backscatter power of tag.

The hypotheses relating to tag type, tag type and tag position, tag type and frequency, as well as tag type, tag position, and frequency, are tested. Based on the SPSS output in APPENDIX J, the p-values tag as well as tag and tag position (i.e., 0.000, and 0.000, respectively) are less than  $\alpha = 0.05$ . Hence, the suspicious tag can be considered as counterfeit tag.

## 6 Accuracy of Fingerprint Matching Methods

### 6.1 Accuracy of T-Test Analysis Based on Received Power (Method I)

Method I is an electronic fingerprint matching method using the t-test to analyze the received power of tag for each of the 8 frequency bands. The FARs and the FRRs obtained from Method I are in the range of 42.7 % to 60.3 % and 21.8 % to 34.5 %, respectively, as indicated in Table 5. Method I is having the lowest FA and FR at the 921.25 MHz frequency band. Method I falsely accepts 427 counterfeit tags as legitimate tags based on the received power measured at the 921.25 MHz frequency band. In addition, Method I rejects 218 legitimate tags according to the p-value obtained from the t-test. Hence, the lowest FA and FR values contribute to the lowest FAR and FRR at 921.25 MHz frequency band, with 42.7 % and 21.8 %, respectively. By contrast, there is the highest FA and FR for the received power of tag which measured at the 920.75 MHz frequency band. In this case, Method I accepts 603

counterfeit tags and rejects a total of 345 legitimate tags based on the t-test analysis result. Therefore, Method I which analyzes the data obtained at the 920.75 MHz frequency band has the highest FAR and FRR (60.3 % and 34.5 %) because of its highest FA and FR obtained.

**Table 5.** FARs and FRRs of Method I

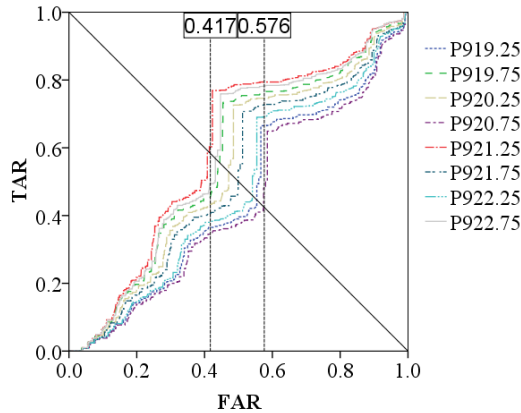
Frequency (MHz)	TA	FR	TR	FA	FAR (%)	FRR (%)
919.25	679	321	418	582	58.2	32.1
919.75	772	228	538	462	46.2	22.8
920.25	743	257	506	494	49.4	25.7
920.75	655	345	397	603	60.3	34.5
921.25	782	218	573	427	42.7	21.8
921.75	722	278	477	523	52.3	27.8
922.25	679	321	433	567	56.7	32.1
922.75	668	232	549	451	45.1	23.2

The AUCs of Method I are in the range of 0.435 to 0.596 as in Table 6, and the EERs obtained from the plotted graph are in the range of 0.417 to 0.576 as in Figure 9. Method I has the highest AUC with 0.596 and the lowest EER with the value of 0.417 at the 921.25 MHz frequency band. This is because Method I has the lowest FAR and FRR values at the 921.25 MHz frequency band. Hence, Method I which analyze the data obtained at the 921.25 MHz frequency band is capable to distinguish legitimate from counterfeit tags more accurately compared to other frequency bands. This is because the received power measured at the 921.25 MHz frequency band has higher values compared to other frequency bands. The higher the signal power value, the lesser it is susceptible to interference [52]. Hence, the received power measured at the 921.25 MHz frequency band has the smallest variation compared to other bands, which is 0.513 dBm. In other word, the probability of counterfeit tag which has the same received power within the limited power range is low [53]. Therefore, Method I is capable to identify counterfeit tags efficiently at the 921.25 MHz frequency band. By contrast, the highest FAR and FRR of Method I at the 920.75 MHz frequency band contributes to the lowest AUC and the highest EER. In the 920.75 MHz frequency band, the AUC and the EER of Method I are 0.435 and 0.576, respectively. Therefore, Method I has the least efficient at the 920.75 MHz frequency band because of the lowest AUC and the highest EER obtained. The result shows there is a large variability of received power of tag at the 920.75 MHz. The variation of the received power at the 920.75 MHz frequency band is 0.748 dBm. This is because the received power of tag at the 920.75 MHz frequency band has lower value compared to other frequency bands. Hence, the received power at the 920.75 MHz tends to has more interference problem [52]. The large variation of received power contributes to the high probability of counterfeit tag to be categorized as legitimate tag [54]. Hence, Method I has lower accuracy in detecting counterfeit tags at the 920.75 MHz frequency band compared to other frequency bands.

**Table 6.** AUCs and EERs of Method I

Test Variable(s)	Result Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval		EER
				Lower Bound	Upper Bound	
919.25	0.454	0.013	0.000	0.429	0.480	0.546
919.75	0.563	0.013	0.000	0.538	0.589	0.445
920.25	0.536	0.013	0.005	0.511	0.562	0.469
920.75	0.435	0.013	0.000	0.410	0.460	0.576
921.25	0.596	0.013	0.000	0.570	0.621	0.417
921.75	0.510	0.013	0.424	0.485	0.536	0.498
922.25	0.475	0.013	0.050	0.449	0.500	0.549
922.75	0.577	0.013	0.000	0.551	0.602	0.457

- a. Under the nonparametric assumption.
- b. Null hypothesis: true area = 0.5.



**Fig. 9.** ROCs of Method I

**6.2 Accuracy of Two-way ANOVA Test Analysis Based on Received Power (Method II)**

Method II is an electronic fingerprint matching method using the two-way ANOVA test to analyze the received power of tag for the 8 frequency bands at 0.3 m. From a total of 1000 counterfeit tags tested, Method II falsely accepts 270 counterfeit tags as legitimate tags based on the p-value obtained from ANOVA test. In the other meaning, Method II has successfully identified 730 counterfeit tags from a total of 1000 counterfeit tags. Besides, Method II is capable to recognize 937 legitimate tags out of a total of 1000 legitimate tags. Although the accuracy of Method II in detecting counterfeit tags is better than Method I, but, the low value of received power is vulnerable to interference [52]. Hence, FAR and FRR obtained for Method II are 27.0 % and 6.3 %, respectively, as shown in Table 7.

**Table 7.** FAR and FRR of Method II

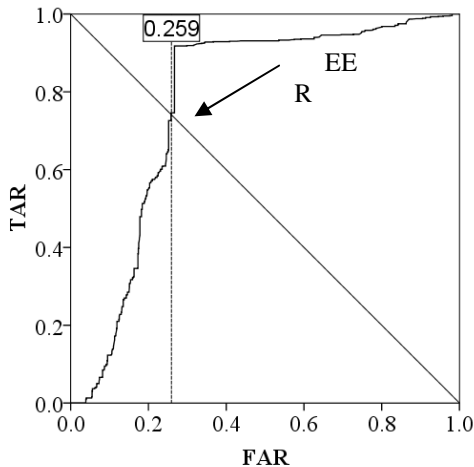
TA	FR	TR	FA	FAR (%)	FRR (%)
937	63	730	270	27.0	6.3

The AUC of Method II is 0.773 in Table 8, and the EER obtained from the plotted graph in Figure 10 is 0.259. The lower FAR and FRR of Method II compared to Method I contributes to the higher AUC and the lower EER obtained. The confidence interval of Method II is between 0.750 and 0.795. Method II has smaller confidence interval than Method I, with the difference of 0.006. Hence, Method II has higher accuracy than Method I with higher AUC, lower EER and narrower confidence interval. Method II uses the two-way ANOVA to evaluate the equality means of received power across 8 frequency bands [55]. The probability of counterfeit tag to have the similar received power at the 8 frequency bands is low. By contrast, Method I analyzes received power of tag at single frequency band using t-test [56]. In short, Method II is capable to distinguish counterfeit from legitimate tags more accurately compared to Method I.

**Table 8.** AUC of Method II

Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.773	0.011	0.000	0.750	0.795

- a. Under the nonparametric assumption.
- b. Null hypothesis: true area = 0.5.



**Fig. 10.** ROC with EER of Method II

**6.3 Accuracy of T-Test Analysis Based on Backscatter Power (Method III)**

The electronic fingerprint matching method using the t-test to evaluate the backscatter power of tag for each of the 8 frequency bands at 0.3 m is categorized as Method III. Table 9 indicates that the FARs and FRRs are in the range of 16.3 % to 30.5 % and 2.1 % to 11.4 %, respectively. Method III has the highest FA and FR at the 920.25 MHz frequency band. The result obtained is because of the backscatter power at the 920.25 MHz frequency band has the largest variation compared to other bands, which is 0.347 dBm. The large variation is caused by the low backscatter power at the 920.25 MHz which is susceptible to interference [52]. Hence, Method III tends to falsely identify counterfeit tag because of the largest variation of backscatter power at the 920.25 MHz frequency band [54]. Method III falsely verifies 305 counterfeit tags as legitimate tags and incorrectly verifies 114 legitimate tags as counterfeit tags. In other words, Method III successfully validates 886 tags out of a total of 1000 legitimate tags and rejects 695 tags out of a total of 1000 counterfeit tags. Hence, the highest FAR and FRR of Method III are found at the 920.25 MHz frequency band because of its highest FA and FR obtained. By contrast, Method III has the lowest FA and FR at the 922.25 MHz frequency band. Method III successfully identifies 979 tags out of a total of 1000 legitimates tags. In addition, Method III is capable to truly reject 837 tags out of a total of 1000 counterfeit tags. The lowest FA and FR contribute to the lowest FAR and FRR of Method III at the 922.25 MHz frequency band, with 16.3 % and 2.1%, respectively. This is because the backscatter power of tag at the 922.25 MHz band has the lowest variation, which is 0.296 dBm. The small variation is caused by the high backscatter power at the 922.25 MHz is less susceptible to interference [52]. Hence, Method III can efficiently identify counterfeit tag because counterfeit tag has low probability to possess similar backscatter power as legitimate tag [57].

**Table 9.** FARs and FRRs of Method III

Frequency (MHz)	TA	FR	TR	FA	FAR (%)	FRR (%)
919.25	935	65	793	207	20.7	6.5
919.75	925	75	759	241	24.1	7.4
920.25	886	114	695	305	30.5	11.4
920.75	941	59	774	226	22.6	5.9
921.25	921	79	742	258	25.8	7.9
921.75	954	47	811	189	18.9	4.7
922.25	979	21	837	163	16.3	2.1
922.75	909	91	726	274	27.4	9.1

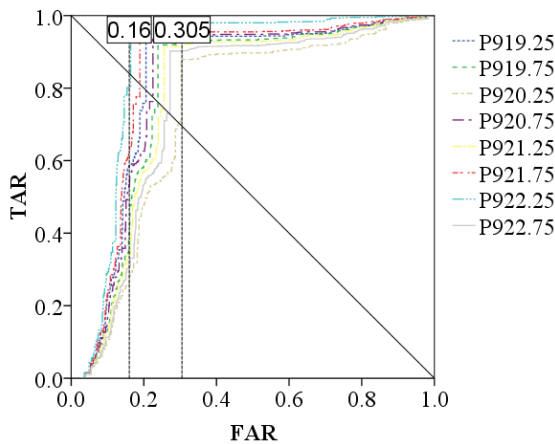
The AUCs of Method III are in the range of 0.732 to 0.868 as shown in Table 10, and the EERs values are in the range of 0.160 to 0.305 as indicated in Figure 11. The highest AUC, with the value of 0.868, is found at the 922.25 MHz frequency band because of its lowest FAR and FRR obtained. The highest AUC of Method III at 922.25 MHz contributes to the lowest EER among the 8 frequency bands, with the

value of 0.160. The lowest AUC and the highest EER of Method III is at 920.25 MHz frequency band. The difference between the highest and lowest of the AUCs and EERs of Method III is 0.136 and 0.145, respectively. In addition, Method III has the widest confidence interval at the 920.25 MHz, with 0.08 higher than the confidence interval obtained at the 922.25 MHz frequency band. The large variation of backscatter powers obtained at the 920.25 MHz contributes to the wide confidence interval [58]. The backscatter power of counterfeit tag tends to has similar value as the legitimate tag because of the wide confidence interval [57]. Therefore, Method III provides low accuracy in detecting counterfeit tag at the 920.25 MHz. The highest FAR and FRR obtained at the 920.25 MHz causes Method III to be the least efficient in distinguish counterfeit from legitimate tags.

**Table 10.** AUCs and EERs of Method III

Test Result Variable(s)	Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval		EER
				Lower Bound	Upper Bound	
919.25	0.817	0.011	0.000	0.796	0.838	0.209
919.75	0.791	0.011	0.000	0.769	0.813	0.238
920.25	0.732	0.012	0.000	0.709	0.755	0.305
920.75	0.810	0.011	0.000	0.788	0.831	0.227
921.25	0.779	0.011	0.000	0.757	0.801	0.246
921.75	0.834	0.010	0.000	0.814	0.855	0.189
922.25	0.868	0.010	0.000	0.849	0.887	0.160
922.75	0.761	0.012	0.000	0.738	0.784	0.269

- a. Under the nonparametric assumption.
- b. Null hypothesis: true area = 0.5.



**Fig. 11.** ROCs of Method III



**6.4 Accuracy of Two-Way ANOVA Test Analysis Based on Backscatter Power (Method IV)**

Method IV is a fingerprint matching method which uses the two-way ANOVA test to evaluate the backscatter power of 8 frequency bands at fixed positions of 0.1 m, 0.2 m and 0.3 m. The FARs obtained are 2.7 %, 3.5 %, and 4.2 %, respectively, for the 3 positions. In addition, the FRRs obtained are 8.0 %, 11.0 %, and 12.0 % at positions of 0.1 m, 0.2 m, and 0.3 m, respectively, as shown in Table 11. Method IV has the highest FAR and FRR at 0.3 m, with the values of 4.2 % and 12.0 %, respectively. In other words, Method IV is capable to identify 880 tags out of a total of 1000 legitimate tags. In addition, Method IV has successfully reject 958 tags from 1000 counterfeit tags, and falsely identifies 42 tags as legitimate tags. By contrast, Method IV produces 2.7 % FAR and 8.0 % FRR at the position of 0.1 m, which are the lowest FAR and FRR among the three positions. Based on the FAR and the FRR values at the position of 0.1 m, Method IV has successfully verifies 40 more legitimate tags and rejects 15 more counterfeit tags compared to the result obtained at the position of 0.3 m. The backscatter power of tag measured at the position of 0.1 m has the highest value, follows by the backscatter power measured at the position of 0.2 m and 0.3 m according to the data obtained. The higher the value of backscatter power of tag, the lesser it susceptible to interference [52]. Hence, the backscatter power of tag measured at 0.1 m has lowest variation, which is 0.231 dBm.

**Table 11.** FARs and FRRs of Method IV

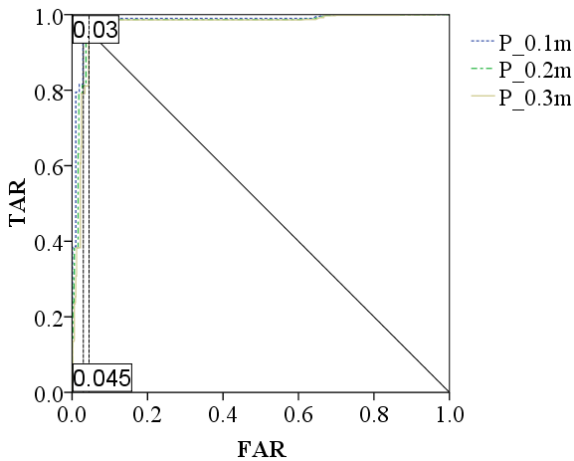
Distances (m)	TA	FR	TR	FA	FAR (%)	FRR (%)
0.1	920	80	973	27	2.7	8.0
0.2	890	110	965	35	3.5	11.0
0.3	880	120	958	42	4.2	12.0

The AUCs of Method IV are 0.984, 0.976, and 0.970 at the positions of 0.1 m, 0.2 m and 0.3 m, respectively as shown in Table 12. The EERs obtained from the plotted graph in Figure 12 are 0.030, 0.038, and 0.045, respectively, for the 3 positions. The AUC of Method IV is the highest at the position of 0.1 m and is the lowest at the position of 0.3 m. On the contrary, the EER of Method IV is the highest at the position of 0.3 m and is the lowest at the position of 0.1 m. In addition, the confidence interval of Method IV at the position 0.1 m is the narrowest, with the difference of 0.03 and 0.05 compared to the confidence interval at the position of 0.2 m and 0.3 m. The high value of backscatter power obtained at 0.1 m contributes to the smallest variation. Hence, the probability of counterfeit tag which has the backscatter power within the limited range is very low [53]. By contrast, the backscatter power at 0.3 m has the largest variation, which is 0.347 dBm. Therefore, the chance of counterfeit tag which has the similar backscatter power range at 0.3 m is high [54]. In short, Method IV is capable to distinguish counterfeit from legitimate tags more efficient at the position of 0.1 m compared to other 2 positions.

**Table 12.** AUCs and EERs of Method IV

Test Result Variable(s)	Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval		EER
				Lower Bound	Upper Bound	
0.1 m	0.984	0.003	0.000	0.978	0.990	0.030
0.2 m	0.976	0.004	0.000	0.969	0.984	0.038
0.3 m	0.970	0.004	0.000	0.962	0.979	0.045

- a. Under the nonparametric assumption.
- b. Null hypothesis: true area = 0.5.



**Fig. 12.** ROCs of Method IV

**6.5 Accuracy of Three-Way ANOVA Test Analysis Based on Backscatter Power – Proposed Method (Method V)**

The electronic fingerprint matching method using the three-way ANOVA test to analyze the backscatter power of tag for 8 frequency bands at three positions is categorized as Method V. The three positions are 0.1 m, 0.2 m, and 0.3 m, respectively. The FAR and the FRR obtained are 0.1 % and 1.3 %, respectively, as shown in Table 13. Method V is capable to reject 999 tags from a total of 1000 counterfeit tags, with only 1 tag is misidentified as legitimate tags. Furthermore, Method V has successfully accepted 987 tags out of a total 1000 legitimate tags, and has falsely rejected 13 legitimate tags as counterfeit tags. Method V uses the three-way ANOVA to analyze the equality means of backscatter power across 8 frequency bands measured at 3 positions [59]. Therefore, the probability of counterfeit tag which has the similar backscatter powers measured at 8 frequency bands and 3 different positions is very low.

**Table 13.** FAR and FRR of Method V

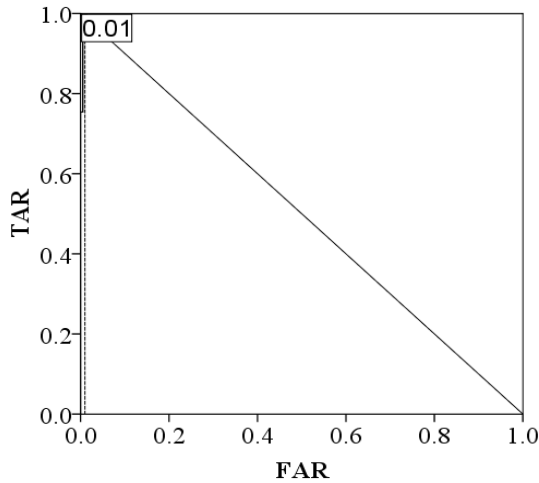
TA	FR	TR	FA	FAR (%)	FRR (%)
987	13	999	1	0.1	1.3

The AUC of Method V is 0.999 is shown in Table 14, and the EER obtained from the plotted graph of Figure 13 is 0.01. Method V has the highest AUC and the smallest EER values compared to other four methods. The confidence interval of Method V is 0, with the lower bound and the upper bound values are both 1.000. The highest AUC, the smallest EER and confidence interval prove that Method V is the most efficient method in detecting counterfeit tags. The three-way ANOVA is used to analyze backscatter power of tag in Method V. The backscatter power of tag has higher value than the received power of tag based on the data obtained. Hence, the backscatter power of tag is less susceptible to interference compared to the received power of tag [52]. In addition, the three-way ANOVA test is capable to analyze the equality mean of all the backscatter powers measured at the 8 frequency bands and 3 positions [59]. Therefore, Method V is an excellent fingerprint matching method.

**Table 14.** AUC and EER of Method V

Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
0.999	0.000	0.000	1.000	1.000

- a. Under the nonparametric assumption.
- b. Null hypothesis: true area = 0.5.



**Fig. 13.** ROC with EER of Method V

## 6.6 Comparisons between Fingerprint Matching Methods

The accuracy of Method I in distinguishing legitimate from counterfeit tags is a failure method. FAR and FRR obtained for the worse case are 60.3 % and 34.5 %, respectively. The AUC and the EER obtained for Method I are 0.435 and 0.576, respectively. Therefore, the two-way ANOVA test is used for Method II to analyze the received power of tag for all frequency bands to improve the accuracy of the fingerprint matching method. Nevertheless, the result indicates that this fingerprint matching method is a fair method based on the test configuration in Table 5.3 because the AUC and the EER obtained are 0.773 and 0.259, respectively. In addition, the FAR and the FRR of the Method II is lower, with values of 27.0 % and 6.3 %, respectively. Method II is able to provide more accurate result compared to Method I because Method II is using the two-way ANOVA to analyze the received power of tag. The two-way ANOVA is capable to test the equality of means across the 8 frequency bands [55]. But, Method I is using the t-test to analyze the received power of tag at one frequency band only [56]. The t-test is unable to analyze the received power at multiple frequency bands. Hence, the t-test has to repeat for 8 times to analyze all the received power of tag at the 8 frequency bands. But, there is a probability to get 5 % of a Type 1 error when a t-test is conducted. Type 1 error is the false rejection of a correct null hypothesis [56]. Therefore, the probability to obtain the Type 1 error will be increased to 40 % if 8 t-tests are conducted on the received power of tag at each frequency bands. By contrast, the two-way ANOVA is capable to control the Type 1 error remains at 5 %. As a result, the analysis result obtained from the two-way ANOVA test is more accurate than multiple t-tests because of lesser Type-1 error [60].

Method III performs better than the Methods I and II. This finding can be attributed to the higher AUC as well as lower EER, FAR, and FRR values of Method III. The AUC and the EER for the worse case of Method III are 0.732 and 0.305, respectively. The FAR and the FRR are in the ranges from 16.3 % to 30.5 % and 2.1 % to 11.4 %, respectively. Method III analyzes backscatter power of tag as unique fingerprint. By contrast, the Method I and II use received power of tag as the unique fingerprint in the fingerprint matching method. Based on the data obtained from this research, the value of backscatter power is lower than the received power of tag from the range of 7.0 dBm to 13.0 dBm. However, the received power of tag is calculated based on the transmitted power of reader. High transmitted power of reader tends to cause interference. Hence, Method III outperforms Method I and II because backscatter power of tag is less susceptible to interference than the received power of tag [52]. However, Method III is still considered as a fair method based on the test configuration in Table 5.3 because of its low AUC value.

Two-way ANOVA test is conducted for Method IV to analyze the backscatter powers measured at all frequency bands. The accuracy for the two-way ANOVA test in analyzing the backscatter power at multiple frequencies at the position of 0.1 m compared to 0.2 m and 0.3 m is higher than the t-test. The AUC and the EER obtained at 0.1 m are 0.984 and 0.030, respectively. Whereas, the AUC obtained at 0.2 m and 0.3 m are 0.976 and 0.970, respectively. The accuracy test of Method IV proves that

the two-way ANOVA test conducted on all backscatter powers at the position of 0.3 m has significantly improved the AUC and the EER compared to the t-test. The AUC for the two-way ANOVA test of 0.3 m is increased by 0.238 from 0.732 to 0.970, and the EER is decreased by 0.260 from 0.305 to 0.045. Method IV has shown improvement to be an excellent method. However, the FAR and the FRR for Method IV are high, with the values of 4.2 % and 12.0 % for the worse case. Method III uses the t-test and Method IV uses the two-way ANOVA to evaluate the backscatter power of tag. Method III has to conduct 8 times of the t-test to analyze all the backscatter power at 8 frequency bands. Hence, the total Type 1 error obtained is 40 % for a total of 8 t-tests conducted [56]. The two-way ANOVA test is able to maintain the Type 1 error as 5 % although the backscatter power at 8 frequency bands are being tested [60]. In addition, the two-way ANOVA test analyzes the equality of the means across all the backscatter power at 8 frequency bands [55]. By contrast, the t-test analyzes the mean of backscatter power for two tag groups at only one frequency band [56]. Hence, Method III which is using the t-test has lesser accuracy than Method IV which is using the two-way ANOVA test.

One of the objectives of this project is to introduce an excellent fingerprint matching method to detect counterfeit tags efficiently. But, Method I is a failure method based on the Table 5.3 because of the low AUC obtained. Hence, Method II and III are presented to improve the detection mechanism. But, the AUC of Method II and III are in the range of 0.70 to 0.80. According to the Table 5.3, both Method II and III are categorized as fair methods. Then, Method IV is introduced to enhance the accuracy of the fingerprint matching method. Although Method IV is proved as an excellent method with the AUC obtained is in the range of 0.90 to 1.00, but, the FAR and the FRR of Method IV are high. Therefore, Method V is presented to increase the accuracy of fingerprint matching method in detecting counterfeit tags.

Method V is using the three-way ANOVA test to analyze the backscatter power of tag at multiple positions and frequency bands. The AUC, EER, FAR, and FRR obtained are 0.999, 0.010, 0.1 %, and 1.3 %, respectively. Hence, Method V has proven to be an excellent method because of the highest AUC obtained with the smallest FAR and FRR values. This method can distinguish counterfeit tag with the slightest error compared with previous methods. Method V outperforms other methods because three-way ANOVA is used in the analysis of backscatter power of tag. The three-way ANOVA is capable to analyze the equality of the means across all the backscatter power at 8 frequency bands measured at 3 positions [59]. In addition, the three-way ANOVA able to maintain the probability of Type 1 error at 5 % by analyzing three factors, including frequency bands, positions and tag groups [61]. By contrast, the two-way ANOVA used in Method II and IV analyze the equality of means across two factors only, namely, frequency bands and tag groups [55]. In addition, Method I and III use the t-test in the fingerprint matching method. The t-test offers least accuracy compared to other test because the t-test compares the means of the power response of tag (received or backscatter power) at only one frequency band [56]. Table 15 shows the comparisons between the fingerprint matching methods.

**Table 15.** Comparison between fingerprint matching methods

Method	Power Response of Tag	Distance (m)	Statistical Test	FAR (%)	FRR (%)	AUC	EER
I	Received Power	0.3	T-test	42.7	21.8-	0.435	0.417
				-	34.5	-	-
				60.3		0.596	0.576
II	Received Power	0.3	Two-way ANOVA	27.0	6.3	0.773	0.259
III	Backscatter Power	0.3	T-test	16.3	2.1 -	0.732	0.160
				-	11.4	-	-
				30.5		0.868	0.305
IV	Backscatter Power	0.1	Two-way ANOVA	2.7	8.0	0.984	0.030
		0.2		3.5	11.0	0.976	0.038
		0.3		4.2	12.0	0.970	0.045
V	Backscatter Power	0.1, 0.2, 0.3	Three-way ANOVA	0.1	1.3	0.999	0.010

### 6.7 Comparison between Conventional Electronic Fingerprint Matching Method and the Proposed Method

Method V is used to distinguish legitimate from counterfeit tags because of its highest accuracy instead of the other four methods. The proposed Method V is compared with the method suggested by [31] as shown in Table 16. Method V is analyzed using three-way ANOVA because it has three factors, which are multiple frequency bands, tag positions, and tag groups. On the contrary, two factors, namely, multiple frequencies and tag groups are tested by using two-way ANOVA in [31]. The accuracy of the proposed method and method by [31] are both excellent, with the same value of 0.999. Although the FAR of Manufacturer 1 has the same value of the proposed method, but, the proposed method offers lower FAR value compared to the FAR of Manufacturer 2. The FAR of the proposed method is 0.1 % lesser than Periaswamy *et al.* method. The results indicate the proposed method has falsely accepted 1 legitimate tag from a total of 1000 counterfeit tags. But, the Periaswamy *et al.* method has falsely accepted 2 tags as Manufacturer 2 from a total of 1000 counterfeit tags. In addition, the proposed method outperforms the method by [31] with a lower FRR value, which is 4.3 % lesser than FRR of Manufacturer 1 and 8.0 % lesser than FRR of Manufacturer 2. The lesser the FRR, the more significance the security improvements of an RFID system in cloning attack. Based on the proposed method and Periaswamy *et al.* method, the accuracy of detecting counterfeit tags is increasing with lesser FRR value. This is because the proposed method is capable to falsely reject less 43 legitimate tags from a total of Manufacturer 1 tags. In addition, the proposed method is able to falsely reject less 80 legitimate tags as counterfeit tags from a total 1000 Manufacturer 2 tags. Hence, the accuracy of the proposed method in detecting counterfeit tags is higher than the method of Periaswamy *et al.*

**Table 16.** Comparison between Method V and method by [31]

	<b>Proposed Fingerprint Matching Method</b>	<b>Fingerprinting Tags [31]</b>	<b>RFID</b>
Statistical Test	Three-way ANOVA Test	Two-way ANOVA	
Physical characteristic of tag	Backscatter power of tag	Minimum power response of tag	
Independent variables	<ul style="list-style-type: none"> <li>• Multiple frequency bands</li> <li>• Multiple positions</li> <li>• Difference tag groups</li> </ul>	<ul style="list-style-type: none"> <li>• Multiple frequency bands</li> <li>• Difference tag groups</li> </ul>	
AUC	0.999	Manufacturer 1: 0.999 Manufacturer 2: 0.997	
EER	0.01	-	
FAR (%)	0.1	Manufacturer 1: 0.1 Manufacturer 2: 0.2	
FRR (%)	1.3	Manufacturer 1: 5.6 Manufacturer 2: 9.3	

### 6.8 Future Works

A few recommendations for future research are suggested due to their significance and importance related to the present study.

- i. It is recommended to further study on the fingerprint matching method in the present study by using other statistical methods. The measurement of the received and backscatter powers of tag should be conducted in an anechoic chamber to obtain a more accurate data.
- ii. It is recommended to explore other physical characteristics of tag can be used as unique electronic fingerprint. It is suggested that a further study on microstructure of tag and minimum power needed to initiate each tag. The fingerprint matching method can be improved by using multiple characteristics of tag as unique electronic fingerprint.

## 7 Conclusion

In this chapter, a detection mechanism is proposed as a digital forensic practice to identify the counterfeit RFID tags. Electronic fingerprint matching method is used as a detection mechanism in detecting counterfeit tags. The electronic fingerprint matching method is presented in the digital forensic investigation model, which consists of seven phases. In identification phase, the power responses of tag, including received and backscatter powers, are proposed to be used as unique electronic fingerprint in the fingerprint matching method. The power response of each tag is measured, examined, and stored in the database for further reference in the collection, examination, and preservation phases. Two statistical tests, namely, t-test and

ANOVA test, are used in analyzing the data obtained in the analysis phase. In presentation phase, the accuracy of the statistical algorithm in identifying a counterfeit tag is verified by analyzing FAR, FRR, AUC, and ERR. A measured tag is proven as counterfeit in the decision phase if the p-value obtained is less than the significance level (0.05).

The fingerprint matching is categorized into five methods based on the types of power response of tag, which measures at different tag positions and different statistical tests used. The accuracy of t-test and ANOVA test for Method I and II which use the received power of tag measured at 0.3 m as a unique electronic fingerprint are a failure and fair methods. The fingerprint matching method that uses the backscatter power of tag as a unique electronic fingerprint is well performed than that using the received power of tag. This finding can be attributed to the higher AUC and lower EER values of the former method. However, the accuracy of t-test and ANOVA test for Method III and IV that use the measured backscatter power of tag at 0.3 m as a unique electronic fingerprint are fair and excellent methods respectively. In addition, the FARs and FRRs for the both methods are high, with 30.5 % and 11.4 % for Method III and 4.2 % and 12.0 % for Method IV. Method V which uses three way ANOVA test analyze backscatter power of tag measured at multiple positions and frequencies has the most accurate results. Method V which is the proposed method has the highest AUC (0.999) and lowest EER (0.01) values. The FAR and FRR obtained are 0.1 % and 1.3 %, respectively. Based on the FAR and FRR values, the proposed method is falsely accepted 1 tag from a total of 1000 counterfeit tags and is falsely rejected 13 tags from a total of 1000 legitimate tags. Hence, this method can distinguish a counterfeit tag with the slightest error compared with the other methods.

The proposed fingerprint matching method is a simple detection mechanism because it can be applied directly to any existing tag without any modification to its computational capabilities. In addition, digital forensic investigation model is shown in this chapter to describe the overall counterfeit tag detection practice in detail. As a result, the simple and detailed proposed fingerprint matching method can be used in digital forensic detection to identify counterfeit RFID tag efficiently.

**Acknowledgement.** The authors would like to thank the USM Research University & PRGS grant secretariat for sponsoring this research. In addition, appreciation also to the Malaysia Ministry of Higher Education (LRGS fund) for financially supporting the development of the in-house built EPC Class 1 Gen 2 UHF RFID devices.

## References

1. Hagl, A., Aslanidis, K.: RFID: Fundamentals and Applications. In: Kitsos, P., Zhang, Y. (eds.) RFID Security, pp. 3–26. Springer US (2009)
2. Piramuthu, S.: Lightweight Cryptographic Authentication in Passive RFID-Tagged Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 38(3), 360–376 (2008)
3. Tuyls, P., Batina, L.: RFID-Tags for Anti-Counterfeiting. In: Pointcheval, D. (ed.) CT-RSA 2006. LNCS, vol. 3860, pp. 115–131. Springer, Heidelberg (2006)



4. King, B., Zhang, X.: RFID: An Anticounterfeiting Tool. In: Kitsos, P., Zhang, Y. (eds.) RFID Security, pp. 27–55. Springer US (2009)
5. Li, T., Lim, T.-L.: RFID Anticounterfeiting: An Architectural Perspective. In: Kitsos, P., Zhang, Y. (eds.) RFID Security, pp. 131–146. Springer US (2009)
6. Staake, T., Thiesse, F., Fleisch, E.: Extending the EPC Network: The Potential of RFID in Anti-Counterfeiting. In: Proceedings of the 2005 ACM Symposium on Applied Computing, Santa Fe, New Mexico, pp. 1607–1612 (2005)
7. MarketsandMarkets: Global Anti Counterfeit Packaging Market Food and Pharmaceuticals (2010)
8. RNCOS: Research Report on Global RFID Market Analysis Till 2010 (2010)
9. Henrici, D.: RFID Security and Privacy. Concepts, Protocols, and Architectures. LNEE, vol. 17. Springer, Heidelberg (2008)
10. Sharif, A., Potdar, V.: A Critical Analysis of RFID Security Protocols. In: Proceedings of the 22nd International Conference on Advanced Information Networking and Applications - Workshops, GinoWan, Okinawa, Japan, pp. 1357–1362 (2008)
11. Stadlober, S.: An Evaluation of Security Threats and Countermeasures in Distributed RFID Infrastructures. Graz University of Technology, Graz (2005)
12. Martínez-Sala, A.S., Egea-López, E., García-Sánchez, F., García-Haro, J.: Tracking of Returnable Packaging and Transport Units with Active RFID in The Grocery Supply Chain. Computers in Industry 60, 161–171 (2009)
13. Razaq, A., Luk, W.T., Shum, K.M., Cheng, L.M., Yung, K.N.: Second-Generation RFID. IEEE Security & Privacy, 21–22 (2008)
14. Juels, A.: Strengthening EPC Tags Against Cloning. In: Proceedings of the 4th ACM Workshop on Wireless Security, pp. 67–76. ACM, Cologne (2005)
15. Rieback, M.R., Crispo, B., Tanenbaum, A.S.: The Evolution of RFID Security. IEEE Pervasive Computing 5(1), 62–69 (2006)
16. Berbain, C., Billet, O., Etrog, J., Gilbert, H.: An Efficient Forward Private RFID Protocol. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, Illinois, USA (2009)
17. Mitrokotsa, A., Rieback, M.R., Tanenbaum, A.S.: Classification of RFID Attacks. In: Proceedings of the 2nd International Workshop on RFID Technology - Concepts, Applications, Challenges, Barcelona, Spain, pp. 73–86 (2008)
18. Wong, K.H.M., Hui, P.C.L., Chan, A.C.K.: Cryptography and Authentication on RFID Passive Tags for Apparel Products. Computers in Industry 57(4), 342–349 (2006)
19. Mirowski, L., Hartnett, J., Williams, R.: An RFID Attacker Behavior Taxonomy. IEEE Pervasive Computing 8(4), 79–84 (2009)
20. Lehtonen, M.O., Michahelles, F., Fleisch, E.: Trust and Security in RFID-Based Product Authentication Systems. IEEE Systems Journal 1(2), 129–144 (2007)
21. Koscher, K., Juels, A., Brajkovic, V., Kohno, T.: EPC RFID Tag Security Weaknesses and Defenses: Passport Cards, Enhanced Drivers Licenses, and Beyond. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, Chicago, Illinois, USA, pp. 33–42 (2009)
22. Juels, A., Molnar, D., Wagner, D.: Security Issues in E-Passports. In: Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communication Networks, Athens, Greece, pp. 74–88 (2005)
23. Khor, J.H., Ismail, W., Younis, M.I., Sulaiman, M.K., Rahman, M.G.: Security Problems in an RFID System. Wireless Personal Communications 59(1), 17–26 (2010)
24. Kamoun, F.: RFID System Management: State-of-the Art and Open Research Issues. IEEE Transactions on Network and Service Management 6(3), 190–205 (2009)

25. Bailey, D.V., Juels, A.: Shoehorning Security into the EPC Tag Standard. In: De Prisco, R., Yung, M. (eds.) SCN 2006. LNCS, vol. 4116, pp. 303–320. Springer, Heidelberg (2006)
26. Zuo, Y.: Survivable RFID Systems: Issues, Challenges, and Techniques. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 40(4), 406–418 (2010)
27. Ghayal, A., Khan, Z., Moona, R.: SmartRF - A Flexible and Light-Weight RFID Middleware. In: *Proceedings of the IEEE International Conference on e-Business Engineering*, Xi'an, China, pp. 317–324 (2008)
28. Lin, F., Chen, B., Chan, C.Y., Wu, C.H., Ip, W.H., Mai, A., Wang, H., Liu, W.: The Design of a Lightweight RFID Middleware. *International Journal of Engineering Business Management* 1(2), 25–30 (2009)
29. Li, T.Y., Lim, T.L.: RFID Anticounterfeiting: An Architectural Perspective. In: Kitsos, P., Zhang, Y. (eds.) *RFID Security*, pp. 131–146. Springer US (2009)
30. Zanetti, D., Danev, B., Capkun, S.: Physical-Layer Identification of UHF RFID Tags. In: *Proceedings of the Sixteenth Annual International Conference on Mobile Computing and Networking*, Chicago, Illinois, USA, pp. 353–364 (2010)
31. Periaswamy, S.C.G., Thompson, D.R., Jia, D.: Fingerprinting RFID Tags. *IEEE Transactions on Dependable and Secure Computing* 8(6), 938–943 (2011)
32. Van Staden, F.R., Venter, H.S.: Adding Digital Forensic Readiness to The Email Trace Header. In: *Information Security for South Africa (ISSA)*, pp. 1–4 (2010)
33. Sammons, J.: *The Basics of Digital Forensics: The Primer for Getting Started in Digital Forensics*. Elsevier (2012)
34. Palmer, G.: A Road Map for Digital Forensic Research. In: *The First Digital Forensic Research Workshop*, Utica, New York (2001)
35. Khor, J.H., Ismail, W., Rahman, M.G.: Prevention and Detection Methods for Enhancing Security in an RFID System. *International Journal of Distributed Sensor Networks* 2012, 8 (2012)
36. Chang, K.: *RF and Microwave Wireless Systems*. Wiley-Interscience (2000)
37. Du, K.-L., Swamy, M.N.S.: *Wireless Communication Systems: From RF Subsystems to 4G Enabling Technologies*. Cambridge University Press (2010)
38. Lu, D., So, D.K.C., Brown, A.K.: Receive Antenna Selection Scheme for V-BLAST with Mutual Coupling in Correlated Channels. In: *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, Cannes, France, pp. 1970–1974 (2008)
39. Lim, S., Ling, H.: Design of Electrically Small Yagi Antenna. *Electronics Letters* 43(5), 256–258 (2007)
40. Thomas, J.R., Nelson, J.K., Silverman, S., Silverman, S.J.: *Research Methods in Physical Activity*. Human Kinetics (2010)
41. Montgomery, D.: *Design and Analysis of Experiments*. John Wiley and Sons (2001)
42. Venkatesan, S.: *Investigation of RFID Readability for License Plates in Static and Motion Testing*. University of Nebraska, Lincoln (2011)
43. Tamhane, A.C.: *Statistical Analysis of Designed Experiments: Theory and Applications*. Wiley-Interscience (2009)
44. Jian, Z., Shirai, H., Takahashi, I., Kuroiwa, J., Odaka, T., Ogura, H.: A Hybrid Command Sequence Model for Anomaly Detection. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) *PAKDD 2007*. LNCS (LNAI), vol. 4426, pp. 108–118. Springer, Heidelberg (2007)

45. Tronci, R., Giacinto, G., Roli, F.: Dynamic Score Combination: A Supervised and Unsupervised Score Combination Method. In: Perner, P. (ed.) MLDM 2009. LNCS, vol. 5632, pp. 163–177. Springer, Heidelberg (2009)
46. Wu, J.C.: Operational Measures and Accuracies of ROC Curve on Large Fingerprint Data Sets. *National Institute of Standards and Technology* 1(1), 1–23 (2008)
47. Ramachandra, A.C., Pavithra, K., Yashasyini, K., Raja, K.B., Venugopal, K.R., Patnaik, L.M.: Cross-Validation for Graph Matching Based Offline Signature Verification. In: Annual IEEE India Conference, Kanpur, India, pp. 17–22 (2008)
48. Kutlu, Y., Kuntalp, D.: A Multi-Stage Automatic Arrhythmia Recognition and Classification System. *Journal of Computers in Biology and Medicine* 41(1), 37–45 (2011)
49. Chevillotte, C.J., Ali, M.H., Trousdale, R.T., Larson, D.R., Gullerud, R.E., Berry, D.J.: Inflammatory Laboratory Markers in Periprosthetic Hip Fractures. *Journal of Arthroplasty* 24(5), 722–727 (2009)
50. Lehmann, E.L., Romano, J.P.: *Testing Statistical Hypotheses*. Springer, New York (2005)
51. Taneja, S., Akcamete, A., Akinci, B., Garrett, J., Soibelman, L., East, E.: Analysis of Three Indoor Localization Technologies for Supporting Operations and Maintenance Field Tasks. *Journal of Computing in Civil Engineering* 26(6), 708–719 (2012)
52. Bockstiegel, K.H., Benko, M.: *Space Law: Basic Legal Documents*. Eleven International Publishing (1990)
53. Hayes, A.F.: *Statistical Methods for Communication Science*. Taylor & Francis (2012)
54. Wooldridge, J.M.: *Introductory Econometrics: A Modern Approach*. South Western, Cengage Learning (2009)
55. Bradley, T.: *Essential Statistics for Economics, Business and Management*. John Wiley & Sons (2007)
56. Rumsey, D.: *Intermediate Statistics For Dummies*. Wiley (2007)
57. Scutchfield, F.D., Keck, C.W.: *Principles of Public Health Practice*. Thomson/Delmar Learning (2003)
58. Sinclair, A., Nantel, P., Catling, P.: Dynamics of Threatened Goldenseal Populations and Implications for Recovery. *Biological Conservation* 123(3), 355–360 (2005)
59. Jackson, S.L.: *Statistics: Plain and Simple*. Cengage Learning (2009)
60. Creighton, T.B.: *Schools and Data: The Educator's Guide for Using Data to Improve Decision Making*. SAGE Publications (2006)
61. Lee, I.: *Mobile Applications and Knowledge Advancements in E-Business*. IGI Global (2012)

### Appendix A

**Independent Samples Test**

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Power Equal variances assumed	.443	.509	1.005	58	.319	.06496	.06465	-.06446	.19438
Equal variances not assumed			.863	11.401	.406	.06496	.07531	-.10008	.23000

### Appendix B

**Tests of Between-Subjects Effects**

Dependent Variable: Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.529 <sup>a</sup>	15	.502	14.782	.000
Intercept	1.812	1	1.812	53.369	.000
Tag	.004	1	.004	.112	.738
Frequency	5.133	7	.733	21.594	.000
Tag * Frequency	.218	7	.031	.917	.492
Error	15.756	464	.034		
Total	26.350	480			
Corrected Total	23.285	479			

a. R Squared = .323 (Adjusted R Squared = .301).

### Appendix C

**Independent Samples Test**

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Power Equal variances assumed	.350	.556	-.335	58	.739	-.01552	.04639	-.10839	.07735
Power Equal variances not assumed			-.309	12.026	.762	-.01552	.05019	-.12484	.09380

### Appendix D

**Tests of Between-Subjects Effects**

Dependent Variable:Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.465 <sup>a</sup>	15	4.764	92.617	.000
Intercept	25598.320	1	25598.320	4.976E5	.000
Tag	6.667E-7	1	6.667E-7	.000	.997
Frequency	40.691	7	5.813	113.002	.000
Tag * Frequency	.401	7	.057	1.113	.354
Error	23.869	464	.051		
Total	46171.997	480			
Corrected Total	95.334	479			

a. R Squared = .750 (Adjusted R Squared = .742).

**Appendix E****Tests of Between-Subjects Effects**

Dependent Variable:Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	13756.152 <sup>a</sup>	47	292.684	6.764E3	.000
Intercept	93348.803	1	93348.803	2.157E6	.000
Tag	.006	1	.006	.142	.706
Distance	7110.646	2	3555.323	8.216E4	.000
Frequency	64.276	7	9.182	212.188	.000
Tag * Distance	.062	2	.031	.721	.486
Tag * Frequency	.081	7	.012	.267	.967
Distance * Frequency	437.311	14	31.237	721.831	.000
Tag * Distance * Frequency	.522	14	.037	.862	.601
Error	60.237	1392	.043		
Total	181901.802	1440			
Corrected Total	13816.390	1439			

a. R Squared = .996 (Adjusted R Squared = .995).

## Appendix F

**Independent Samples Test**

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Power Equal variances assumed	5.813	.019	4.556	58	.000	.26586	.05836	.14905	.38267
Equal variances not assumed			6.840	24.636	.000	.26586	.03887	.18575	.34597

## Appendix G

**Tests of Between-Subjects Effects**

Dependent Variable: Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.685 <sup>a</sup>	15	.512	15.212	.000
Intercept	5.821	1	5.821	172.817	.000
Tag	1.273	1	1.273	37.790	.000
Frequency	3.031	7	.433	12.854	.000
Tag * Frequency	.466	7	.067	1.975	.037
Error	15.628	464	.034		
Total	28.277	480			
Corrected Total	23.314	479			

a. R Squared = .330 (Adjusted R Squared = .308).

## Appendix H

**Independent Samples Test**

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Power	Equal variances assumed	.120	.730	-4.180	58	.000	-.19232	.04601	-28441	-10023
	Equal variances not assumed			-3.988	12.338	.002	-.19232	.04822	-29707	-.08757

## Appendix I

**Tests of Between-Subjects Effects**

Dependent Variable: Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	66.367 <sup>a</sup>	15	4.424	95.399	.000
Intercept	25752.833	1	25752.833	5.553E5	.000
Tag	.233	1	.233	5.029	.025
Frequency	31.709	7	4.530	97.670	.000
Tag * Frequency	.497	7	.071	1.532	.015
Error	21.520	464	.046		
Total	46257.164	480			
Corrected Total	87.887	479			

a. R Squared = .755 (Adjusted R Squared = .747).



## Appendix J

### Tests of Between-Subjects Effects

Dependent Variable: Power

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	13550.596 <sup>a</sup>	47	288.311	6.511E3	.000
Intercept	92050.826	1	92050.826	2.079E6	.000
Tag	4.884	1	4.884	110.307	.000
Distance	6780.389	2	3390.195	7.656E4	.000
Frequency	63.739	7	9.106	205.634	.000
Tag * Distance	7.832	2	3.916	88.439	.000
Tag * Frequency	.183	7	.026	.591	.764
Distance * Frequency	408.765	14	29.197	659.379	.000
Tag * Distance * Frequency	1.266	14	.090	2.041	.053
Error	61.638	1392	.044		
Total	180916.909	1440			
Corrected Total	13612.234	1439			

a. R Squared = .995 (Adjusted R Squared = .995).

# Learning Remote Computer Fingerprinting

João P. Souza Medeiros<sup>1</sup>, João B. Borges Neto<sup>1</sup>,  
Agostinho M. Brito Júnior<sup>2</sup>, and Paulo S. Motta Pires<sup>2</sup>

<sup>1</sup> Elements of Information Processing Laboratory (LabEPI),  
Department of Exact and Applied Sciences (DCEA),  
Federal University of Rio Grande do Norte (UFRN), Caicó, RN, Brazil  
{joaomedeiros, joaoborges}@ufrnet.br

<sup>2</sup> Security Information Laboratory (LabSIN),  
Department of Computer Engineering and Automation (DCA),  
Federal University of Rio Grande do Norte (UFRN), Natal, RN, Brazil  
{ambj, pmotta}@dca.ufrn.br

**Abstract.** The process of remote characterization and identification of computers has many applications in network security and forensics. On network forensics, this process can be used together with intrusion detection systems to characterize suspicious machines of remote attackers. The characterization of remote computers is based on the analysis of network data originated from the remote machine. The classical approach is to exploit peculiar characteristics of different implementations of network protocols at each layer of the protocol stack, i.e. link, network, transport and application layers. Recent works show that the use of computational intelligence techniques can improve the identification performance when compared to classical classification algorithms and tools. This chapter presents some advances in this area and surveys the use of computational intelligence for remote identification of computers and its applications to network forensics.

**Keywords:** Network Stack Fingerprinting, Intelligent Detection System, Remote Computer Fingerprinting.

## 1 Introduction

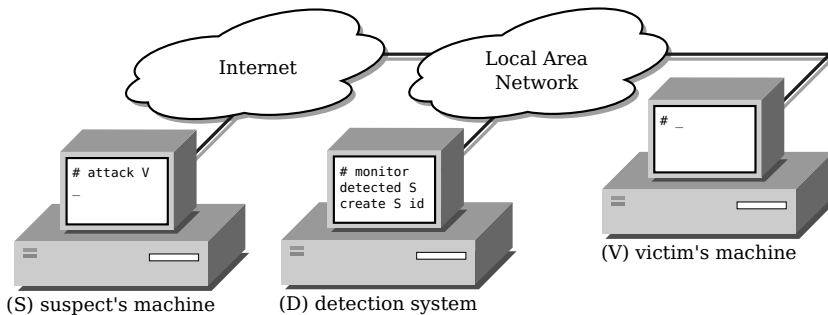
In computer forensics, the identification of machines that are used to perform illegal activities is an important step that can help to track criminals. This identification can be used in network forensics investigations to create an evidence of the use of a specific computer device in a cyber crime. In this case, the identification system shall create a fingerprint of a remote networked machine used by the suspect and, when an equipment is seized, the computer expert will be able to check the previous fingerprint with the one generated by the seized equipment. This application was exemplified by [56, 55] to capture digital evidences of criminal activity in chat rooms and web sites. This specific identification process is known in literature as Remote Computer Fingerprinting (RCF).

**Definition 1 (Remote Computer Fingerprinting).** The process of feature extraction from network traffic originated by a remote computer to create a fingerprint which enables its future identification.

The network forensics application of Remote Computer Fingerprinting is on the automatic creation of fingerprint evidences triggered by a system which detects suspicious network activity. For illustration, consider the general scenario presented in Fig. 1, where a detection system senses a Local Area Network (LAN) to:

1. identify suspicious activity from the analysis of traffic which has as destination devices belonging to the local area network; and
2. create a fingerprint of the remote suspect's machine to be used as evidence in eventual criminal investigations.

As concluded by [22], autonomous systems that are able to detect and present outliers, as well as other elements that seem out-of-place, are of fundamental importance to cope the challenges of digital forensics. For network forensics, this conceptual detection system can be implemented as some integration of two network security tools, namely, an Intrusion Detection System (IDS) to detect suspicious activity, and a RCF system to create evidences.



**Fig. 1.** A conceptual network forensics system which detects, acquires and preserves evidences of suspicious network activity in a local area network

The resulting evidence consists of a series of descriptors extracted from network data related to traffic originated from the suspect's machine. This traffic can be used to characterize remote services, systems and devices of suspect's machine. This diversity of characteristics can be used to more accurately discriminate the RCF according to the characteristic being considered. Based on the concepts presented on [5], we can identify three kinds of remote computer fingerprinting, according to groups of protocols that are being analyzed:

1. **Remote Service Fingerprinting (RSF):** for techniques which use data from network traffic associated to service application on the remote machine;
2. **Remote Network Stack Fingerprinting (RNSF):** for techniques which deal with data from transport and internetwork layers;
3. **Remote Link Fingerprinting (RLF):** for techniques focused on data associated to the link layer.

All these three types of remote fingerprinting are well explained further in this chapter. It is important to notice that the definition of RNSF is widely spread in computer security literature as Remote Operating System Fingerprinting (ROSF), since these layers are implemented on most operating systems. However, we use the term Remote Network Stack Fingerprinting because it is more appropriated and it avoids ambiguity with the other kinds of RCF.

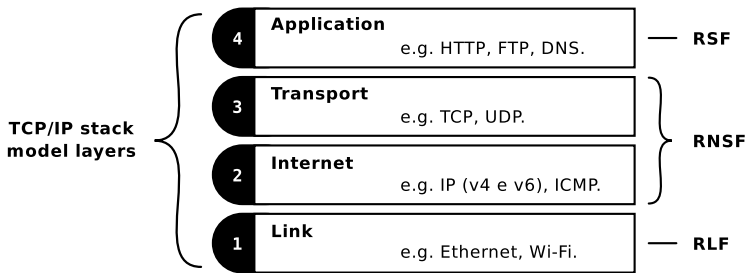
In this text we focus on recent advances of computational intelligence in the context of RNSF. The exclusion of RSF and RLF is due to the lack of data availability in most common scenarios, a fact which is detailed in Sect. 2. Since network forensics needs sophisticated tools to acquire, preserve, examine, analyze and present digital evidence [57], understanding the techniques which makes this identification possible is of great importance.

We introduce the reader to a better understanding of the current techniques that are used by forensic field experts. Some fundamentals on key computer network concepts and its relation to the process of remote identification are also presented. Following, we present the several techniques that are used to perform RCF.

**Organization of the Chapter.** Basic concepts and other prerequisites necessary to the understanding of RCF are detailed in Sect. 2. Sect. 3 presents the state of the art on the use of computational intelligence to RNSF. Also, the applicability and perspectives about the use of computational intelligent techniques in RCF are presented in Sect. 3. No technique or tool can be successful employed in the context of forensics without considering the legal implications. Therefore, in Sect. 4 we consider the admissibility of RCF evidences into a court of law. Finally, in Sect. 5 we discuss the current achievements and future perspectives which remote computer fingerprinting aided by computational intelligence brings to network forensics.

## 2 Remote Computer Fingerprinting Fundamentals

Unlike human fingerprint and deoxyribonucleic acid (DNA) sequences analysis, which is regarded to unique identify human beings, a remote computer fingerprint cannot guarantee an unique identification for a given computer device. However, RCF can aid cyber crime investigations supporting identity for suspicious network activity evidences. As pointed out in Sect. 1, there are different ways to characterize remote machines in a network. More specifically, it is possible to identify hardware or software components of a computer device according to the network model layer to which the acquired data is associated. To make it clear, we use the terminology of computer network literature, which conveniently conceptualize network protocols functions into a protocol stack model. On the Internet, the *de facto* standard is the use of IP, the Internet Protocol [59]. Devices uses IP to carry data from a source to a destination and TCP, the Transmission Control Protocol [61], to provide transport service for applications that requires end-to-end reliability, re-sequencing, and flow control. These are the main protocols that characterize the TCP/IP stack model. This model is currently used worldwide to deliver data among machines and its protocols are classified according to their functions into four layers: Application, Transport, Internetwork and Link



**Fig. 2.** The protocol stack model which should be adopted by TCP/IP networked devices according to [5]. At layers (2) and (3), and occasionally on layer (4) by indirect ways, it is possible to determine the remote operating system. At layer (1) it is possible to infer about network interface cards. The data from each layer could be used to perform RLF (layer 1), RNSF (layers 2 and 3) and RSF (layer 4).

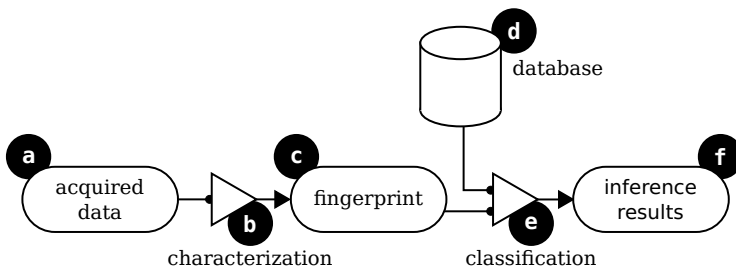
layer [5]. Fig. 2 depicts this stack model with examples of network protocols associated to each layer.

Using network data from layer 4 (Application) it is possible to perform RSF. On the Application layer, RSF can be used to identify service implementations by name and version [40]. For example, a RSF system can characterize different implementations of a File Transfer Protocol (FTP) [63], HyperText Transfer Protocol (HTTP) [17], and Domain Name System (DNS) [54] servers. The services usually provided by a default installation of an operating system cannot be effectively used to distinguish computers. However, additional and uncommon services can reveal the identity of a remote computer. In addition to the name and version of services, web pages of HTTP servers and anonymous directory listing of FTP servers can also be used as evidence and are likely to provide identity, as well as DNS table entries.

The data provided by layers 2 and 3 can be used to characterize different TCP/IP stack implementations. The protocols implemented on these layers are IP, versions 4 and 6 [14], the Internet Control Message Protocol (ICMP) [60] at layer 2, TCP [61] and User Datagram Protocol (UDP) [58] at layer 3. The implementation of these protocols specifications may vary from an operating system to another. In addition to the operating system itself, there are fewer other possibilities for identification that are few addressed by popular tools. These cases are discussed further in this chapter.

Finally, using data from the Link layer, it is possible to identify wireless devices, such as Wireless Access Points (WAP) and Network Interface Cards (NIC) [6]. Therefore, using data from different layers of the TCP/IP stack model it is possible to distinguish, or identify, a variety of elements of a remote computer. Quite apart from the network data used to perform RCF, the process of characterization and classification follows a common sequential procedure. A conceptual diagram for any RCF process is presented in Fig. 3.

The acquisition of network data is the first step in any RCF system (see Fig. 3). In fact, according to the way this raw network data is captured, it is possible to classify RCF systems into two categories, namely, active and passive fingerprinting, which are defined as follows.



**Fig. 3.** The Remote Computer Fingerprinting illustrated: (a) the network data captured is used to characterize the computer system through (b), producing the system description or fingerprint in (c); the fingerprint is then compared to other descriptions in (d) through a matching algorithm in (e); and, the result is presented in (f)

**Definition 2 (Active Fingerprinting).** The device which performs fingerprinting can send messages and inspect network traffic originated from the target to obtain meaningful information.

**Definition 3 (Passive Fingerprinting).** The device which performs fingerprinting can only inspect network traffic originated from the target machine to eventually find meaningful information.

In Active Fingerprinting the machine which performs RCF can send specially crafted packets to gain as much as possible information from the remote machine. In contrast, in Passive Fingerprinting, the machine which collects network data does not communicate with the remote computer. Therefore, it is natural that through Active Fingerprinting it is possible to gain more information about the remote machine. However, one may consider undesirable this proactive behavior and prefer the passive approach, since it is more probable to be detected or even damage the remote computer with the crafted packets [48].

After the acquisition step, it is necessary to perform feature extraction from raw network data. This step, named characterization, should produce a fingerprint which represents some aspect of the remote computer. This fingerprint is stored in a database of labeled fingerprints, which may be associated with suspicious activities records. Considering the later access to the remote computer, when it is seized, it is possible to create a new fingerprint and compare to that created by the detection system. This comparison is done in a classification step using adequate procedures.

According to the goal of the identification process, it is worth clarify that the fingerprint can be used to identify both a single computational networked element or a class of elements. For the single element case, we can cite the suspect's machine fingerprint, which can be used to differentiate it from another machines. When considering a class of elements, some examples includes the manufacturer of a machine or its operating system.

For the identification of a single element the goal is to extract data related to the machine's hardware components, in which its uniqueness can be used to differentiate

that machine from another. Even if two machines are visually identical, share the same manufacturer, same hardware components types, and configuration, there are unavoidable and unique differences between them, due to fabrication variations [36]. One can argue that a unique identification can be accomplished by gathering the remote Media Access Control (MAC) address of each NIC. However, the MAC address of each NIC is a hard-coded information and is supposed to be unique, in most operating systems it is also quite simple to change the original MAC address and hide the link layer identity.

Regarding the process to identify the class of a element, the extracted data must be related to the characteristics that are common to elements belonging to the same class. The uniqueness characteristics at this point rely on the behavior of that element, that are inserted, generally, by the manufacturer design or implementation decisions. For example, the specification of IEEE 802.11 Wireless Local Area Network (WLAN) standard permits that the firmware of different vendors have slight different control messages [6]. Hereafter, we present examples of features used to perform RSF, RNSF and RLF.

## 2.1 Remote Service Fingerprinting

The information provided by the services running on a remote machine can aid to reveal the identity of the remote computer owner, if it is suspect of some malicious or criminal cyber activity. RSF is perhaps the only fingerprinting method which can directly reveal personal information about the computer owner. For example, the files accessible on a FTP [63] service could have some identification, e.g. its user name. The files on a HTTP server also can have personal information of computer owner, e.g. personal home pages. However, such cases seems not be the rule and other mechanisms should be employed.

A complete report about service applications names and versions could be of great value to characterize remote computers. For example, consider the existence of a FTP server on a remote suspicious machine. If the server has some banner information, it can be used to compose the service fingerprint. We illustrate this in Fig. 4, showing a FTP session started from a Telnet [62] client to a public service.

To produce this FTP service data, a detection system could verify the availability of TCP port 21 on the remote system, or in other ports by more sophisticated means, as presented in [40, chap. 7]. Next, the detection system could verify the possibility of anonymous login. Finally, if possible, it retrieves the output of other convenient data from the server. In the FTP session just presented there are various server characteristics which can be used to build up a service fingerprint for the remote machine. More specifically, the following information can be extracted:

1. The existence of the FTP service itself, since its absence is also something to take into consideration;
2. The TCP port associated to the service, since some users can avoid to use default ports to hide services, or bypass firewall rules based on default ports;
3. The initial banner text, as found on line 1 of Fig. 4, which can reveal information about the computer owner, and the name and version of the server application;

```

1 220 ftp.NetBSD.org FTP server (NetBSD-ftpd 20100320) ready.
2 USER anonymous
3 331 Guest login ok, type your name as password.
4 PASS anonymous@example.com
5 230 Guest login ok, access restrictions apply.
6 SYST
7 215 UNIX Type: L8 Version: NetBSD-ftpd 20100320
8 HELP
9 214-
10 The following commands are recognized.
11 ('-' = not implemented, '+' = supports options)
12 USER REIN- TYPE ALLO MKD HELP MIC MLST+ MSND-
13 PASS PORT STRU REST PWD NOOP+ CONF MLSD MSOM-
14 ACCT- LPRT MODE RNFR LIST AUTH ENC MAIL- XCUP
15 CWD EPRT RETR RNTD NLST ADAT FEAT MLFL- XCWD
16 CDUP PASV STOR ABOR SITE PROT OPTS MRCP- XMKD
17 SMNT- LPSV STOU DELE SYST PBSZ MDTM MRSQ- XPWD
18 QUIT EPSV APPE RMD STAT CCC SIZE MSAM- XRMD

```

**Fig. 4.** An FTP session with a NetBSD server. The server allows anonymous authentication and reveals important information in the messages from the target machine.

4. The server application name and version, which can be obtained using specific FTP commands or using some inference based on commands availability, for example the implemented FTP commands;
5. The possibility of anonymous login, since this is not necessarily permitted;
6. The output of some FTP commands: (i) *SYST*, which reveal information about the operating system; and (ii) *HELP*, which describe the implemented commands.
7. A proper hash representation, using MD5 message-digest algorithm by [69], and the complete FTP session stream itself to verify authenticity.

This characterization could be expanded, for example, using the *STAT* command to describe the tree directory structure and reveal more information about the server status [63]. In fact, the effective characterization of services depends much on the understanding of the service itself. To illustrate the effectiveness of this proposed characterization, consider the FTP session presented in Fig. 5 taken from another public server.

In comparison with the server answers presented in Fig. 4, the answers in Fig. 5 differs in the characterization information from 3, 4, 6 and 7 descriptions. In addition, the possibility of customization of server greeting messages makes the RSF an effective tool for remote computer fingerprinting. Therefore, if the suspicious computer is a server, then RSF is probably applicable. Unfortunately, this seems to be usually improbable if the remote computer is a end-user machine, where the expectation is that no services are running.

**Passive Fingerprinting.** In addition to the Active Fingerprinting example presented, it is possible to detect not only server applications, but also the clients who connect to them. To achieve this, it is necessary to intercept data from the suspect machine to services. This is the case of HTTP connections, in which browser applications often describe its name and version, and in some cases even the client operating system,



```

1 220 beastie.tdk.net FTP server (Version 6.00LS) ready.
2 USER anonymous
3 331 Guest login ok, send your email address as password.
4 PASS anonymous@example.com
5 230 Guest login ok, access restrictions apply.
6 SYST
7 215 UNIX Type: L8 Version: BSD-199506
8 HELP
9 214- The following commands are recognized (* =>'s unimplemented).
10 USER PORT TYPE MLFL* MRCP* DELE SYST XMKD XCUP
11 PASS LPRT STRU MAIL* ALLO CWD FEAT RMD STOU
12 ACCT* EPRT MODE MSND* REST XCWD STAT XRMD SIZE
13 SMNT* PASV RETR MSOM* RNFR LIST HELP PWD MDTM
14 REIN* LPSV STOR MSAM* RNTO NLST NOOP XPWD
15 QUIT EPSV APPE MRSQ* ABOR SITE MKD CDUP

```

**Fig. 5.** An FTP session with a FreeBSD server. The server allows anonymous authentication and reveals important information in the messages from the target machine.

using the User-Agent header field [17]. Since this data should be initiated by the suspect machine, this identification could only be done passively.

**Availability and Reliability.** The ease of customization of services, which can expose the identity of remote computer owner, is also a drawback. Since it can be easily modified by the computer owner itself. For this reason and, probably, the nonexistence of services on end-user machines, RSF is not the most applicable and reliable RCF type. In fact, on the lower network layers, which is less controllable by the user than services and client applications, the reliability and availability increases.

## 2.2 Remote Network Stack Fingerprinting

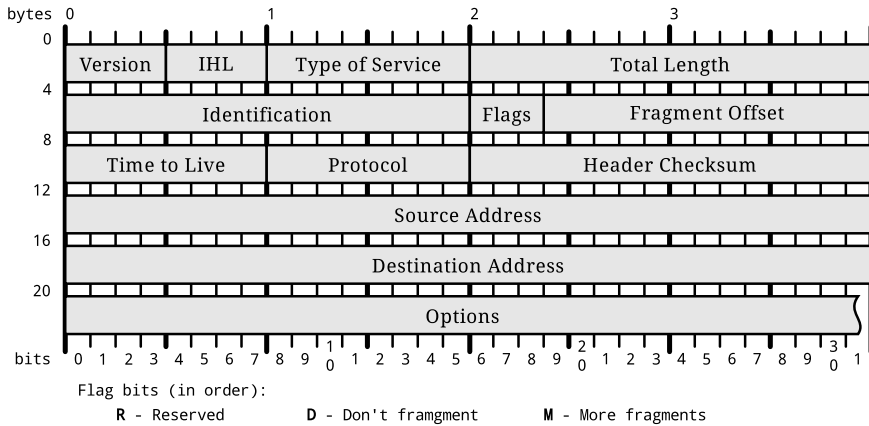
The first comprehensive work concerning with the RNSF process is due to [39], also known by his alias Fyodor. He published on the Phrack electronic hacker magazine a survey of techniques used to perform RNSF. Most of the techniques surveyed by [39] were implemented in the early versions of the tool named Nmap [40], a network security scanner [38]. Since then, Nmap was already improved, and the additional techniques supported in the later versions are presented in [40].

Network stack fingerprinting in TCP/IP networks is also known as remote TCP/IP fingerprinting, since the characterization of remote computers depends mainly on their implementations of TCP, UDP, IP and ICMP. In addition to the data these protocols should deliver, there are always a header in each message used by protocol implementations to guarantee its correct functioning. These headers often carry some sort of interpretation idiosyncrasies of protocol specifications. Also, there are some aspects that some specifications do not describe in sufficient detail to avoid different practices. To illustrate this, we consider the specification of IP and exemplify differences found at distinct implementations.

### 2.2.1 The Specificity of Protocol Headers

Protocols on the network layer of TCP/IP stack model (layer 2 on Fig. 2) should carry data from source host to destination host. Moreover, according to [5], IP is a

datagram internetwork service, providing no end-to-end delivery guarantees. To achieve this objective, the IP version 4, or IPv4 for short, uses a header in each message. This header is presented in Fig. 6 [60].



**Fig. 6.** IPv4 header used to carry data from source host to destination host providing no end-to-end delivery guarantees. The IPv4 datagram payload is preceded by a minimum of 20 bytes long header.

Although all IP header fields have a description of its structure and function by [59], there are some free parameters. To illustrate the possibility of using IPv4 header to characterize a remote machine consider these three fields: (i) identification, a value assigned by the sender to aid in assembling the fragments of a datagram; (ii) flags, more specifically the don't fragment bit, used to specify if a datagram is not to be fragmented; and (iii) time to live (TTL), which indicates the maximum time the datagram is allowed to remain in the Internet system. Hereafter we describe how the values of these three fields could be used to distinguish different IPv4 implementations. To support this claim, we present a case study based on the answer messages of two remote machines, which are replies to ICMP echo request messages [60]. The answers from a web search site are shown in Fig. 7.

```
IP (tos 0x0, ttl 52, id 2506, offset 0, flags [none], proto ICMP (1), length 84)
  searchengine.example.com > local: ICMP echo reply, id 21213, seq 1, length 64
IP (tos 0x0, ttl 52, id 2506, offset 0, flags [none], proto ICMP (1), length 84)
  searchengine.example.com > local: ICMP echo reply, id 21213, seq 2, length 64
IP (tos 0x0, ttl 52, id 2506, offset 0, flags [none], proto ICMP (1), length 84)
  searchengine.example.com > local: ICMP echo reply, id 21213, seq 3, length 64
```

**Fig. 7.** A description of three ICMP echo reply messages sent from a remote free web search engine address. It is convenient to highlight that all three IP identification field have the same value.

The messages are presented in the output format of Tcpcdump [29], a standard and classical tool for network traffic analysis [44]. The output was changed to hide the real addresses and time stamps. To compare with the answers of the search engine in Fig. 7, the answers of a free e-mail service are presented in Fig. 8.

```

IP (tos 0x0, ttl 241, id 62891, offset 0, flags [DF], proto ICMP (1), length 84)
  emailservice.example.com > local: ICMP echo reply, id 21224, seq 1, length 64
IP (tos 0x0, ttl 241, id 7584, offset 0, flags [DF], proto ICMP (1), length 84)
  emailservice.example.com > local: ICMP echo reply, id 21224, seq 2, length 64
IP (tos 0x0, ttl 241, id 16983, offset 0, flags [DF], proto ICMP (1), length 84)
  emailservice.example.com > local: ICMP echo reply, id 21224, seq 3, length 64

```

**Fig. 8.** A description of three ICMP echo reply messages sent from a remote free e-mail service address. It is convenient to highlight that the don't fragment bit was set on all reply messages.

For the sake of simplicity, we call the search engine in Fig. 7 by *G*, and the free e-mail service in Fig. 8 by *H*. Based on figures descriptions, we could highlight three differences: (i) the identification field of *G*'s messages are 2506, whereas in *H*'s messages they seem random; (ii) the don't fragment bit is set on *H*'s messages, and unset in *G*'s messages; and (iii) the initial TTL of *G*'s messages seems to be 64, where for *H*'s its seems to be 256, since initial TTL values are, commonly, powers of two [40, chap. 8]. Although this distinguishable field are exemplified with IP, these differences also exist in other protocols. For a complete survey about these techniques cf. [39] and [2].

### 2.2.2 Clock Skew Methods

This technique, introduced by [31], explores microscopic deviations in device's hardware, the clock skews. As one can note, even if two clocks are produced by the same manufacturer, they are not physically identical. High precision clocks are not required for common end-user devices, and due to cost constraints, most of the machines are built in with a simple clock, generally a quartz crystal clock. For a simple machine with 1 gigahertz processor, that clock oscillates in a granularity of  $10^{-9}$ , which means that the time's notion of this clock increases in the order of nanoseconds units.

Considering an ideal notion of time  $t$ , called real time, for typical clocks, its increasing value is not perfectly synchronized with  $t$ . This means that a clock can deviates from  $t$  in the course of time, in average, about 1 microsecond per second. Furthermore, considering that none of two clocks are physically identical, each of them can differ apart increasing by different speeds, the clock skew [27]. The clock skew can be described as the amount of deviation, between the system clock reference and the real time, introduced by the clock in the course of time. In other words, the clock skew corresponds to the first derivative of the offset in respect to  $t$ .

[31] propose a mechanism to infer about the clock skew of a machine system clock, by gathering its system clock at multiple points in time. To achieve this, they exploit the characteristics of the TCP option [28] which specifies that, when requested by the initiator of a TCP connection, all the packets for that flow will, necessarily, contain a

32-bit time stamp generated by the sender of the packet, included in the TCP option field. An interesting conclusion about the nature and behavior of the machine clock skew phenomena is that, for a single device clock, its clock skew is approximately constant over time. This enforces the knowledge about the remote device identification, but, as claimed by [31], this technique is not able for uniquely identify a device. Thus, its use is encouraged to be combined with other fingerprinting techniques.

This strategy is also addressed and improved by more recent developments. For example, [1] take into consideration the vulnerabilities of the previous method, including clock skew spoofing attacks, and proposes some improvement to the skew measurements and arithmetic. Also, in other studies the clock skew can be extracted from both the link [30] and service [27] layers.

## 2.3 Remote Link Fingerprinting

As defined by [5], the Link layer implements the protocols that are necessary to a host to communicate with its neighbors, in its directly-connected network. These implementations are made both in software and hardware to control the media-access and physical transmissions.

Several techniques can be applied to extract data from the link layer, each one by exploiting some physical characteristic or behavior of the suspect's machine. The RLF techniques discussed in this chapter will rely generally on the hardware properties of the analyzed machines and also the protocol implementations behavior.

### 2.3.1 Hardware Properties

Within the scope of the link layer, it is clear that the detection system must stay on the same local network as the suspect machine, to be able to collect some data. Although is reasonable to consider that the suspect machine will be connected by wire to the victim machine's network, this is not the general case. In most cases, the attacker will try to access the victim by exploiting its wireless network connectivity, since it only requires the access point or wireless router stay in the same operating range.

To address situations just like that one described in the last paragraph, most of the research are related to techniques for fingerprinting the wireless card of a suspect machine, mainly from its radio physical properties. These techniques rely on the observation of unique characteristics of the wireless device due to slight differences on the fabrication process which makes possible that each device has its own singularities. This information can be used both to distinguish a specific machine from another or to classify that machine as belonging to a class of devices (e.g. identifying its manufacturer, model or version). According to [36], some important contributions to this area consider various physical characteristics, such as carrier frequency, pulse width, pulse duration, pulse shape, pulse repetition interval, angle of arrival, amplitude, and radio signal transient.

**Electromagnetic Signatures.** For wireless networks based on IEEE 802.11, an important security requirement is to protect the privacy and anonymity of its users. This is one of the goals of authentication and encryption protocols, such as the

Wired Equivalent Privacy (WEP) protocol and the Wi-Fi Protected Access (WPA). However, user anonymity and privacy can be compromised if a node can be identified, or at least differentiated from another nodes. To demonstrate this, [68] performed a study of the measurement of distinctive radio-frequency (RF) electrical characteristics of six different IEEE 802.11b WLAN cards, and demonstrated what they call as electromagnetic signatures. At this initial and controlled experiment, they showed the feasibility of the electromagnetic fingerprinting, which can be used to a specific emitter identification.

**Radio Frequency Fingerprinting.** [74] present a security oriented study about the RF fingerprinting. The authors propose a classification system consisting of preprocessing, detection, feature extraction, and classification stages. From eight different IEEE 802.11b devices, the fingerprints are obtained by extracting the instantaneous amplitudes and phase angles of detected wave forms. An interesting point about this work is the insertion of a classification stage. At this stage a probabilistic neural network was used as a classifier of each transmitter. Once the neural network is trained with a set of known RF fingerprints, its classification was used to discover some unknown device. For the scenarios where a single emitter is transmitting with multiple antennas, by Multiple-Input Multiple-Output (MIMO) channels, the work of [36] proposes an algorithm for specific emitter identification. The authors take into account the unique behavior of the radio frequency front-end non-linearity, as they prove that its estimation are unavoidable and unique. For the wireless sensor node identification, [67] present a scheme based on the extraction of features of the nodes radio signal. The low-cost devices are very suitable to extract unique radio properties, since its composition and fabrication process are thought to be as cheap as possible. More RF fingerprinting techniques was proposed in the literature, with some of them addressed by [13].

**Hardware Properties Fingerprint Reliability.** Although the remote fingerprinting techniques on the hardware properties of devices are feasible and provides high identification accuracy, they are not free of defeating attacks. [12] demonstrate that impersonation attacks on the physical characteristics of hardware are a realistic possibility, both by replaying feature and signal of a victim's radio transmitter. However, these attacks are somewhat difficult to be performed, once the radio features are hard to record, since they can be channel and antenna dependent. A classification of these attacks are presented by [13], including their feasibility, limitations, and implications.

### 2.3.2 Protocol Implementation

Link layer protocols are driven by some standard rules, its specifications, which defines what the implementations must do. These standards do not, however, dictate how several of the services are to be implemented [9]. This allows manufacturers to make their implementations according to its needs and project designs, but not losing the protocol interoperability. To exemplify this, different vendors are free to establish its technique for adjusting transmission rates, reserving the link, polling for packets to conserve power, and probing the network for connectivity.

**Behavioral Fingerprinting.** With the objective to identify different kinds of wireless NICs, [9] present a temporal behavior analysis of a wireless stream, extracting subtle differences of its implementation design. The authors claim that differences in the composition of a wireless NIC influence data transmission patterns in a manner that is observable through traffic analysis. The implementation behavior in which their work focuses is the rate switching algorithm for cards manufactured by different vendors. This strategy which is defined by the IEEE 802.11 protocol specification performs a dynamic rate switching. It was designed to improve the performance of data transmission when changes are identified in channel conditions mainly originated by noise. Different rate switching algorithms impact on the duration of the frame transmission, arrival rate of frames, inter-frame delay, and are capable of being observable in the traffic of wireless stream. These temporal behaviors was analyzed by [9] and they performed the correct identification of three different card types, showing their proposal is a feasible fingerprinting technique for a class of wireless cards.

Following the idea of extracting a fingerprint of a wireless card according to the implementations of some link layer protocol, another unique behavior that can be extracted for the IEEE 802.11 is the mechanism of active scanning made by different vendors [10]. The scanning occurs when a node needs to discover available wireless networks to join. In the active scanning mechanism, unlike the passive mechanism, the client nodes are able to solicit an immediate response from an Access Point (AP), without waiting for beacon transmissions. For this, it broadcasts some probe request frames and waits for some probe response. The authors presented their results showing that the parameters that can vary per vendor, including the number of probe request frames to transmit per channel, the delay between probe request frames on the same channel, among others, are a useful behavioral fingerprint for that cards.

Another kind of behavioral fingerprint for the IEEE 802.11 implementations relies on the hypothesis that different implementations would react differently to non-standard events or malformed frames. [6] proposes an active method which consists of sending a stimulus frame to a suspect node, and observing its response or the lack of that. These stimulus consisted of frames with unusual combinations of fields, according to the protocol specification. By analyzing its responses, the authors were able to identify unique behaviors for each implementation, mainly for access points, which are considered to be defined only by its manufacturer. But even with these results, it is necessary to emphasize that this approach relies on a weak behavior, which can be reproduced by a malicious attacker.

**Timing Analysis.** Following a similar strategy to that presented by [10], [37] use the timing analysis of IEEE 802.11 probe request frames in active scanning mode to fingerprint some wireless device. The authors get the uniqueness for each device by the combinations of: (i) the time interval between frames of a machine, (ii) the wireless NIC driver, and (iii) the device operating system. They consider as a metric the periodic probe request intervals, which are different between link layer protocol implementations. [21] define an interesting strategy to extract a fingerprint to identify a specific access point type. To extract the signature of the AP, they consider sending a sequence of data through the AP, which they call a packet train, with another wireless node as destination. The forwarding of these packets by the AP generates the same

packet train, but now shifted in time, due to the internal architecture of the AP, its processing time. By capturing these forwarded packets, the packet inter arrival time can be extracted, and these values will be used as a sampling signal which corresponds to the fingerprint information to differentiate an AP to another.

## 2.4 Qualitative Considerations

The previous subsections showed that there are lots of different ways to performing remote computer fingerprinting. To categorize and assess this diversity of techniques, it is convenient to define some criteria. The following definitions create a basis to assess the performance of the different types of RCF techniques.

**Definition 4 (Availability).** Consider the existence of data necessary to perform Remote Computer Fingerprinting.

**Definition 5 (Efficacy).** Concerns how the data, and the quality of characterization and classification can be used to distinguish computers in the Remote Computer Fingerprinting process.

**Definition 6 (Efficiency).** Take into account the amount of data and time necessary to perform Remote Computer Fingerprinting, where efficiency is inverse to the amount of data and processing time needed.

**Definition 7 (Detectability).** Consider the possibility of identification of a Remote Computer Fingerprinting process.

**Definition 8 (Reliability).** Concerns with the data veracity and the general reliability of the Remote Computer Fingerprinting process.

**Definition 9 (Tractability).** Consider the qualitative aspect of Remote Computer Fingerprinting to not damage the expected function of the remote computer.

These definitions can be used as a qualitative performance measure associated to the use of different remote computer fingerprinting techniques. For network forensics, the main qualitative requirements are related to availability, efficacy and reliability. In Table 1, we summarize the expected qualitative performance measures for techniques of each RCF method.

The performance measures for each RCF type presents some variability. For the purpose of computer fingerprinting, the reliability and availability measures are very important. Therefore, RNSF seems to be the most suitable fingerprinting method to create digital evidence of suspicious network activity. However, the classical algorithms used to classify fingerprints can be defeated by fingerprinting countermeasure tools [73] and lose performance when applied to environments with protocol scrubbing [76, 77]. Also, some network stack data are underestimated, since they contain hidden features that could be used to characterize the network stack implementation. This fact was recently demonstrated by [52] on a survey concerning with the effectiveness of RNSF tools. In addition, [49] show that using intelligent methods it is possible to perform

**Table 1.** Expected performance of each Remote Computer Fingerprinting type. (a) The data certainly exists if the remote computer is a server and unlikely if it is an end user machine. (b) Some specially crafted packets can be easily detected. (c) Some specially crafted packets can make the remote operating systems unexpectedly stop. (d) If the remote machine is on the same LAN of the detection system, the data will exist. If it is not, there is also no data.

Performance measure	RSF	RNSF	RLF
Availability	depend <sup>(a)</sup>	high	depend <sup>(d)</sup>
Efficacy	high	high	high
Efficiency	low	high	high
Detectability	depend <sup>(b)</sup>	depend <sup>(b)</sup>	depend <sup>(b)</sup>
Reliability	low	high	high
Tractability	high	depend <sup>(c)</sup>	high

RNSF even in the presence of protocol scrubbing and anti-fingerprinting tools, e.g. Honeyd [64, 65].

To overcome this limitation, an interesting approach is to use data mining to perform feature selection and produce improved fingerprints, and then use pattern classification algorithms to identify them. The next section presents the current state of the art discussion about the use of computational intelligence to perform RNSF.

### 3 Intelligence in Remote Computer Fingerprinting

Given the process presented in Fig. 3, it is possible to highlight the following possible applications of computational intelligence: (i) forensics data preparation and feature extraction from raw network data, (ii) methods for classification of fingerprints, and (iii) inference results projection, representation and visualization.

When considering the application presented in Fig. 1, it is important to note that we focus on RNSF instead of Remote Operating System Detection (ROSD), which concerns to identify the operating systems of a remote machine. The main difference is that in ROSD, the fingerprint is used to determine the operating system, while with RNSF we aim to identify the subtle differences in the network communication properties that are intrinsic to the machine itself. Fortunately, the data and fingerprints used to perform ROSD are the same thing. Therefore, we should take advantage of the tools to support RNSF, since new information can be obtained with an extra classification. Hereafter, we present the state of the art about the use of computational intelligence in ROSD.

**Beverly [4]** developed what seems to be the first scientific reference on the use of computational intelligence for identifying remote systems. The author uses probabilistic learning to build up a Bayesian Classifier (BC) [8] which can identify remote operating systems in a passive fashion. Also, he verified improvement in operating system classification when compared to rule based systems, e.g. early versions of p0f [78, chap. 9].



**Burroni and Sarraute [7]** is the first work to deal with computational intelligence applied to active Remote Operating System Fingerprinting. An extended version of their work as published later in [71]. The authors deal with ROSD as an inference problem and classify the operating system of remote hosts as the most likely to generate the captured traffic. To achieve this, they use a neural network of Multi-Layer Perceptrons (MLP) [70] to classify the fingerprints according to known patterns. The proposed neural network provided a more reliable classification mechanism when compared with Nmap [40].

**Li et al [35]** provides similar results using data from a passive ROSF fingerprinting tool p0f [78, chap. 9], apparently, not aware of the previous work of [4]. However, instead of probabilistic learning, they use a back-propagation algorithm [70] with the Levenberg-Marquardt algorithm (LMA) [34, 43] to train a MLP network which produces better classification results.

**Gagnon et al [20]** discusses how Answer Set Programming (ASP) [42] can be used to address the problem of ROSF by logically specifying the problem and providing solutions through automated reasoning. This work is also presented in [19]. The results presented in these papers support two interesting assumptions: (i) using a knowledge base to keep previously deduced information enhance the accuracy; and (ii) a fully passive tool may not be sufficient in the context of intrusion detection.

**Greenwald and Thomas [24, 25]** use concepts of information theory [72] to evaluate the effectiveness of fingerprinting probes based on information gain. For example, while Nmap [40] transmits 16 different probe packets, they demonstrated successful fingerprinting with one to three packets. This result is quite important to perform ROSD efficiently and with efficacy. Furthermore, these packets are valid TCP synchronization packets to open ports, which are less likely to be detected and are tractable.

**Zhang et al [79]** propose a method to perform ROSD by using a Support Vector Machine (SVM) [11], which is a high generalization neural network because of its ability to simultaneously minimize the empirical classification error and maximize the geometric margin classification space. Experimental results on identification of signatures in the fingerprint database of different Nmap [40] versions show that the method is effective in the discovery of new signatures not included in the older database.

**Medeiros et al [45, 46]** propose a new method to improve the classification effectiveness of automation devices, where operating systems are usually proprietary or unknown. The proposal makes use of Self-Organizing Maps (SOM) to build a contextual feature map [32] that organizes operating systems according to the similarities of their TCP/IP fingerprints. This map is used to identify devices under test based on its operating system and may help security experts to select security tests according to the class the device belongs.

**Medeiros et al [47, 51]** used data mining techniques to propose three new ways of representation of the Nmap [40] ROSD database that can express how operating

systems are similar to each other according to their TCP/IP stack implementations. More specifically, they use a SOM [33], a Growing Neural Gas (GNG) [18], and the k-means algorithm [41] to assess the abilities of these algorithms on fingerprint database processing. In addition, they highlight applications in: (i) improvement on the capability of identifying unknown operating systems; (ii) compression of fingerprint databases; and (iii) fingerprint corruption evaluation.

**Medeiros et al [48, 50]** propose a method that uses only TCP synchronization messages to collect TCP ISN (Initial Sequence Number) samples, motivated by the fact that Nmap operating system detection [40, chap. 8] may harm sensitive TCP/IP stack implementations (such as those of some automation devices), causing the communication interruption. They use signal processing tools to classify the operating systems based on these samples. Using this technique, they show that it is possible to recognize operating systems using only one open TCP port on the target machine without compromise the device operation. Also, they present results which shows that their technique cannot be fooled by Honeyd [64] or protocol scrubbing [76, 77].

In general, the current state of the art in intelligent ROSD consists in the use of classifiers which outperforms the rule based matching algorithms of passive and active fingerprinting tools, i.e. p0f [78, chap. 9] and Nmap [40], respectively. The exception is the work of [48, 50] which does not use any previous database to characterization and classification of remote machines. The contributions of this literature to the process of intelligent Remote Network Stack Fingerprinting are summarized in Table 2.

**Table 2.** Description of the contributions of computational intelligence to Remote Network Stack Fingerprint. The acronyms are: ASP, answer set programming; BC, Bayesian classifier; GNG, growing neural gas; LMA, Levenberg-Marquardt algorithm; MLP, multi-layer perceptrons (trained with back-propagation algorithm); SOM, self-organizing map; and SVM, support vector machine.

<b>Contribution</b>	<b>References</b>	<b>Learning theory</b>
Characterization	Medeiros et al [48, 50]	SOM, Kohonen [32]
Classification	Beverly [4]	BC, Cooper and Herskovits [8]
	Sarraute and Burroni [71]	MLP, Rumelhart et al [70]
	Li et al [35]	MLP, Rumelhart et al [70] and LMA, Levenberg [34], Marquardt [43]
	Gagnon and Esfandiari [19]	ASP, Marek and Truszczyński [42]
	Zhang et al [79]	SVM, Cortes and Vapnik [11]
	Medeiros et al [45, 46]	SOM, Kohonen [32]
Data mining	Medeiros et al [51]	SOM, Kohonen [32], GNG, Fritzsche [18] and K-means, MacQueen [41]
Feature selection	Greenwald and Thomas [24, 25]	Information theory, Shanon [72]

The contributions presented in these works could be used to design or improve various performance aspects of a detection system which create evidences based on RNSF. Hereafter, we discuss how computational intelligence can be used to create more accurate fingerprints and perform more effective and reliable classification.

### 3.1 Creation and Classification of Fingerprint Evidences

Almost all the contributions concerning with the use of computational intelligence in RNSF are based on the fingerprint database of Nmap [40] and p0f [78, chap. 9]. A forensic detection system such as presented in Fig. 1 could incorporate all contributions that are available on each tool's database, but the focus of such forensic system is to identify machines uniquely, not only its operating system. Unfortunately, with current technologies, such feature is still very hard to be ensured. However, if there exists some information in the fingerprints database which can be used to make a machine distinguishable, the classification method used by the forensic detection system should accomplish this.

**Fingerprint Scrubbing.** Protocol scrubbers are transparent mechanisms that aims to compromise network scans and attacks at various protocol layers. The use of fingerprint scrubbing may compromise fingerprint's reliability [76]. This follows from the fact that some fields of the original message header are changed to prevent ROSD. Fingerprint scrubbers are designed specifically to avoid ROSD, specially with Nmap and p0f. Fortunately, they do not seem to be an obstacle for intelligent classification. For example, [35] and [52] verified that current fingerprint scrubbers do not prevent ROSD.

**Firewall Influence.** Some firewall rules can drop specially crafted packets used to perform ROSD. For example, some packets sent by Nmap are not usual, and can easily be detected. Furthermore, a common firewall configuration practice involves the filtering of closed ports. Since Nmap uses also closed ports to create a fingerprint, its classification will be defeated in this scenario. To illustrate this, consider the Nmap fingerprint of a remote machine in Fig. 9.

```

1 SEQ(SP=105%GCD=1%ISR=10A%TI=Z%TS=8)
2 OPS(O1=M5B4ST11NW6%O2=M578ST11NW6%O3=M280NNT11NW6%O4=M22CST11NW6%
. O5=M218ST11NW6%O6=M109ST11)
3 WIN(W1=16A0%W2=16A0%W3=16A0%W4=16A0%W5=16A0%W6=16A0)
4 ECN(R=Y%DF=Y%TG=40%W=16D0%O=M5B4NNSNW6%CC=Y%Q=)
5 T1(R=Y%DF=Y%TG=40%S=0%A=S+%F=AS%RD=0%Q=)
6 T2(R=N)
7 T3(R=N)
8 T4(R=Y%DF=Y%TG=40%W=0%S=A%A=Z%F=R%O=%RD=0%Q=)
9 U1(R=N)
10 IE(R=N)

```

**Fig. 9.** The Nmap fingerprint of a remote machine through the Internet. The machine is protected by a firewall which prevents the response of some Nmap probes.

The remote machine is accessible through the Internet and is located in a protected network of an university laboratory. The fingerprint in Fig. 9 is a detailed description of the responses of sixteen probe packets sent by Nmap to the remote machine [40, chap. 8]. The first six probe packets are used to build up the lines 1, 2, 3 and 5 of the fingerprint, which each line describe: (i) 'SEQ', a description of the way TCP time stamp, TCP ISN and IP ID sequences are generated; (ii) 'OPS', a description of the received TCP options; (iii) 'WIN', a description the received TCP window sizes; (iv) 'T1', a general description of the response of the first packet; and In addition, the line 'ECN' describes the packet sent in response to a TCP open port to test the implementation of Explicit Congestion Notification (ECN) [66].

The responses for the next six packets are described in lines 'T2' to 'T7'. As presented in Fig. 9, there was no responses for the packets associated to lines 'T2' and 'T3', which are send to open TCP ports. Moreover, the packets associated to lines 'T5', 'T6', and 'T7', which are send to closed ports, were not sent because a firewall filtered all closed ports of the remote host. An UDP packet is sent to a closed port to receive an ICMP port unreachable message (the response to this message is described in the line 'U1'). Finally, two ICMP echo request packets are sent, and the responses are described in the 'IE' line. As we can verify in the fingerprint of Fig. 9, there was no response for these last three packets also.

Simulating the seizure of the machine, the Nmap fingerprinting process was repeated in controlled environment. The local network was configured to guarantee a favorable scenario to Nmap. The information obtained from these new responses, presented in Fig. 10, is similar to the information in Fig. 9 for the received responses.

```

1 SEQ(SP=106%GCD=1%ISR=10C%TI=Z%CI=Z%II=I%TS=8)
2 OPS(O1=M5B4ST11NW6%O2=M5B4ST11NW6%O3=M5B4NNT11NW6%O4=M5B4ST11NW6%
. O5=M5B4ST11NW6%O6=M5B4ST11)
3 WIN(W1=16A0%W2=16A0%W3=16A0%W4=16A0%W5=16A0%W6=16A0)
4 ECN(R=Y%DF=Y%T=40%W=16D0%O=M5B4NNSNW6%CC=Y%Q=)
5 T1(R=Y%DF=Y%T=40%S=0%A=S+%F=AS%RD=0%Q=)
6 T2(R=N)
7 T3(R=Y%DF=Y%T=40%W=16A0%S=0%A=S+%F=AS%O=M5B4ST11NW6%RD=0%Q=)
8 T4(R=Y%DF=Y%T=40%W=0%S=A%A=Z%F=R%O=%RD=0%Q=)
9 T5(R=Y%DF=Y%T=40%W=0%S=Z%A=S+%F=AR%O=%RD=0%Q=)
10 T6(R=Y%DF=Y%T=40%W=0%S=A%A=Z%F=R%O=%RD=0%Q=)
11 T7(R=Y%DF=Y%T=40%W=0%S=Z%A=S+%F=AR%O=%RD=0%Q=)
12 U1(R=Y%DF=N%T=40%IPL=164%UN=0%RIPL=G%RID=G%RIPCK=G%RUCK=G%RUD=G)
13 IE(R=Y%DFI=N%T=40%CD=S)

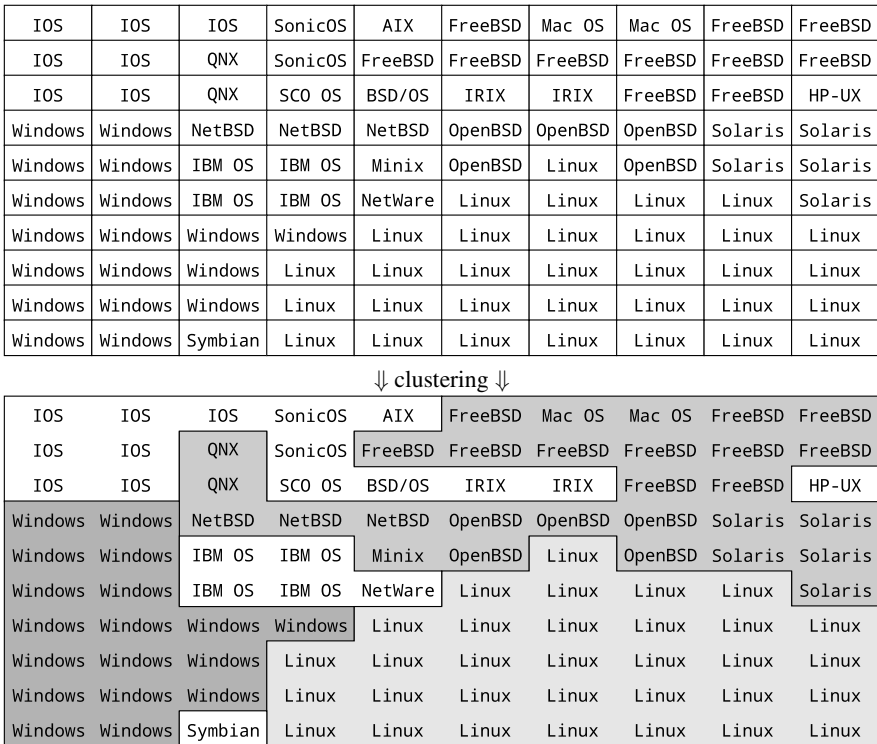
```

**Fig. 10.** The Nmap fingerprint of the same machine of Fig. 9 in a controlled LAN environment. In this case, the machine is supposed to answer all Nmap probes.

Comparing the two fingerprints we may conclude that the remote system did not reply only the packet probe associated to 'T2', which is a TCP null packet (a TCP packet with no flags), which is expected to be dropped by some operating systems. The remaining unresponsiveness is due to filtering rules of the internal network security mechanisms. However, it is important to note that the intersection of the existent responses from Fig. 9 and Fig. 10 are almost equal. Therefore, it is still possible to

perform network stack fingerprint in the presence of firewalls. Moreover, according to [24], valid TCP SYN messages not blocked for open TCP ports encodes most relevant information to distinguish different network stack implementations.

**Clustering Similar Fingerprints.** The performance of classical RNSF is negatively affected by the presence of firewalls and use of protocol scrubbing. Fortunately, the use of computational intelligence overcome this limitation. However, since the amount of data which is reliable on the worst-case scenario is at a premium, its is convenient to build up specialized classifiers for systems with similar fingerprints. The clustering of similar systems, according to its TCP/IP stack implementation, is firstly presented by [45, 51]. The authors present a labeled self-organizing map which groups operating systems based on its network stack fingerprint. Moreover, a visual analysis of the location of each fingerprint and its label, can guide the creation of three main clusters, as depicted in Fig. 11.



**Fig. 11.** Illustration of Kohonen’s self-organizing map of Nmap fingerprints. The clusters are used for the general classification of fingerprints in order to use the specialized neural network for specific classification of similar fingerprints.

The clustering of similar fingerprints to build specialized classifiers based on any classification work presented in the beginning of Sect. 3 can improve the classification of already used characterization databases, e.g. the databases of Nmap and p0f. However, TCP SYN negotiation mechanisms used by some operating systems (e.g. FreeBSD) can invalidate the reliability of classical data used to perform RNSF.

**TCP Synchronization Mechanisms.** To minimize the impact of TCP SYN flooding attacks, there are some mechanisms which interposes between the client and the server synchronization. Among them, the SYN cache and SYN cookies [15, 16], and the TCP SYN proxy, implemented by the OpenBSD Packet Filter (PF) [26]. Fortunately, these mechanisms can be detected or do not influence on specific informations used to perform RNSF such as the use of data present in an already established TCP communication, which is the case of clock skew based on TCP time stamps. In addition, a recent study by [50, 52] shows the viability of RNSF and the possible identification of such mechanisms using TCP ISN sequences. Hereafter, we explore the characterization process proposed in that work.

**The Generation of Sequence Numbers.** [78] presented an original method for characterization of TCP ISN pseudo random number generators. From [3], and [23], we can extract the following recommendation for the generation of TCP initial sequence numbers

$$s(c_{id}, t) = m(t) + f(c_{id}, t), \quad (1)$$

where  $s(c_{id}, t)$  is the initial sequence number for a connection identity  $c_{id}$  (source address and port, and destination address and port) at instant  $t$ ,  $m(t)$  is a random incremental function, generally represented by

$$m(t) = m(t-1) + r(t), \quad (2)$$

where  $r(t)$  is a random number generator. The function  $f(c_{id}, t)$  is a connection dependent term that is constant most of the time, represented by

$$f(c_{id}, t) = h(c_{id}, k(t)), \quad (3)$$

where  $h(\cdot)$  is a hash function which the second argument  $k(t)$  is an optional secret key input which may change over time. In their work [50] used the Pseudo Random Number Generator (PRNG)  $r(t)$  function to classify operating systems. Using Equations 1, 2 and 3, it is possible to recover samples of  $r(t)$  using the relation

$$\hat{r}(t) = s(c_{id}, t) - s(c_{id}, t-1), \quad (4)$$

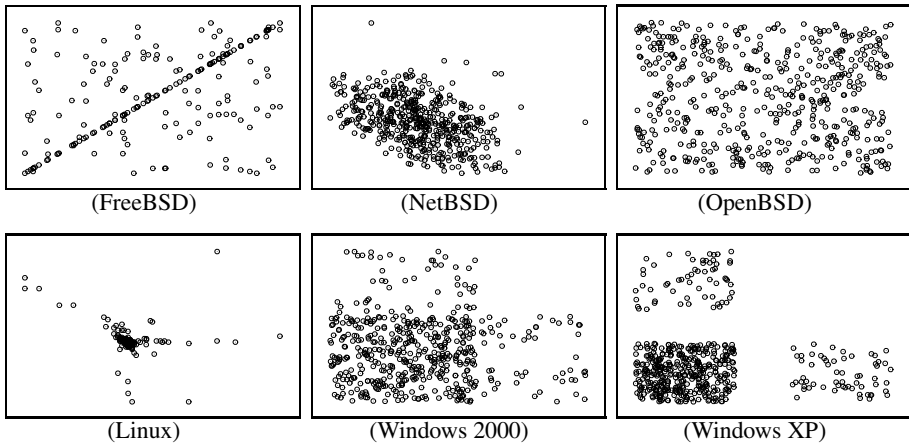
because  $f(c_{id}, t)$  will only change eventually, and will be constant for the most part of time. For the majority of cases,  $\hat{r}(t)$  will be equal to  $r(t)$ , however, when  $k(t)$  changes, the estimated value  $\hat{r}(t)$  is different of  $r(t)$ .

Not surprisingly, there exists operating systems which do not follow the initial recommendation of [3] or the standard by [23]. There are two common other alternative approaches to generate ISN: (i) use constant increments instead of random increments, and (ii) use directly pseudo random sequences instead of the incremental approach.

To visualize these sequences in a graphical plot, it is possible to use delayed coordinates. Each point in the plot is given by the relation

$$[x, y] = [\hat{r}(t), \hat{r}(t - 1)], \quad (5)$$

where  $t$  is iterated from the unity to the size of the sequence. The visualization of the sequences of six general purpose operating system are presented in Fig. 12.



**Fig. 12.** Portraits of the pseudo random number generators of six general purpose operating systems. All the generators are distinguishable and can be used to characterize the operating system of a remote machine. The linearity of FreeBSD portrait illustrates the presence of a SYN cache.

A visual analysis of the portraits presented in Fig. 12 can easily conclude that the pseudo random number generators of the TCP implementations of the six analysed operating systems are distinguishable. To automate this classification process [50] first used the SOM neural network to create reduced representations of the portrait. Furthermore, to classify these representations, the authors used a metric to compute the distance between two set of points. To create this fingerprint of the remote machine it is sufficient an open TCP port on the remote machine and hundreds of TCP ISN samples. This amount of samples is required based on the current developments of the classification system. Although, to detect TCP SYN mechanisms the amount of samples could be reduced to a few samples.

**PRNG Samples via Passive Fingerprinting.** In some types of computer network attacks, the machine used to perform the criminal activity sends a considerably number of TCP SYN packets. For example, in cyber attacks of the type of Denial of Service (DoS), it is possible to passively gather a sequence of TCP ISN from the network traffic and use it to characterize the PRNG of the attacker's machine.

### 3.2 The Architecture of an Intelligent Detection System

The definitions, case studies and techniques presented and discussed so far can ground the extension of the proposed detection system illustrated in Fig. 1. This extension is proposed as an architecture in which an incident detection system should create and preserve fingerprints of the machine used to perform the cyber crime. When some machine is seized, the machine would be inserted in a controlled network to verify if it was used in any previous detected cyber crime, similar to a human fingerprint database system. This whole process is described in Fig. 13.

Architectures similar to that presented in Fig. 13 are already explored in network forensics scientific literature. For example, in [53] the authors propose a passive fingerprinting method to automate chat room monitoring. Remote network stack fingerprinting is already used by [56, 55] to aid cyber sex crimes investigations. Despite the success of its use, classical remote computer fingerprinting systems may become unreliable if the suspect's machine is protected by firewalls, protocol scrubbers, TCP SYN mechanisms, or even fingerprinting courier measures such as Honeyd. To overcome this undesirable limitation, it is possible to use computational intelligence methods applied to ROSD. These methods can be effective even if we still use classical characterization databases and just apply new classification mechanisms. The use of clustering to build specialized classifiers is a natural step toward the creation of a more effective classification system, as described in Fig. 14.

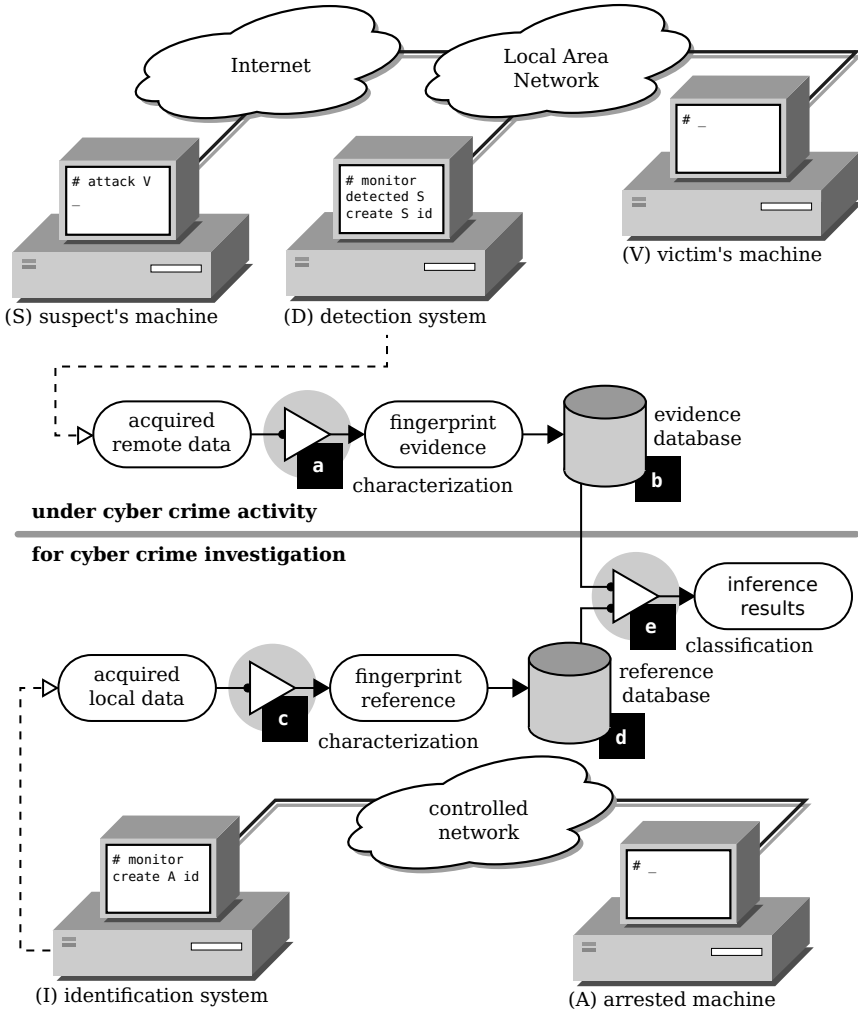
This classification mechanism must be considered general and should be adapted to the learning strategies used to classify and characterize fingerprints. The chosen learning algorithms must consider the possibility of use of computer fingerprint evidences in cyber crime investigations and judgments and its necessity of advances in reliability and efficacy of both data and inference. These advances are crucial to support the admissibility of these evidences, which is discussed in next section.

## 4 Legal Issues

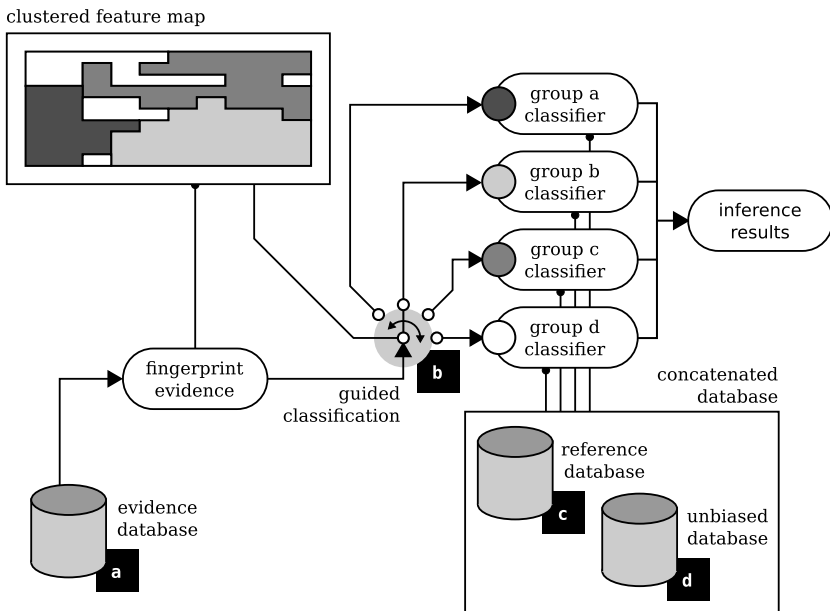
In the early days of digital forensics, the cyber crime investigations concerns with the use of law enforcement agents to monitor the cyber crimes and acquire evidences to arrest the suspect of the illegal activities. However, due to the simple and naive techniques used by the detectives, the veracity of these evidences was eventually challenged by defense attorneys. [56] discuss about a pedophilia case where an undercover detective, posing as a minor in chat rooms, talks to a man interested in sexual relations. The detective simply logged the online conversations, cut and paste the chat room transcripts into computer files, which was considered inadmissible as evidence.

This example lead us to take into consideration the question of how the evidences should be collected. While the current techniques for acquiring evidences, as those discussed in this chapter, are much more robust and feasible than the previous ones, the legal issue that arises concerns to the legality of these evidences, i.e. if it can be admitted into a court of law. This possibility can be used to determine another quality to RCF





**Fig. 13.** A conceptual network forensics system which detect, acquire, characterize (a) and preserve evidences (b) of suspicious network activity in a local area network. Furthermore, the characterization of the seized machine in a controlled network (c) is preserved (d) to possible identification (e) of previous use in criminal activities.



**Fig. 14.** Classification of evidences of Remote Network Stack Fingerprinting based on computational intelligence. Iteration over the evidence database (a) used to perform a classification mechanism guided by a clustering of similar fingerprints (b) classified by specialized classifiers trained with a concatenated database of seized machine fingerprints (c) and unbiased fingerprints from non-suspect machines (d) to prevent false positives.

techniques, the admissibility, defined next. The admissibility depends on the country or state in which the prosecution occurs. But, as a general rule, the admissibility depends on: (i) the reliability of the evidence, and (ii) if the method to gather the evidence does not violate the person's *reasonable expectation of privacy* [55].

**Definition 10 (Admissibility).** Concerns to the legality of these evidences, i.e. if it can be admitted into a court of law. Generally, it depends on reliability, and if does not violate suspect's reasonable expectation of privacy.

In the case of Remote Computer Fingerprinting, the reliability of the acquired evidences depends on the scientific method used, i.e. if is proven that it yields a correct and possible unique fingerprint to the suspect's machine. Regarding the privacy issues, a remote fingerprint technique employed must consider extracts the evidence either from public domain, e.g. an information available on the Internet, and using a technology that is in the general public use. For the cases where there is no such public information or technology, one can try to explore the exceptions to the rules, but it will be safer to rely on the one supported by a court-issued warrant. In addition, [75] highlight that a forensic testimony, to be admissible in court, must be based on scientifically valid methodology, i.e. methods that: (i) are peer reviewed, (ii) based on testable hypotheses, (iii) have a known error rate (false positives and true negatives), (iv) follow an existing

set of standards, and (v) are generally accepted within the scientific community. In addition, it should not violate suspect's reasonable expectation of privacy.

Therefore, in addition to the technical challenges associated to the process of remote computer fingerprinting, there are a variety of legal issues that must be considered. As said by [75], the use of techniques or inspection of traffic without a warrant, or making manipulations to the protocols beyond its normal behavior, which is the case for some of the active fingerprinting techniques, are violation to the law enforcement, and will invalidate the acquired evidences. These are just a few of legal issues to be concerned when dealing to forensics research, more discussion about this and jurisprudence example cases can be found in [57, 75, 22] and [40, chap. 1].

## 5 Conclusion

This chapter introduced the concept of Remote Computer Fingerprinting. Although the remotely creation of computer fingerprints are practically developed since the middle of ninety decade, concerning mainly with computer security, its use to aid cyber crime investigations is a recent research subject. Due to substantial differences on its application to network security, the use of computer fingerprints in network forensics should revise the related scientific literature, since its main application was in remote operating system detection. This distinction of applications in network security and forensics is well explored by [75].

To correctly classify the methods presented in the literature, we introduced a new taxonomy for remote computer fingerprinting methods. More specifically, we classify the methods according to the network data they use to characterize the remote machine. Considering its practical use, this classification is based on the Internet stack model. Therefore, Remote Computer Fingerprinting was divided into three classes: Remote Service Fingerprinting, Remote Network Stack Fingerprinting and Remote Link Fingerprinting. To enforce the connection between computer scientist and forensic field expert, we introduced the basic mechanisms behind the different techniques used to perform computer fingerprints remotely.

The different techniques which characterize machines remotely were evaluated. The main result is the performance classification presented in Table 1. This analysis concludes that network data from layers 2 and 3 of the TCP/IP stack model (cf. Fig. 2) is the most valuable in various aspects. This fact guided the work to the application of computational intelligence in Remote Network Stack Fingerprinting.

After the presentation of necessary definitions and the adaptation of the concepts of Remote Computer Fingerprinting from network security literature, the use of computational intelligence was justified by two main contributions: the improvement in classification and the use of data mining techniques to perform feature extraction and clustering. These contributions were summarized in Table 2, which supports the proposed architecture to aid the creation and classification of remote computer fingerprints. Moreover, to enhance the possibility of use of computer fingerprint evidences in cyber crime investigations and judgments, it is necessary advances in the study of reliability and efficacy of both data and inference.

Furthermore, the application possibilities of remote identification of systems and devices in digital forensics can partially supply the need of sophisticated tools to

acquire, preserve, examine, analyze and present digital evidences in cyber crime investigations. Therefore, the use of remote computer fingerprinting to aid cyber crime investigations must evolve, since its applications were just initially explored, it should become a new research area seeking for more sophisticated tools.

## References

- [1] Arackaparambil, C., Bratus, S., Shubina, A., Kotz, D.: On the reliability of wireless fingerprinting using clock skews. In: Proceedings of the Third ACM Conference on Wireless Network Security (WiSec), pp. 169–174 (2010), doi:10.1145/1741866.1741894
- [2] Arkin, O., Yarochkin, F.: ICMP based remote OS TCP/IP stack fingerprinting techniques. Phrack Magazine 11(57) (2001)
- [3] Bellovin, S.: RFC 1948 (Informational), Defending Against Sequence Number Attacks. Internet Engineering Task Force (IETF) (1996)
- [4] Beverly, R.: A robust classifier for passive TCP/IP fingerprinting. In: Barakat, C., Pratt, I. (eds.) PAM 2004. LNCS, vol. 3015, pp. 158–167. Springer, Heidelberg (2004)
- [5] Braden, R.: RFC 1122 (Standard), Requirements for Internet Hosts – Communication Layers. Internet Engineering Task Force (IETF) (1989)
- [6] Bratus, S., Cornelius, C., Kotz, D., Peebles, D.: Active behavioral fingerprinting of wireless devices. In: Proceedings of the First ACM Conference on Wireless Network Security (WiSec), pp. 56–61 (2008), doi:10.1145/1352533.1352543
- [7] Burrioni, J., Sarraute, C.: Using neural networks for remote OS identification. In: Proceedings of the 3rd Pacific Security Conference (PacSec) (2005)
- [8] Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning 9(4), 309–347 (1992), doi:10.1007/BF00994110
- [9] Corbett, C.L., Beyah, R.A., Copeland, J.A.: A passive approach to wireless NIC identification. In: Proceedings of IEEE International Conference on Communications (ICC), pp. 2329–2334 (2006), doi:10.1109/ICC.2006.255117
- [10] Corbett, C.L., Beyah, R.A., Copeland, J.A.: Passive classification of wireless NICs during active scanning. International Journal of Information Security 7(5), 335–348 (2008), doi:10.1007/s10207-007-0053-7
- [11] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995), doi:10.1007/BF00994018
- [12] Danev, B., Luecken, H., Capkun, S., Defrawy, K.E.: Attacks on physical-layer identification. In: Proceedings of the Third ACM Conference on Wireless Network Security (WiSec), pp. 89–98 (2010), doi:10.1145/1741866.1741882
- [13] Danev, B., Zanetti, D., Capkun, S.: On physical-layer identification of wireless devices. ACM Computing Surveys 45(1) (2012), doi:10.1145/2379776.2379782
- [14] Deering, S., Hinden, R.: RFC 2460 (Draft Standard), Internet Protocol, Version 6 (IPv6) Specification. Internet Engineering Task Force (IETF) (1998)
- [15] Eddy, W.M.: Defenses against TCP SYN flooding attacks. The Internet Protocol Journal 9(4), 2–16 (2006)
- [16] Eddy, W.M.: RFC 4987 (Informational), TCP SYN Flooding Attacks and Common Mitigations. Internet Engineering Task Force (IETF) (2007)
- [17] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., Berners-Lee, T.: RFC 2068 (Proposed Standard), Hypertext Transfer Protocol – HTTP/1.1. Internet Engineering Task Force (IETF) (1999)
- [18] Fritzke, B.: A growing neural gas network learns topologies. In: Tesauro, G., Touretzky, D., Leen, T. (eds.) Advances in Neural Information Processing Systems, vol. 7, pp. 625–632. MIT Press (1995)

- [19] Gagnon, F., Esfandiari, B.: Using answer set programming to enhance operating system discovery. In: Erdem, E., Lin, F., Schaub, T. (eds.) LPNMR 2009. LNCS, vol. 5753, pp. 579–584. Springer, Heidelberg (2009)
- [20] Gagnon, F., Esfandiari, B., Bertossi, L.: A hybrid approach to operating system discovery using answer set programming. In: Proceedings of the 10th IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 391–400 (2007), doi:10.1109/INM.2007.374804
- [21] Gao, K., Corbett, C., Beyah, R.: A passive approach to wireless device fingerprinting. In: Proceedings of the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), pp. 383–392 (2010), doi:10.1109/DSN.2010.5544294
- [22] Garfinkel, S.L.: Digital forensics research: The next 10 years. *Digital Investigation* 7, S64–S73 (2010), doi:10.1016/j.diin.2010.05.009
- [23] Gont, F., Bellovin, S.: RFC 6528 (Standards Track), Defending Against Sequence Number Attacks. Internet Engineering Task Force (IETF) (2012)
- [24] Greenwald, L.G., Thomas, T.J.: Toward undetected operating system fingerprinting. In: Proceedings of the First USENIX Workshop on Offensive Technologies (WOOT) (2007)
- [25] Greenwald, L.G., Thomas, T.J.: Understanding and preventing network device fingerprinting. *Bell Labs Technical Journal* 12(3), 149–166 (2007), doi:10.1002/bltj.20257
- [26] Hartmeier, D.: Design and performance of the OpenBSD stateful packet filter (pf). In: Proceedings of the FREENIX Track: USENIX Annual Technical Conference, pp. 171–180 (2002)
- [27] Huang, D.J., Yang, K.T., Ni, C.C., Teng, W.C., Hsiang, T.R., Lee, Y.J.: Clock skew based client device identification in cloud environments. In: Proceedings of the IEEE 26th International Conference on Advanced Information Networking and Applications (AINA), pp. 526–533 (2012), doi:10.1109/AINA.2012.51
- [28] Jacobson, V., Braden, R., Borman, D.: RFC 1323 (Proposed Standard), TCP Extensions for High Performance. Internet Engineering Task Force (IETF) (1992)
- [29] Jacobson, V., Leres, C., McCanne, S.: TCPDUMP/LIBPCAP public repository, version 4.3.0 (2012), <http://www.tcpdump.org/> (released on June 2012)
- [30] Jana, S., Kasera, S.K.: On fast and accurate detection of unauthorized wireless access points using clock skews. *IEEE Transactions on Mobile Computing* 9(3), 449–462 (2010), doi:10.1109/TMC.2009.145
- [31] Kohno, T., Broido, A., Claffy, K.: Remote physical device fingerprinting. *IEEE Transactions on Dependable and Secure Computing* 2(2), 93–108 (2005), doi:10.1109/TDSC.2005.26
- [32] Kohonen, T.: Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43(1), 59–69 (1982)
- [33] Kohonen, T.: *Self-Organizing Maps*, 3rd edn. Springer (2001)
- [34] Levenberg, K.: A method for the solution of certain non-linear problems in least squares. *Quarterly of Applied Mathematics* 2, 164–168 (1944)
- [35] Li, W., Zhang, D.-F., Yang, J.: Remote OS fingerprinting using BP neural network. In: Wang, J., Liao, X.-F., Yi, Z. (eds.) ISNN 2005. LNCS, vol. 3498, pp. 367–372. Springer, Heidelberg (2005)
- [36] Liu, M.W., Doherty, J.F.: Wireless device identification in MIMO channels. In: Proceedings of the 43rd Annual Conference on Information Sciences and Systems (CISS), pp. 563–567 (2009), doi:10.1109/CISS.2009.5054783
- [37] Loh, D.C.C., Cho, C.Y., Tan, C.P., Lee, R.S.: Identifying unique devices through wireless fingerprinting. In: Proceedings of the First ACM Conference on Wireless Network Security (WiSec), pp. 46–55 (2008), doi:10.1145/1352533.1352542
- [38] Lyon, G.F.: The art of port scanning. *Phrack Magazine* 7(51) (1997)
- [39] Lyon, G.F.: Remote OS detection via TCP/IP fingerprinting. *Phrack Magazine* 8(54) (1998)

- [40] Lyon, G.F.: Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning. Insecure.Com LLC (2009)
- [41] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967)
- [42] Marek, V.W., Truszczyński, M.: Stable models and an alternative logic programming paradigm. In: Apt, K.R., Marek, V.W., Truszczyński, M., Warren, D.S. (eds.) *The Logic Programming Paradigm: A 25-Year Perspective*, pp. 375–398. Springer (1999), doi:10.1007/978-3-642-60085-2\_17
- [43] Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11(2), 431–441 (1963), doi:10.1137/0111030
- [44] McCanne, S., Jacobson, V.: The BSD packet filter: A new architecture for user-level packet capture. In: Proceedings of the USENIX Winter 1993 Conference, pp. 259–269 (1993)
- [45] Medeiros, J.P.S., Cunha, A.C., Brito Jr., A.M., Motta Pires, P.S.: Application of kohonen maps to improve security tests on automation devices. In: Lopez, J., Hämmerli, B.M. (eds.) *CRITIS 2007. LNCS*, vol. 5141, pp. 235–245. Springer, Heidelberg (2008)
- [46] Medeiros, J.P.S., Cunha, A.C., Brito, A.M., Pires, P.S.M.: Automating security tests for industrial automation devices using neural networks. In: Proceedings of the 12th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 772–775 (2007), doi:10.1109/EFTA.2007.4416854
- [47] Medeiros, J.P.S., Brito Jr., A.M., Pires, P.S.M.: A data mining based analysis of Nmap operating system fingerprint database. In: Herrero, Á., Gastaldo, P., Zunino, R., Corchado, E. (eds.) *CISIS 09. AISC*, vol. 63, pp. 1–8. Springer, Heidelberg (2009)
- [48] Medeiros, J.P.S., Brito, A.M., Pires, P.S.M.: A new method for recognizing operating systems of automation devices. In: Proceedings of the 14th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4 (2009), doi:10.1109/ETFA.2009.5347095
- [49] Medeiros, J.P.S., Santos, S.R., Brito, A.M., Pires, P.S.M.: Advances in network topology security visualisation. *International Journal of System of Systems Engineering* 1(4), 387–400 (2009), doi:10.1504/IJSSE.2009.031347
- [50] Medeiros, J.P.S., Brito Jr., A.M., Motta Pires, P.S.: An effective TCP/IP fingerprinting technique based on strange attractors classification. In: Garcia-Alfaro, J., Navarro-Arribas, G., Cuppens-Boulahia, N., Roudier, Y. (eds.) *DPM 2009. LNCS*, vol. 5939, pp. 208–221. Springer, Heidelberg (2010)
- [51] Medeiros, J.P.S., Brito, A.M., Pires, P.S.M.: Using intelligent techniques to extend the applicability of operating system fingerprint databases. *Journal of Information Assurance and Security* 5(4), 554–560 (2010)
- [52] Medeiros, J.P.S., de Medeiros Brito Júnior, A., Motta Pires, P.S.: A qualitative survey of active TCP/IP fingerprinting tools and techniques for operating systems identification. In: Herrero, Á., Corchado, E. (eds.) *CISIS 2011. LNCS*, vol. 6694, pp. 68–75. Springer, Heidelberg (2011)
- [53] Meehan, A., Manes, G., Davis, L., Hale, J., Sheno, S.: Packet sniffing for automated chat room monitoring and evidence preservation. In: Proceedings of the 2001 IEEE Workshop on Information Assurance and Security, pp. 285–288 (2001)
- [54] Mockapetris, P.: RFC 1035 (Internet Standard), Domain Names – Implementation and Specification. Internet Engineering Task Force (IETF) (1987)
- [55] Novotny, J., Schulte, D., Manes, G., Sheno, S.: Remote computer fingerprinting for cyber crime investigations. In: di Vimercati, S.D.C., Ray, I., Ray, I. (eds.) *Data and Applications Security XVII. IFIP*, vol. 142, pp. 3–15. Springer, Boston (2004)

- [56] Novotny, J.M., Meehan, A., Schulte, D., Manes, G.W., Sheno, S.: Evidence acquisition tools for cyber sex crimes investigations. In: Proceedings of the SPIE, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Defense and Law Enforcement, vol. 4708, pp. 53–60 (2002), doi:10.1117/12.479292
- [57] Pollitt, M., Caloyannides, M., Novotny, J., Sheno, S.: Digital forensics: Operational, legal and research issues. In: di Vimercati, S.D.C., Ray, I., Ray, I. (eds.) Data and Applications Security XVII. IFIP, vol. 142, pp. 393–403. Springer, Boston (2004)
- [58] Postel, J.: RFC 768 (Internet Standard), User Datagram Protocol. Internet Engineering Task Force (IETF) (1980)
- [59] Postel, J.: RFC 791 (Internet Standard), Internet Protocol – DARPA Internet Program, Protocol Specification. Internet Engineering Task Force (IETF) (1981)
- [60] Postel, J.: RFC 792 (Internet Standard), Internet Control Message Protocol – DARPA Internet Program, Protocol Specification. Internet Engineering Task Force (IETF) (1981)
- [61] Postel, J.: RFC 793 (Internet Standard), Transmission Control Protocol – DARPA Internet Program, Protocol Specification. Internet Engineering Task Force (IETF) (1981)
- [62] Postel, J., Reynolds, J.: RFC 854 (Internet Standard), Telnet Protocol Specification. Internet Engineering Task Force (IETF) (1983)
- [63] Postel, J., Reynolds, J.: RFC 959 (Internet Standard), File Transfer Protocol (FTP). Internet Engineering Task Force (IETF) (1985)
- [64] Provos, N.: A virtual honeypot framework. In: Proceedings of the 13th USENIX Security Symposium (2004)
- [65] Provos, N., Holz, T.: Virtual Honeypots: From Botnet Tracking to Intrusion Detection. Addison-Wesley (2008)
- [66] Ramakrishnan, K., Floyd, S., Black, D.: RFC 3168 (Proposed Standard), The Addition of Explicit Congestion Notification (ECN) to IP. Internet Engineering Task Force (IETF) (2001)
- [67] Rasmussen, K.B., Capkun, S.: Implications of radio fingerprinting on the security of sensor networks. In: Proceedings of the Third International Conference on Security and Privacy in Communications Networks and the Workshops (SecureComm), pp. 331–340 (2007), doi:10.1109/SECCOM.2007.4550352
- [68] Remley, K., Grosvenor, C., Johnk, R., Novotny, D., Hale, P., McKinley, M.: Electromagnetic signatures of WLAN cards and network security. In: Proceedings of Fifth IEEE International Symposium on Signal Processing and Information Technology, pp. 484–488 (2005), doi:10.1109/ISSPIT.2005.1577145
- [69] Rivest, R.: RFC 1321 (Informational), The MD5 Message-Digest Algorithm. Internet Engineering Task Force (IETF) (1992)
- [70] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* 323(6088), 533–536 (1986), doi:10.1038/323533a0
- [71] Sarraute, C., Burrioni, J.: Using neural networks to improve classical operating system fingerprinting techniques. *Electronic Journal of SADIO* 8(1), 35–47 (2008)
- [72] Shannon, C.E.: A mathematical theory of communication. *Bell System Technical Journal* 27(3), 379–423 (1948)
- [73] Smart, M., Malan, G.R., Jahanian, F.: Defeating TCP/IP stack fingerprinting. In: Proceedings of the 9th USENIX Security Symposium (2000)
- [74] Ureten, O., Serinken, N.: Wireless security through RF fingerprinting. *Canadian Journal of Electrical and Computer Engineering* 32(1), 27–33 (2007), doi:10.1109/CJECE.2007.364330
- [75] Walls, R.J., Levine, B.N., Liberatore, M., Shields, C.: Effective digital forensics research is investigator-centric. In: Proceedings of the 6th USENIX Conference on Hot Topics in Security (HotSec) (2011)

- [76] Watson, D., Smart, M., Malan, G., Jahanian, F.: Protocol scrubbing: network security through transparent flow modification. In: Proceedings of the DARPA Information Survivability Conference and Exposition II (DISCEX), pp. 108–118 (2001), doi:10.1109/DISCEX.2001.932163
- [77] Watson, D., Smart, M., Malan, G., Jahanian, F.: Protocol scrubbing: network security through transparent flow modification. *IEEE/ACM Transactions on Networking* 12(2), 261–273 (2004), doi:10.1109/TNET.2003.822645
- [78] Zalewski, M.: *Silence on the Wire: A Field Guide to Passive Reconnaissance and Indirect Attacks*, 1st edn. No Starch Press (2005)
- [79] Zhang, B., Zou, T., Wang, Y., Zhang, B.: Remote operation system detection base on machine learning. In: Proceedings of the International Conference on Frontier of Computer Science and Technology, pp. 539–542 (2005), doi:10.1109/FCST.2009.21



# Signature-Based Biometric Authentication

Srikanta Pal<sup>1</sup>, Umapada Pal<sup>2</sup>, and Michael Blumenstein<sup>1</sup>

<sup>1</sup> School of Information and Communication Technology,  
Griffith University, Gold Coast, Australia

<sup>2</sup> CVPRU, Indian Statistical Institute, Kolkata, India  
srikanta.pal@griffithuni.edu.au

**Abstract.** In a modern, civilized and advanced society, reliable authentication and authorization of individuals are becoming more essential tasks in several aspects of daily activities and as well as many different important applications such as in financial transactions, access control, travel and immigration, healthcare etc. In some situations, when individual equipment is required for confirmation of one's identity to other groups of people in order to make use of services or to achieve access to physical places, it is always necessary to declare self-identity and to prove the claim. Traditional authentication methods, which are based on knowledge (password-based authentication) or the utility of a token (photo ID cards, magnetic strip cards and key-based authentication), are less reliable because of loss, forgetfulness and theft.

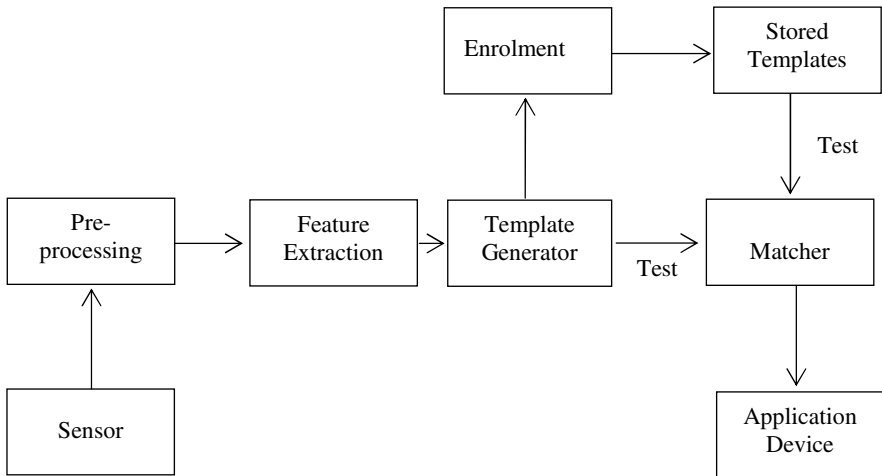
These issues direct substantial attention towards biometrics as an alternative method for person authentication and identification. The word 'biometric' has been derived from the Greek words "Bio-metriks", "Bio" which means life and "metriks" which means measures. Therefore a biometric is the measurement and statistical analysis of unchanging biological characteristics. Biometrics evaluate a person's unique physical or behavioural traits to authenticate their identity. As biometric identifiers are unique to persons, they are more reliable in verifying identity than token-based and knowledge-based methods. In the last few years, substantial efforts have been devoted to the development of biometric-based authentication systems. Biometrics provide an expected and successful solution to the authentication problem, as it offers the construction of systems that can identify individuals by the analysis of their physiological or behavioural characteristics [1]. In fact, the field of biometrics is the science of using digital technologies and the intention of biometric systems is to perform the recognition or authentication of people based on some biological characteristics that are intrinsically unique for each individual. The effectiveness of a biometric system is measured mainly by the distinguishing attributes that are used to verify the identity. A large number of biometric traits have been investigated and some of them are nowadays used in several applications. Common physical traits include fingerprints, ear, hand or palm geometry, vein, retina, iris and facial characteristics [2]. Behavioural traits include voice, signature, keystroke pattern and gait.

A biometric scheme can either verify or identify the authentication of an individual. In verification mode, it authenticates the person's identity on the basis of his/her claimed identity. In identification mode, it establishes the person's identity (among those enrolled in a database) without the subjects

having to claim their identity [3]. Among all other biometric traits, signature verification occupies an important and a very special place in the field of biometrics.

## 1 Overview of Biometric Traits and Technologies

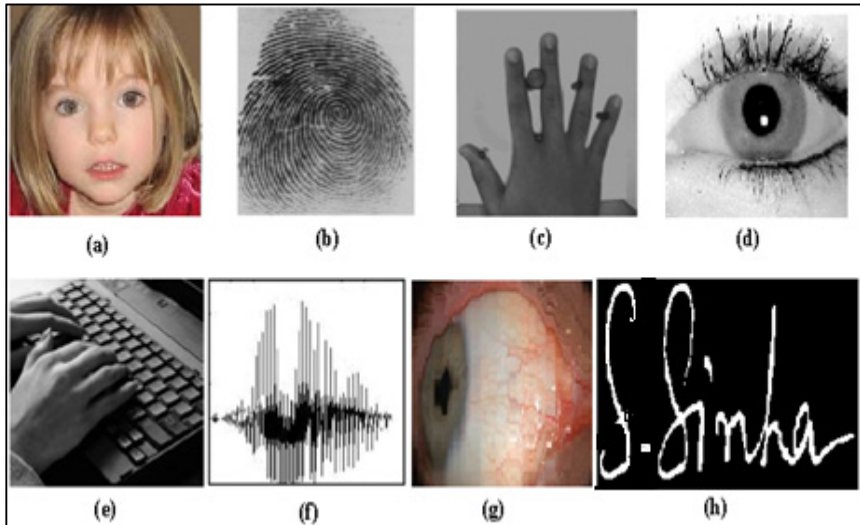
Biometric systems are a constantly growing technology, which have been widely used in many official and commercial identification applications. A biometric method is essentially a pattern recognition system which makes a personal identification decision by determining the authority of specific physiological or behavioural traits [4]. Nowadays a large number of biometric traits have been investigated and some of them are used in several applications. The diagram of a generic biometric system as specified in [39] is shown in Figure 1.



**Fig. 1.** A Generic Biometric System

Each biometric technology has its strengths and limitations. It is not expected that one biometric trait will efficiently fulfil the needs of all the applications. The match between a specific biometric and an application is determined depending upon the requirements of the application and the properties of the biometric characteristic. A number of biometric characteristics have been in use for different applications [5]. Each biometric characteristic has its effectiveness and disadvantages, and the choice depends on the specific application. No single biometric is expected to successfully meet all of the requirements (e.g., accuracy, practicality, and cost) of all applications (e.g., digital right management, access control, and welfare distribution) [6]. In other words, no biometric is “optimal” although a number of them are “admissible.” The suitability of a specific biometric for a particular application is determined depending upon the requirements of the application and the properties of the biometric

characteristic. A large number of biometric traits have been explored and a brief description of some important and commonly used biometric traits is discussed as follows. Examples of different biometric characteristics are shown in Figure 2.



**Fig. 2.** Examples of Some Biometric Characteristics: (a) Face, (b) Fingerprint, (c) Hand geometry, (d) Iris, (e) Keystroke, (f) Voice, (g) Sclera, (h) Signature

- Face recognition investigates facial characteristics, and facial images are one of the common biometric characteristics to undertake personal recognition. A facial recognition method is an application of computer for automatically identifying or verifying a person from a digital image. A face recognition scheme generally consists of four modules: face detection and tracking, facial feature finding, face representation, and matching [7]. In a research by Jain et al. [10], authors have indicated that the applications of facial recognition range from a static, controlled authentication to a dynamic, uncontrolled face identification process. The face recognition approaches [8] are usually based on either: (a) the position and shape of facial characteristics (eyes, eyelid, eyebrows, nose, lips, and chin and their spatial relationships) or (b) the investigation of the face image samples. The main objective of face recognition system is security related, but there are some kind of applications related to personal use, convenience and productivity enhancement.
- The pattern of ridges and valleys on the surface of a fingertip are considered as fingerprint. Fingerprints are one of the forms of biometric applied to recognize individuals and verify their identity. Fingerprints recognition system is one the most common biometric authentication systems. Fingerprints remain constant throughout life of a person and it has been used by human for personal identification for many decades [9]. Over the last few decades in the field of

biometric authentication, no two fingerprints have ever been detected to be similar, not even those of identical twins. According to Jain et al. [10] fingerprint recognition or fingerprint authentication refers to the automated technique of verifying a match between two individual fingerprints [10]. Comparison of several features of the print pattern is generally required for the analysis of fingerprints matching. Three basic patterns of fingerprint ridges are the arch, loop, and whirl. Arch: The ridges enter from a side of the finger, move towards the centre making an arc, and then exit the other part of the finger. Loop: The ridges come from one side of a finger, create a curve, and then exit on that same side. Whirl: Ridges create circularly around a central point on the finger.

- Hand geometry is a biometric that uses the geometric shape of the hand with its shape, size of palm, and lengths, widths of the fingers [11] for authenticating a user's identity. This biometric offers a good balance of performance characteristics and is comparatively easy to use. It might be appropriate where there are more users or where users access the system infrequently and are perhaps less disciplined in their approach to the system. Environmental factors, such as dry weather or personal anomalies such as dry skin, do not provide to have any negative effects on the authentication accuracy of hand geometry-based methods. The geometry of the hand is not known to be very unique and hand geometry-based recognition systems cannot be scaled up for techniques requiring identification of an individual from a huge population. Hand geometry information may not be constant during the period of growth of children. In addition, a person's adornments (e.g., rings) may pose further challenges in extracting the correct hand geometry information.
- An iris scan presents an investigation of the rings, furrows and freckles in the coloured ring of the eye. The iris is the annular area of the eye surrounded by the pupil and the sclera (white of the eye) on either side. Iris-based biometrics, involve analysing features found in the coloured ring of tissue that surrounds the pupil. The iris patterns are formed six months after birth and become stable after about one year. After that, the patterns remain unchanged for life. The complex iris texture carries very unique information which is useful for personal recognition [12]. Each iris is supposed to be unique and, like fingerprints, even the irises of identical twins are expected to be different. It is very hard to surgically tamper the texture of the iris. Although the early iris-based recognition systems required considerable user participation and were expensive, the new methods have become more accessible and cost-effective. Iris biometrics work with glasses and contact lenses in place and are one of the few devices that can work well in identification mode.
- Retina-based biometrics involve analysing the coating of blood vessels located at the back part of the eye. A recognized technology, this technique involves using a low-intensity light source through an optical coupler to scan the unique patterns of the retina [13]. Retinal scanning can be quite accurate but does require the user to look into a receptacle and focus on a given point. This is not convenient if glasses are used or concerned about having close contact with the reading device. For these argues, retinal scanning is not well accepted by all users.

- Voice is a combination of physical and behavioural biometrics. Voice authentication is not based on voice recognition but on voice-to-print authentication, where advanced technology converts voice into text. Voice biometrics has a good potential for growth; because it needs no new hardware as most PCs already contain a microphone. According to J. P. Campbell [14], features of a person's voice are based on the shape and size of the appendages (e.g., vocal tracts, mouth, nasal cavities, and lips) that are employed in the synthesis of the sound. These physical features of human speech are invariant for an individual, but the behavioural part of the speech of a person changes over time due to age, health conditions (such as cold), emotional state, etc. [14]. Voice is also not very distinctive and may not be appropriate for large-scale identification.
- It is hypothesized that each person types on a keyboard in a characteristic way. This behavioural biometric is not expected to be unique to each individual but it is expected to offer sufficient discriminatory information that permits identity verification [15]. Keystroke dynamics is a behavioural biometric; for some persons, one may expect to detect huge variations in typical typing patterns. Moreover, the keystrokes of an individual using a system could be monitored unobtrusively as that person is keying in information. However, this biometric allows 'continuous verification' of an individual over a period of time.
- Among the various biometric techniques, sclera recognition is considered as one of the important traits. As the sclera area is a highly-protected portion of the eye, it is very difficult to spoof. Identification of a person by the vessel patterns of the sclera is possible because firstly, these patterns possess a high degree of randomness, which is never the same for any two individuals, even for identical twins and this makes it ideal for personal identification. Secondly, the patterns remain stable throughout a person's lifetime [16], these patterns even differ for the right and the left eye of the same individual. Additionally this trait can be easily combined with iris biometrics. It is interesting to note that humans are the only mammals with extensive exposed sclera, which is amenable to imaging of the encompassing conjunctival vasculature. The various challenges in sclera recognition include accurate segmentation of the sclera area, sclera vessel enhancement and the extraction of discriminative features of the sclera vessel pattern for authentication and identification purposes. The task becomes more difficult, as frequently a complete sclera image is not obtained but it is occluded by portions of the eyelid and eyelashes. Moreover different lighting conditions can change the appearance of the texture patterns by accentuating and attenuating various grey tones. Also, the authentication system should work in real-time so that extraction, representation and comparison of texture images should not consume large computational resources. After that, a classification system uses the mathematical model of the sclera texture to compare with other sclera images to identify specific individuals or identify an individual.

## 2 Signature Biometrics

For security and control in recognition of human identity, most biometric identifiers require a special type of device/equipment or sensor system. However, biometric authentication using signatures can be realized with no additional sensor except a pen and a piece of paper. Nakanishi et al. [17] have shown that every human being has limited biometrics and if the biometric data are leaked out or accessed inadvertently, and the identity of the person whose biometrics they belong to is disclosed, they can never be used for authentication again. So, to deal with this problem, cancellable biometric techniques have been introduced. Among various biometric modalities, only the signature is considered cancellable from a viewpoint of spoofing [17]. Even if a signature shape is known by others, it is possible to cope with the problem by changing the shape. Among all of the biometric authentication systems that have been proposed and implemented, automatic handwritten signatures are considered as the most legally and socially accepted attributes for personal identification. The most challenging aspect in the automation of signature-based authentication is the need for obtaining high accuracy results in order to avoid false authorization or rejections.

Handwritten signature authentication is based on systems for signature verification and signature identification. Whether the given signature belongs to a particular person or not is decided through a signature identification system, whereas the signature verification system decides if a given signature belongs to a claimed person or not. Signature-based authentication can be either static or dynamic. In the static mode (referred to as off-line), only the digital image of the signature is available. In the dynamic mode, also called “on-line”, signatures are acquired by means of a graphic tablet or a pen-sensitive computer display.

A signature is a biometric attribute created by a complex process originating in the signer’s brain as a motor control “program”, implemented through the neuromuscular system and left on the writing surface by a handwriting device [18]. Consequently, signature-based identification and verification is also considered as an important authentication technique among all of the most popular biometric-based authentication methods in the area of personal identification.

A signature also has a high legal value, since it has always played a role in document authentication and it is accepted both by governmental institutions and for commercial transactions as a mean of identification. Moreover, contrary to the majority of other biometrics, a signature can be reissued, in the sense that, if compromised, with a certain degree of effort the user can change his signature. On the other hand, it can be influenced by physical and emotional conditions and it exhibits a significant variability that must be taken into account in the authentication process.

Signature verification analyses the way a user signs his/her name [19]. Signing features such as speed, acceleration, velocity, and pressure are as important as the completed signature’s static shape. Signature based authentication enjoys a synergy with existing processes that other biometric based authentication methods do not. People are generally used to signatures as a means of transaction-related identity authentication. Signature verification devices are reasonably accurate in operation and obviously lend themselves to applications where a signature is an accepted personal

identifier. Remarkably, comparatively limited significant signature applications have emerged compared with other biometric methodologies.

Of the many possible biometrics available, the handwritten signature perhaps has the longest history, and is the best established biometric mechanism both for identity, cheque and transaction authorisation, and is the most widely accepted by the general public [20]. In many practical situations there are advantages in using the simple handwritten signature as a means of confirming identity and authorising system access, yet signatures are notoriously variable and difficult to characterise uniquely, are also prone to forgery and misuse, and the technological challenges presented by automatic signature verification are significant [21].

Signature based biometrics authentication is widely used in forensic applications. The goal of forensic study is that of determining whether observed evidence can be attributed to an individual. The main aim of signature based biometric authentication with forensic application is the prevention of crime.

Jain and Ross [22] have shown that an inherent advantage of a signature-based biometric system is that the signature has been established as an acceptable form of personal identification method and can be incorporated transparently into the existing business processes requiring signatures such as credit card transactions. A block diagram of a generic signature authentication system is shown in Figure 3. This figure follows the classical signature verification model steps, that is, data acquisition, pre-processing, feature extraction, comparison (which is usually called 'verification' in the signature identification and verification field) and performance evaluation.

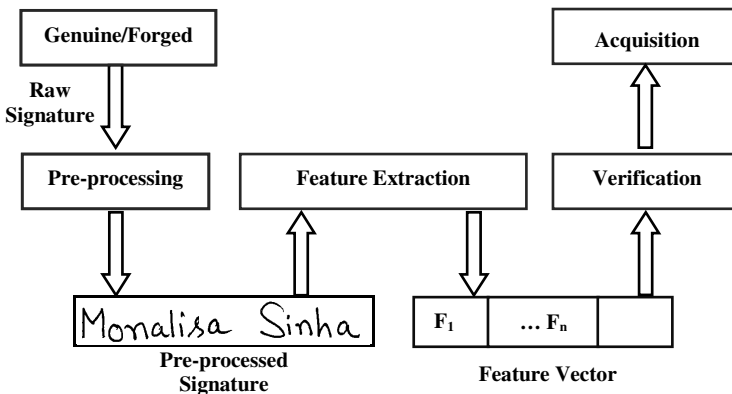


Fig. 3. Block Diagram of an Automatic Signature Verification System

### 3 Signature Verification Concept

Pattern recognition is one of the most important and active fields of research. During the past few decades, there has been a considerable growth of interest in problems of pattern recognition. In the last few years many methods have been developed in the area of pattern recognition.

Signature verification is a significant area of study in the field of pattern recognition. Many techniques of verification have been built up specifically to address the off-line signature verification problem. In general, to deal with the problem of off-line signature verification, researchers have investigated a commonly used approach which is based on analysing two different patterns of classes, class 1 and class 2, where class 1 represents the genuine signature set, and class 2 represents the forged signature set. When the performance of the off-line signature verification system is calculated, usually two types of errors [23] are considered: the False Rejection, which is called a Type-1 error and the False Acceptance, which is called a Type-2 error. Hence, there are two types of error rates: False Rejection Rate (FRR) which is the percentage of genuine signatures treated as forgeries, and False Acceptance Rate (FAR) which is the percentage of forged signatures treated as genuine. The Average Error rate (AER) is the average of FAR and FRR. When we deal with the experiments of a system, we must make a trade-off between FRR and FAR based on the application and other aspects of where and how the system is used. Conversely, if the decision threshold of a system is set to have the percentage of false rejections approximately equal to the percentage of false acceptances, the Equal Error Rate (EER) is calculated. During the enrolment phase, the input signatures are processed and their personal features are extracted and stored into the knowledge base. During the classification phase, personal/salient features extracted from an accepted signature are compared against the information in the knowledge base, in order to judge the authenticity of the applied signature.

According to Ismail et al. [24], an automatic signature verification system should meet the following requirements:

- Reliability: The forgeries should be rejected and the genuine signatures should be accepted if there is adequate distinction between the input samples and the original patterns.
- Adaptability: Genuine signatures should be accepted even with slight variations
- Practicality: It is possible to implement such systems in real-time.

## 4 Multi-script Signature Verification Concept

Although significant research has already been undertaken in the field of signature-based authentication, particularly when single-script signatures are considered, however conversely, less attention has been devoted to the task of multi-script signature-based authentication. In the signature-based personal identification and verification area, introduction of multi-script challenges is a very recent concept and a novel scheme.

As a multi-script and multi-lingual country, India doesn't have the concept of a single language. The country has a set of official regional scripts and languages recognized for some of its individual states for official communications. There are ten major scripts in India for the documentation of its official languages. They are Devanagari, Bangla, Gurumukhi, Gujarati, Oriya, Kannada, Telugu, Tamil, Malayalam and Urdu (Nastaliq). Most of the Indian scripts have originated from an



ancient script called Brahmi through various transformations [25]. Devanagari script is being used for writing many languages namely Hindi, Marathi, Nepali, Sanskrit, Konkani, Maithili, Santali, Sindhi, and Kashmiri. Hindi, written in Devanagari script, is the national language of India.

When a country deals with two or more scripts and languages for reading and writing purposes, it is known as a multi-script and multi-lingual country. Multilingualism is a widespread phenomenon as there exist more than 6500 languages around the world. Most countries have only a single language but very few countries have more than one script for reading and writing purposes. In India, there are officially 23 (Indian constitution accepted) languages and 11 different scripts. In such a multi-script and multi-lingual country like India, languages are not only used for writing/reading purposes but also applied for reasons pertaining to signing and signatures. In such an environment in India, the signatures of an individual with more than one language (regional language and international language) are essentially needed in official transactions (e.g. in a passport application form, an examination question paper, a money order form, bank account application form etc.). To deal with these situations, signature verification techniques employing single-script signatures are not sufficient for consideration. Consequently in a multi-lingual and multi-script scenario, signature verification methods considering more than one script are in great demand.

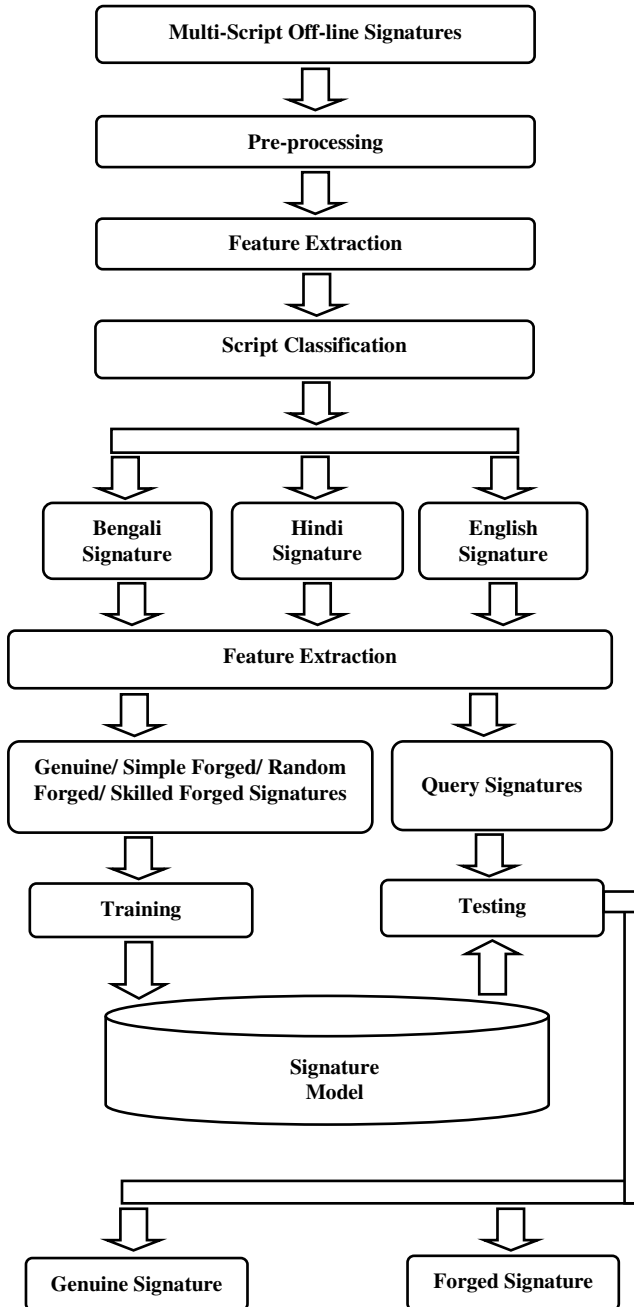
Development of a general multi-script signature-based authentication system, which can verify the identity of people using signatures of all scripts, is very complicated and it is not possible to develop such a method in the Indian scenario. The verification accuracy in such multi-script signature environments will not be desirable compared to single script signature verification. To achieve the necessary accuracy for multi-script signature authentication, it is first important to identify signatures based on the type of script and then use individual single script signature verification for the identified signature script.

India is a union of 29 states and most of the regions use three different languages (international language, national language and regional language). West Bengal is a state where Bangla (local language), Hindi as a national language and English as an international language are generally used for official transactions. A generic multi-script signature verification system considering these three different scripts of signatures (Bangla, Hindi and English) is shown in Figure 4.

## 5 Classification of Biometric Signatures

For increasing the reliability of biometric-based authentication methods and systems, different groups of biometric signatures have been undertaken in research and scientific discussion to make the authentication systems more protected.

Arslan Bromme [26] has presented that the usage of biometric signatures within the biometric enrollment, authentication and de-enrollment processes shows mainly two classes of biometric signatures in use for mono-modal biometric processes: i. mono-modal biometric signatures for single biometric signatures and ii. mono-modal biometric templates representing sets/classes of single biometric signatures.



**Fig. 4.** Multi-script Signature Verification System

Taking multimodality into account, two more classes of biometric signatures are considered: multi-modal biometric signatures as lists of mono-modal biometric signatures for more than one biometric method used, and multi-modal biometric templates for lists of mono-modal biometric templates.

On the superset level with regard to mono-modal or multi-modal biometric processes for changing environmental conditions or changing biological characteristics (e.g. aging), other high level classes will arise: mono-modal biometric multi-templates for sets of mono-modal biometric templates and multi-modal biometric multi-templates for lists of mono-modal biometric multi-templates.

On the other hand, different combinations of biometric traits have also been taken into account in research to make the authentication systems more secure. Biometric systems using a single biometric trait either for identification or for verification is called a unimodal biometric system [27]. According to Teddy Ko [28] an unimodal biometric authentication system sometimes fails to be accurate enough for the identification of a large user population due to some problems such as noisy dataset, non-universality, limited degrees of freedom, spoof attacks, intra-class variations, and undesirable error rates.

However, no biometric trait is truly universal [29]. Biometric systems based solely on a single biometric may not always meet security requirements. Thus multi-biometric systems are emerging as a trend which helps in overcoming limitations of single biometric solutions. So the problems associated with unimodal biometric systems can be overcome by multimodal biometric systems. A multimodal biometric system unifies the information presented by multiple biometric sources. Multiple biometric sources include multiple sensors, multiple instances, multiple samples, multiple algorithms, or multiple biometric traits [30].

## 6 Signature Stability

Stability analysis of handwritten signatures is very relevant for automatic signature verification and this is also a very important characteristic for investigating the intrinsic human properties related to the handwriting generation processes concerning human psychology and biophysics. Each handwritten signature strongly depends on a large number of factors such as the psychophysical state of the signer and its social and cultural environment as well as the conditions under which the signature apposition process occurs [31]. In addition, its study can provide new insights for a more accurate treatment of signatures for verification purposes, hence contributing to the design of more effective signature verification systems. For these reasons, it is not surprising that the scientific community has been devoting much effort to the analysis of signature stability.

A lot of approaches estimate signature stability by the analysis of a specific set of characteristics, when dynamic signatures are considered. In general, these approaches have shown that there is a set of features which remain stable over long time periods, while others can change significantly in time [32]. More precisely, a comparative study of the consistency of certain features of dynamic signatures has demonstrated

that position, velocity and pen inclination can be considered to be among the most consistent, when a distance-based consistency model is applied [33]. Other results, based on personal entropy, demonstrated that position is a stronger characteristic than pressure and pen inclination in both short and long-term variability. Moreover, although pressure may give better performance results in a short-term context, it is not recommended for signature verification in the long-term.

When static signatures are considered, the degree of stability of each region of a signature can be estimated by a multiple pattern-matching technique [34]. The basic idea is to match corresponding regions of genuine signatures in order to estimate the extent to which they are locally different. A preliminary step is used there to determine the best alignment of the corresponding regions of signatures, in order to diminish any differences among them.

Although stability has been observed in signatures, the signing process does not lend itself to the production of repeatable, perfectly accurate and identical characteristic data issued from successive trials. The only certainty in this domain is that when two signatures are identical and one of them is a forgery, i.e. probably a copy. In fact, a great deal of variability can be observed in signatures, depending on country, age, time, habits, psychological or mental state, physical and practical conditions. Two types of signature variability have to be clearly distinguished: intraclass or intrapersonal variability, i.e. the variation observed within a class of genuine signature specimens of one individual, interclass or interpersonal variability, i.e. differences which exist between genuine signature classes produced by two different writers. In theory, intraclass scatter must be as low as possible and interclass scatter extensive enough to be used for class separation.

The stability comes from the intrinsic properties of rapid human movements that somehow constitute the basic element of each signature [32]. In fact, a number of major psychophysical phenomena have been observed on a regular and consistent basis during the study of these movements. The most remarkable is without doubt what is known as the invariance of velocity profiles. A technique for the analysis of stability in static signature images has been presented by Impedovo et al. [35]. The technique uses an equimass segmentation approach to non-uniformly split signatures into a standard number of areas. Consecutively, a multiple matching technique is adopted to estimate stability of each area, based on cosine similarity.

## 7 Signature Verification vs. Identification

In the field of automatic signature recognition, two different types of signature recognition systems are considered such as signature verification and signature identification systems. Signature-based biometric technologies are used for either one of those two purposes, verification or identification, and the implementation and selection of the technology and related procedures are closely tied to this aim. Technologies differ in their capabilities and effectiveness in addressing these purposes. Verification (Am I whom I claim I am?) includes confirming or rejecting a person's demanded identity. In identification, someone has to establish a person's

identity (Who am I?). Each one of these approaches has its own complexities and could probably be solved by a signature authentication system. These two examples illustrate the difference between the two primary uses of biometrics: identification and verification.

**Signature identification (1:N, one-to-many, recognition):** A signature identification system must recognise a signature from a list of N signatures in the template database. The process of determining a person's identity by performing matches against multiple templates. Identification systems are designed to determine identity based solely on signature information.

**Signature Verification (1:1, matching, authentication):** A signature verification system simply decides whether a given signature belongs to a claimed signature or not. It is the process of establishing the validity of a claimed identity by comparing a verification template to an enrollment template. Verification needs that an individuality be claimed, after which the individual's enrollment template is located and compared with the verification template. Verification responses the query, 'Am I who I claim to be?'

## 8 Dynamic and Static Signature Verification

The biomechanical processes involved in the production of the human signature are very complex. In vastly simplified terms, the main excitation is thought to take place in the central nervous system, more specifically in the human brain, with predefined intensity and duration describing the intent of the movement. The signal of the intent (or the movement plan) is passed through the spinal cord to the particular muscles which are activated in the intended order and intensity. As a result of such activation and relaxation of the muscles and whilst holding a pen, the resultant arm movement is recorded in the form of a trail on paper as a handwritten signature.

Based on the handwritten signature data acquisition method, two types of systems for handwritten signature verification can be identified: static (off-line) systems and dynamic (on-line) systems.

### 8.1 Dynamic Signature Verification

Whenever handwriting is captured as a user writes for the purpose of recognition or analysis, it is called on-line handwriting recognition. This process requires special devices, such as stylus or digitizer pen and tablet, to capture the writing information on-the-fly. The temporal stream of information which is extracted as the writing is produced is called on-line features, which include local pressure, acceleration, speed, number of strokes, and order of strokes. The signature image can be simulated with high accuracy using this temporal-spatial feature information.

Dynamic signature verification system uses a digitizer or an instrumented pen to give a representation of the written signature generating one or several signals which vary with time. The raw data are then pre-processed to remove spurious information, to filter the significant signals and to validate the acquisition process. The next step

involves what is referred to as the feature extraction process. Specific and discriminant functions or parameters are computed from the filtered input data and are used to represent a signature.

Dynamic signature verification methods can be classified in two principal groups. In the first group of dynamic signature verification, the techniques deal with functions as features. In this case, the complete signals (i.e. position, pressure, velocity, acceleration vs. time, etc.) are regarded as, mathematical time functions where the values directly constitute the feature set [36]. In the second group of dynamic signature verification, the techniques refer to several parameters as features. These parameters are computed from the measured signals. Both global and local information are either explicitly or implicitly taken into explanation, separately or together.

- Number of strokes: This feature is the total number of lines contained in the entire signature. One line is from the time since the signer put down the pen to the contact surface until it is filed or until pen-up occur.
- Number of pen-ups: This feature shows how many times writer picked up a pen during signing a signature. It should be noted that the last lifting of pen is not counted because it marks the end of the signing.
- Signature aspect ratio: This feature considers the width of the signature (signature size on the x-axis) expressed in pixels of a tablet and normalized on pixels of the screen and the height of a signature (the size of signatures on the y-axis) expressed in the same way that puts them in proportion. The assumption is that the user will sign each time the same in terms of creating a signature in one, two or more lines and that the size of signatures each time will be approximately the same.
- Signing time: Feature expresses the total time needed to get a person to sign, usually in milliseconds, since the beginning of the signing. It is assumed that the time for a trained signature will always be nearly equal.
- Time-down ratio: It describes how much of total signing time the pen was in contact with the signing surface. To a person who has trained signature, this characteristic will be fairly constant because it is directly related to the signing time.
- Time-up ratio: This feature opposite to Time-down ratio, and indicates how long of total signing time a pen was separate from the signature area. This feature is used for authentication algorithm, which may favour one of two opposing feature in relation to represent a more stable signature characteristic of the person.
- Signature speed: This feature is derived from the total length of the signature and the time in which the pen was in direct contact with the signing area. It tells the speed of signing expressed in pixels per millisecond. Trained signature should not have significant differences over the time. However, this feature strongly depends on the physical and mental condition of the person.
- Velocity along the x-axis: This feature represents speed expressed in number of pixels per millisecond, which indicates how quickly people sign if only the x coordinate is considered in the system. It calculates the total length which pen passed along the x-axis and is divided by the total time in which the pen was

lowered to the signing area. This feature depends on the physical and mental condition of the person and is often used less than other characteristics.

- **Velocity along the y-axis:** Here the feature represents speed expressed in number of pixels per millisecond, which indicates how quickly people sign if only the y coordinate is considered in the system. It calculates the total length which pen passed along the y-axis and is divided by the total time in which the means for writing was lowered to the signing area. This feature depends on the physical and mental condition of the person and is often used less than other characteristics.
- **Average pressure:** This feature is obtained by monitoring the level of pressure which pen leaves on the signing area. In order to obtain the average pressure it is necessary to add up all levels of the received pressure to one variable and divide by the total number of packages. The biggest influence on this characteristic has signers body.
- **Strongest pressure moment:** This feature can be characterized as the only local characteristics of signatures which can be global as it is unique in the entire signature. In order to extract this feature, monitoring the level of pressure which pen leaves on the signing area is needed. The highest level is observed and the time of its creation is recorded. It is assumed that the signature always has nearly the same moment of the strongest pressure.
- **Speed:** When a signature is captured with a digitizer, the pen motions (dynamics) are recorded. According to Zimmerman et al. [36], when signing, the hand can operate in a rule known as ballistic movement, where the muscles are not controlled by sensual feedback. Ballistic motions are usually fast, practiced motions whose accurateness rises with speed [36]. In the on-line signature there is an significant feature that can be extracted, which is the speed of the signature. During the signing process, the speed of the pen ball is changing at every point of the signature. These changes are repeated in a fixed way every time a person signs again. To find out the speed of the signature it is needed to record the time at which a specific point is sampled. Here,  $\text{speed} = \text{Distance} / \text{Time}$ .
- **Acceleration:** Acceleration produced by pen movements while one is writing or signing provide useful information for handwriting research, particularly for applications like automatic signature verification. Measurement of pen acceleration is usually done with accelerometers integrated into a pen or with devices that either derive pen acceleration from other physical measures or sense physical quantities equivalent to pen acceleration [37]. Acceleration signals are characterized in terms of phase, amplitude and frequency. This characterization makes possible the extraction from the accelerometer output those signal components relevant to the handwriting process.

## 8.2 Static Signature Verification

When the recognition is undertaken using only the static images of handwriting, the process is called off-line recognition. Despite the unique advantage over its on-line counterpart, as no specialized capture device is required, the amount of information obtained from off-line recognition is two orders greater, but much less meaningful

and more difficult to interpret. Moreover, the traces of dynamic information are very difficult to compute. Traditionally, the recovery of such information requires professional skills and techniques whose implementation on computers is not easy [38]. Several evaluations performed by expert document analysts concluded that the detection of forgeries of high skill require not just static information but also dynamic information, a survey by Plamondon and Lorette [37] reported. Some rare attempts to extract direct pressure information were made by Ammar et al. [40]. With the distinct characteristics mentioned above, on-line recognition systems are able to achieve better results than their off-line counterparts [41].

Off-line systems utilize the classic method of on-paper signatures for person verification. The signature obtained is digitized by an optical scanner or camera. An alternative is to input the image through a tablet or any other suitable device. Subsequently, respective applications determine the match of the person's signature with a reference sample by comparing the overall trace (image) of the signature. Based on this particular principle, the current very unreliable methods, commonly practiced in banking and retail for example, are utilized to verify handwritten signatures, relying on the human factor in the form of a calligraphy expert.

## 9 Types of Forgeries

There are usually three different types of forgeries to take into account. According to Coetzer et al. [42], the three basic types of forged signatures are indicated below:

- Random forgery: The forger is not familiar and has no access to the genuine signature (not even the author's name) and reproduces a random one.
- Simple forgery: The forger is familiar with the author's name, but has no access to a sample of the signature.
- Skilled forgery: The forger has access to one or more samples of the genuine signature and is able to reproduce it. But based on the various skilled levels of forgeries, it can also be divided into six different subsets.

The paper [43] shows various skill levels of forgeries and these are shown below.

- A forged signature can be another person's genuine signature. Justino et al. [44] categorized this type of forgery as a Random Forgery.
- A forged signature is produced with the knowledge about the genuine writer's name only. Hanmandlu et al. [45] categorized this type as a Random Forgery whereas Justino et al. [44] categorized this type as a Simple Forgery. Weiping et al. categorized this type as a Casual Forgery [46].
- A forged signature imitating a genuine signature's model reasonably well is categorized as a Simulated Forgery by Justino et al. [44]
- Signatures produced by inexperienced forgers without the knowledge of their spelling after having observed the genuine specimens closely for some time are categorized as Unskilled Forgeries by Hanmandlu et al. [45]



- Signatures produced by forgers after unrestricted practice by non-professional forgers are categorized as Simple Forgery/Simulated Simple Forgery by Ferrer et al. [47], and a Targeted Forgery by Huang and Yan [48].
- Forgeries which are produced by a professional imposter or person who has experience in copying Signatures are categorized as Skilled Forgeries by Hanmandlu et al.[45]

## 10 Related Work on Signature-Based Biometric Authentication

Many techniques have been developed in the field of signature-based biometric authentication. Some examples of biometric verification and identification approaches and optimised schemes are discussed below:

### 10.1 Single-script Signature-Based Biometric Authentication

Arslan Bromme [26] has shown that every human being has static, dynamic, physiological and behavioural biological characteristics, which can be used for biometric person recognition. Handwritten signatures are one of the behavioural biological characteristics. Biometric signatures can be used for classes of biometric systems which are similar to those used within the core processes of biometric authentication systems. The main objective of that paper was the classification of biometric signatures.

Rabasse et al. [49] described a method for the generation of synthetic handwritten signatures, in the form of sequences of time-stamped pen data channels, for use in on-line signature verification experiment. The method presents modelled variability within the generated data based on variation that is naturally found within genuine source data. Experimentation using the SVC2004 [50] dataset and a commercial signature verification engine shows that the synthesized data achieves comparative verification performance to the use of genuine data. The method uses two seed signatures from a signer with captured data in the form of time stamped vectors. Rather than a simple interpolation between the two seed signature, our method deploys an intelligent mapping and introduces naturally occurring variability within each signature. Derivative Dynamic Time Warping (DTW) to find minimum Euclidean edit distance between points within two seed signatures.

A new proposal for score normalization in biometric signature recognition based on client threshold prediction was proposed by Pascual et al. [51]. The use of score normalization in biometric based recognition system is a very important part, particularly in those based on behavioural traits, such as written signature. The score normalization techniques can be classified as: i) Test dependent and ii) Target dependent. The first is used mainly in speaker verification, while the second approach is use for signature verification techniques. This work focuses on target dependent techniques.

Biometric security system based on signature verification using neural networks is presented by Kumar et al. [52]. The global and grid features are combined to generate

new set of features for the verification of signature. The Neural Network is also used as a classifier for the authentication of a signature. Random, unskilled and skilled signature forgeries along with genuine signatures were considered for performance analysis of the system. Some common global features such as i. Aspect Ratio, ii. Signature Height, iii. Image Area, iv. Pure Width and v. Pure Height have been used for the experiments. A number of 600 signature samples collected from 20 signers were considered for their experiments.

Maiorana et al. [53] proposed a signature-based biometric authentication system, where water marking techniques have been used to embed some dynamic signature features in a static representation of the signature itself. A multi-level verification scheme, which is able to provide two different levels of security, has been obtained. These proposed watermarking techniques are based on the properties of the Radon transform which well fits to the signature images. In order to test the authentication performances of this approach, 50 signatures have been acquired from each of 30 users, taking for each of them 10 signatures in five different sessions during a week time span. Some approaches for the protection and authentication of biometric data using watermarking have been proposed in another report [54] where robust watermarking techniques are used to embed codes or timestamps.

Another approach [55] discusses a protected on-line signature-based biometric authentication system, where the biometrics considered are secured by means of non-invertible alterations, able to produce templates from which retrieving the original information is computationally as hard as random guessing it. The benefits of using a protection technique based on non-invertible transforms are exploited by presenting three different matching strategies in the converted domain, and by suggesting a multi-biometrics method based on score-level fusion to improve the performances of the considered system. The experiments were evaluated on the public MCYT signature database.

Another on-line signature-based biometric authentication system is presented by Maiorana et al. [56]. In this proposed technique, the non-invertible transformations are applied to the acquired signature functions, creating impossible to derive the original biometrics from the kept templates, while keeping the same recognition performances of an unprotected method. Specifically, the possibility of producing cancelable templates from the same original dataset, thus offering a proper solution to privacy concerns and security issues, is intensely explored.

Nagasundara et al. [27] presented an authentication approach based on hand geometry, palmprint and signature. The aim of that paper is to exploit the best possible combinations of hand geometry, palmprint and static signatures for multimodal biometric systems by integrating the information at score level fusion. Primarily, Zernike moments are extracted for each biometric trait of a person and study the identification accurateness. Consequently, the effect of identification accuracy using score level fusion of multiple traits of a person is investigated. Experimentations are accompanied on GPDS hand geometry dataset, PolyU two dimensional palmprint dataset and UOM offline signature database to assess the actual advantage of the fusion of multiple biometric traits performed at score level fusion.

In an approach by Mhatre and Maniroja [57] a signature based authentication by using two different algorithms was introduced. Before extracting different features from the signature, some pre-processing of the signature is performed. In pre-processing, the signature is colour normalized and scaled into a standard format. The process is pretty different and it deals with extraction of features based on moment, standard deviation and mean. The process uses Euclidean distance classifier for comparing test signature with database. The algorithm has shown promising results while dealing with random forgeries and simple forgeries; also it gives good recognition rate.

Maiorana [58] introduced a set of noninvertible conversions, which can be employed to any biometrics whose template can be represented by a set of sequences, in order to produce multiple transformed versions of the template. Once the transformation is made, recovering the original data from the transformed template is computationally as hard as random guessing. As a proof of perception, the suggested method is applied to an on-line signature recognition scheme, where a hidden Markov model-based matching approach is applied. The performance of a secured on-line signature recognition system employing the proposed BioConvolving approach is calculated, both in terms of verification rates and renewability capacity, employing the MCYT signature dataset. The reported extensive set of experimentations showed that protected and renewable biometric templates can be properly generated and used for recognition.

## 10.2 Multi-script Signature-Based Biometric Authentication

A different signature verification technique considering multi-script signatures has been proposed by Pal et al. [59]. This multi-script signature verification method involving English and Hindi signatures is very significant in multi-script signature environment. This multi-script signature identification and verification technique has never been used for the task of signature verification and this task was the first report in signature verification area. In that paper, the multi-script signatures were identified first on the basis of signature script type and afterward verification experiments were investigated based on the identified script result. Two different results for identification and verification were calculated and analysed.

In another approach by Pal et al. [60] the performance of signature script identification was reported. An experiential contribution towards the understanding of multi-script signature identification was presented. In that proposed signature identification technique, the signatures of Bengali (Bangla), Hindi (Devanagari) and English are considered for the identification process. The aim of that paper was to identify whether a claimed signature belongs to the group of Bengali, Hindi or English signatures. In a multi-script signature verification environment, signature script identification plays an important role. If the signatures are identified based the script used for writing signatures, subsequently the individual signature verification can be done based on the identified script result. Zernike Moment and histogram of gradient were employed as two different feature extraction methods. In the proposed scheme, Support Vector Machines (SVMs) were considered as classifiers for signature identification.

In another report by Pal et al. [61] a script identification scheme of signatures was investigated. In their paper, a technique for a bi-script off-line signature identification method is proposed. In this signature identification system, the signatures of English and Bengali (Bangla) are considered for the identification procedure. Different features like, modified chain-code direction features, under-sampled bitmaps and gradient features computed from both background and foreground components are employed for this purpose. SVMs and Nearest Neighbour (NN) techniques are considered as classifiers for signature identification in the proposed scheme. A dataset of 1554 English signature samples and 1092 Bengali signature samples are used to generate the experimental results. Different results based on different features are calculated and analysed.

An investigation of the performance of a signature identification system involving English and Chinese off-line signatures was presented by Pal et al. [62]. In that paper, a foreground and background based technique was proposed for identification of scripts from bi-lingual (English/Roman and Chinese) off-line signatures. The aim of the system was to identify whether a claimed signature belongs to the group of English signatures or Chinese signatures. The identification of signatures samples based on its script is a major contribution in a multi-script signature verification environment. Two background information extraction techniques were used to produce the background components of the signature images. Gradient-based technique was used to extract the features of the foreground as well as background components. Zernike Moment feature was also used on signature samples. (SVMs are used as the classifier for signature identification in the proposed system.

A two-stage approach for English and Hindi off-line signature identification and verification was proposed by Pal et al. [63]. The main aim of their approach was to demonstrate the significant advantage of signature script identification in a multi-script signature verification environment. In their proposed signature verification technique the performance of a multi-script off-line signature identification system, considering a joint dataset of Hindi and English signatures, was initially investigated and subsequently a verification task was explored separately for English signatures and Hindi signatures based on the identified script result. The gradient feature, water reservoir feature, loop feature and aspect ratio were employed and SVMs were considered for verification.

An experimental contribution in the direction of multi-script off-line signature identification and verification using a novel technique involving off-line English, Hindi (Devnagari) and Bangla (Bengali) signatures is introduced by Pal et al.[64]. In the first stage of the proposed signature verification technique, the performance of a multi-script off-line signature verification scheme, considering a joint dataset of English, Hindi and Bangla signatures, was investigated. In the second stage of experimentation, multi-script signatures were identified based on the script type, and subsequently the verification task was explored separately for English, Hindi and Bangla signatures based on the identified script result. The chain code and gradient features were employed, and Support Vector Machines (SVMs) along with the

Modified Quadratic Discriminate Function (MQDF) were considered in this scheme. From the experimental result achieved, it is noted that the verification accuracy obtained in the second stage of experiments (where a signature script identification method was introduced) is better than the verification accuracy produced following the first stage of experiments. Experimental results indicated that an average error rate of 20.80% and 16.40% were obtained for two different phases of verification.

## 11 Signature Database Availability

Although research into signature verification has been pursued for several decades, there has been only a limited number of standard public off-line signature databases created. The scarcity of standard databases has partly resulted from the privacy aspect of the collection of handwritten signatures and a number of constraints that a standard database should meet. Due to the lack of availability of significant and public signature databases, developments of signature verification systems have been negatively affected. It has been difficult to make a significant comparison among different approaches presented in the literature due to the use of custom databases by researchers, and that these databases are not publicly available. The design and construction of an off-line signature corpus involves a long and complex procedure in which aspects such variability of drawing surface, changes of the writing instrument, differences between sessions, number of signers, number of genuine signs per person, forgery procedure and number of forgeries per person, etc. should be taken into account [65]. It is not easy to build a corpus which has considered all the above-mentioned variables mainly due to the difficulties in recruiting appropriate signers.

One of the major problems that can be found in the performance evaluation of signature verification systems, in both identification and verification modes, is the lack of publicly available large signature databases. The quality of available datasets also differs, as there has been no standard collection procedure. Besides, it is very costly to create a large corpus with different types of forgeries, especially skilled forgeries. So, the research in automatic signature verification has long been constrained by the limited availability of a standard database. Presently there are only a few publicly-available databases. Some of them are summarized in Table 1.

**Table 1.** Handwritten Signature Corporuses Available

Corpus Name	Signers	Genuine	Forgeries
GPDS signature [66]	160	24	30
SVC2004 [50]	40	20	20
MCYT-100 [67]	100	25	25
MCYT-75 [68]	75	15	15
GPDS signature [65]	960	24	30

## 12 Signature Data Acquisition and Pre-processing

On the basis of the handwritten signature data acquisition method, two types of systems for handwritten signature verification have been identified (as mentioned previously): static (off-line) systems and dynamic (on-line) systems. In static mode, handwritten signature data is converted to digital form by scanning the signature from the signature collection paper. In this mode, the handwritten signatures are represented as a gray level image. On the other hand, one can deal with the signature data acquisition in online method by using a special pen on an electronic surface. The most conventional online data acquisition devices are digitizing tablets [36]. Electronic pens are also able to detect position, velocity, acceleration, pressure, pen inclination, and writing forces etc.

Once the signature has been acquired, either off-line or on-line, some pre-processing techniques are usually needed. The pre-processing step is vital in order to ensure that only the desired data is fed to the feature extraction module. Normally, acquired signature images are of different formats and resolutions and need to be processed to enable accurate feature extraction. The acquired images may contain unexpected marks, stains, or noise which would cause negative effects on the recognition accuracy. Pre-processing includes steps eliminating such noise and converting the image to a suitable format for feature extraction. Other important pre-processing techniques (signature size normalization, binarization, thinning, smearing, skew correction, skeleton extraction) are also considered in static signatures for accurate feature extraction. Typical pre-processing algorithms for dynamic signature verification involve filtering, noise reduction and smoothing. Another vital pre-processing step that strongly influences all the successive phases of signature verification in both static and dynamic modes is segmentation [69]. Some of the pre-processing steps are carried out in the following sub-steps.

- **Thinning:** This is the transformation of a digital image into a simplified form, but the image should be topologically equivalent. It is a kind of topological skeleton, but computed by means of mathematical morphology operators that are used to remove selected foreground pixels from binary images.
- **Filtering process:** Generally, digital image might contain speckles, smears, scratches or other forms of unwanted noise that might thwart feature extraction. Thus, median filtering is used to eliminate the existing noises.
- **Binarization:** The process by which the image is converted into black and white is called binarization.
- **Width normalization:** All signature images have been reduced to a standard size so as to ease the process of feature extraction.

## 13 Feature Extraction Techniques

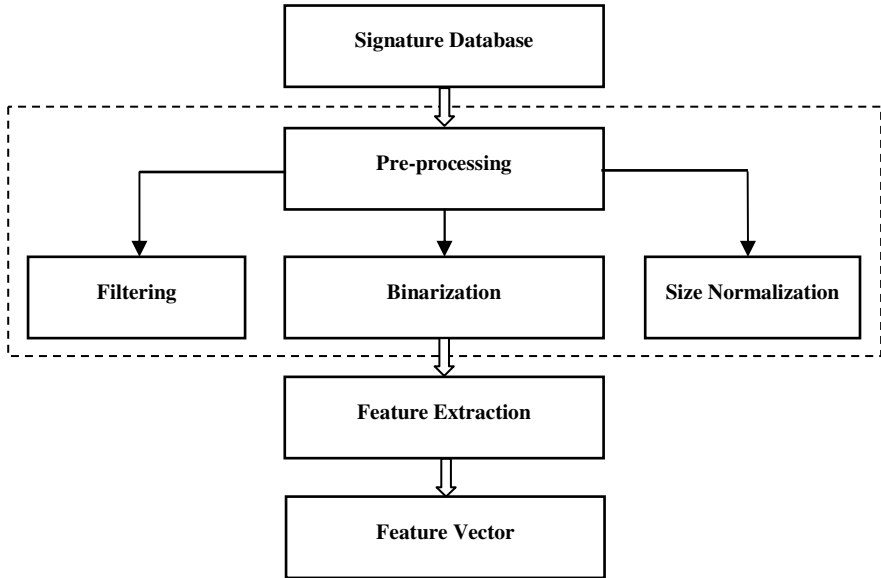
Feature Extraction is an important part of any pattern recognition system. The process in which digital information is modified, simplified, combined so that the salient

information can be classified, is called feature extraction [70]. To be successful, a feature extraction technique should be justifiable using rules that govern the formation of the class of pattern being considered. As features are refined as inputs for the learning process and the decision process, feature extraction techniques are crucial to the success of the whole process of automated pattern recognition [71]. Good features are those that enable the system to identify a pattern's class with the least amount of errors. Baltzakis and Papamarkos [72] commented that the selection of features must be appropriate for the application and the approach. Klement et al. [73] summarized the three requirements that concerned the feature selection process: (i) Speciality (minimizing intra-class variability and maximizing inter-class variability); (ii) Universality (can be applied to any writer); (iii) environmental independence (with respect to writing instruments and materials). In other words, it is essential that a feature extraction technique could minimize or even eliminate the negative effects from variations such as rotation, shift, or dilation of the pattern being considered.

In general, two types of features can be considered for signature verification: i. parameter-based features ii. function-based features. In the case of function-based features, [74] signatures are usually characterized in terms of a time function and the values of the time function constitute the feature set. Conversely, when parameter features [75] are considered, the signatures are characterized as a vector of elements; each one represents the value of a feature. It has been shown by Plamondon and Lorette [76] that function features generally provide better performance as compared to parameter features, but they usually need time-consuming procedures for matching. In addition, parameters are generally grouped into two main categories: i. global parameters and ii. local parameters. The whole signature is considered for global parameters. Usual global parameters are total time duration of a signature, number of pen ups and downs, number of components, global orientation of the signature, etc. Local parameters concern features extracted from a few exact parts of the signature.

### 13.1 Feature Vector Generation

In computer vision and image processing the concept of feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application. A feature vector is an  $n$ -dimensional vector of numerical features that represent some object. The flowchart in Figure 3 shows the process of feature vector generation [77]. It consists of mainly two steps, pre-processing and feature extraction. As previously mentioned, pre-processing is performed on the signature images from a database so as to prepare it for the process of feature extraction and to ensure that all the signature images are of the same dimensions so that it is easier and convenient to extract the features. A flowchart of feature vector generation is shown in Figure 5.



**Fig. 5.** Feature Vector Generation Flowchart

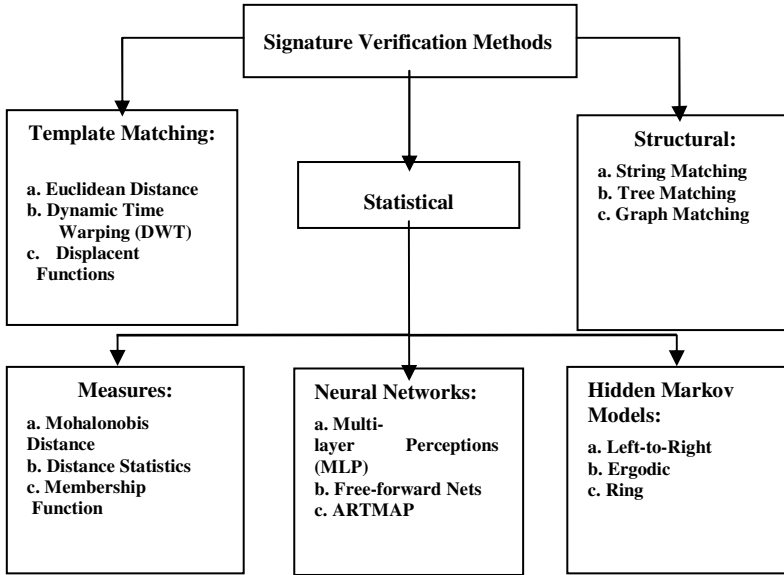
## 14 Classification

In the signature verification method, the authenticity of the test signature is evaluated by matching its features against those stored in the knowledge base developed during the enrolment stage. This process generates a single response that states the authenticity of the test signature samples. Some of the most relevant approaches to signature verification as mentioned in [69] are shown in Figure 6.

When template matching methods are considered in verification, a questioned signature sample is matched against templates of authentic/forged signatures. Dynamic Time Warping (DTW) is used for the most common approaches in this situation for signature matching. DTW is a Template Matching technique used for measuring similarity between two sequences of observations. DTW allows the compression or expansion of the time axis of two time sequences representative of the signatures to obtain the minimum of a given distance value [78].

An Artificial Neural Network (ANN) is a massively parallel distributed system composed of processing units capable of storing knowledge learned from experience (examples) and using it to solve complex problems. The ANN-based approaches have been widely used for a long time in signature verification area, due to their learning and generalizing capability [79].





**Fig. 6.** Signature Verification Techniques (Classification approaches)

In recent times, more attention has been dedicated to the use of Hidden Markov Models for both offline and online signature verification. These models have found to be well suited for signature modelling since they are highly adaptable to personal variability [80].

SVMs are another promising statistical approach to signature verification. SVMs are a relatively new classification technique in the field of statistical learning theory and they have been successfully applied in many pattern recognition approaches. An SVM can map input vectors to a higher dimensional space in which clusters may be determined by a maximal separating hyper plane. SVMs have been used successfully in both offline [66] and online signature verification [81].

## 15 Conclusions

This chapter presented a detailed study on signature-based biometric authentication. Automatic signature verification is a very interesting area of research from the scientific point of view. In recent years, along with the continuous enhancement of security requirements, the field of automatic signature-based authentication is being explored with renewed interest. Up to date outcomes achieved in worldwide competitions using benchmark databases have confirmed that signature authentication systems can have an accuracy level similar to those achieved by other biometric systems [82]. Although, a significant amount of work has been undertaken in order to solve the authentication problem, there are still many challenges to be faced. Hence in this Chapter a detailed description of signature-based biometric authentication has

been presented and hopefully it will be helpful to the researcher as reference materials.

## References

- [1] Boyer, K.W., Govindaraju, V., Ratha, N.K.: Introduction to the Special Issue on Recent Advances in Biometric Systems. *IEEE Trans. Systems, Man, and Cybernetics, Part B* 37(5), 1091–1095 (2007)
- [2] Samal, A., Iyengar, P.A.: Automatic Recognition and Analysis of Human Faces and Facial Rexpressions: A Survey. *Pattern Recognition* 25(1), 65–77 (1992)
- [3] Jain, A.K., Hong, L., Pankanti, S.: Biometric Identification. *Communications of the ACM* 43(2), 91–98 (2000)
- [4] Zhang, D., Campbell, J.P., Maltoni, D., Bolle, R.M.: Special Issue on Biometric Systems. *IEEE Trans. Systems, Man, and Cybernetics, Part C* 35(3), 273–275 (2005)
- [5] Wayman, J.L., Jain, A.K., Maltoni, D., Maio, D.: *Biometric Systems: Technology, Design and Performance Evaluation*. Springer (2005)
- [6] Jain, A.K., Bolle, R., Pankanti, S. (eds.): *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers (1999)
- [7] Klein, D.V.: Foiling the Cracker: A Survey of Improvements to Password Security. In: *Proc. 2nd USENIX Workshop Security*, pp. 5–14 (1990)
- [8] Li, S.Z., Jain, A.K.: *Handbook of Face Recognition*. Springer (2004)
- [9] Maltoni, D., Maio, D., Jain, A.K., Prabhakar, S.: *Handbook of Fingerprint Recognition*. Springer (June 2003)
- [10] Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. *IEEE Transactions on Information Forensics and Security* 1(2), 125–143 (2006)
- [11] Sanchez-Reillo, R., Sanchez-Avila, C., Gonzales-Marcos, A.: Biometric Identification through Hand Geometry Measurements. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(10), 1168–1171 (2000)
- [12] Daugman, J.: The Importance of being Random: Statistical Principles of Iris Recognition. *Pattern Recognition* 36(2), 279–291 (2003)
- [13] Hill, R.: Retina Identification. In: Jain, A.K., Bolle, R.M., Pankanti, S. (eds.) *BIOMETRICS: Personal Identification in Networked Society*, ch. 4, 2nd Printing. Kluwer Academic Publishers (1999)
- [14] Campbell, J.P.: Speaker Recognition: A Tutorial. *Proc. IEEE* 85(9), 1437–1462 (1997)
- [15] Monrose, F., Rubin, A.: Authentication via Keystroke Dynamics. In: *Proc. 4th ACM Conference on Computer and Communications Security*, pp. 48–56 (1997)
- [16] Joussem, A.M.: Vascular plasticity - the role of the Angiopoietins in Modulating Ocular Angiogenesis. *Graefe's Archive for Clinical and Experimental Ophthalmology* 239(12), 972–975 (2001)
- [17] Isao, N., Shouta, K., Yoshio, I., Shigang, L.: *DWT Domain On-Line Signature Verification*, pp. 183–196. Tottori University, Japan
- [18] Plamondon, R.: A kinematic Theory of Rapid Human Movements: Part III: Kinetic Outcomes. *Biol. Cybern.* (January 1997)
- [19] Nalwa, V.S.: Automatic on-line Signature Verification. *Proc. IEEE* 85(2), 213–239 (1997)
- [20] Brault, J.J., Plamondon, R.: A Complexity Measure of Handwritten Curves: Modelling of Dynamic Signature Forgery. *IEEE Transactions on Systems, Man, and Cybernetics* 23(2), 400–413 (1993)

- [21] Plamondon, R.: The Design of an On-line Signature Verification System: from Theory to Practice. *International Journal on Pattern Recognition and Artificial Intelligence* 8(3), 795–811 (1994)
- [22] Jain, A.K., Ross, A.: Multi-biometric Systems. *Communications of the ACM* 47(1), 35–40 (2004)
- [23] Fairhurst, M.C.: Signature Verification Revisited: Promoting Practical Exploitation of Biometric Technology. *Electronics and Communication Engineering Journal* 9(6), 273–280 (1997)
- [24] Ismail, M.A., Gad, S.: Off-line Arabic Signature Recognition and Verification. *Pattern Recognition* 33(10), 1727–1740 (2000)
- [25] Chaudhuri, B.B., Pal, U.: An OCR System to Read two Indian Language Scripts: Bangla and Devnagari (Hindi). In: *Proceedings of 4th ICDAR*, pp. 1011–1015 (1997)
- [26] Bromme, A.: A Classification of Biometric Signatures. In: *International Conference on Multimedia and Expo, ICME*, pp. 17–20 (2003)
- [27] Nagasundara, K.B., Manjunath, S., Guru, D.S.: Multimodal Biometric System based on Hand Geometry, Palmprint and Signature. In: *5th ACM Computer Conference: Intelligent & Scalable System Technologies*, Article No. 4 (2012)
- [28] Ko, T.: Multimodal Biometric Identification for Large User population Using Fingerprint, Face and Iris Recognition. In: *Proceedings of the 34th Applied Imagery and pattern recognition Workshop*, pp. 218–223 (2005)
- [29] Ross, A., Jain, A.K.: Multimodal Biometrics: An Overview. In: *Proceedings of the 12th European Signal Processing Conference (EUSIPCO)*, pp. 1221–1224 (2004)
- [30] Ross, A., Nandakumar, K., Jain, A.K.: *Handbook of Multibiometrics*, vol. 6. Springer (2006)
- [31] Plamondon, R.: The Handwritten Signature as a Biometric Identifier: Psychophysical Model and System Design. In: *European Convention on Security and Detection*, May 16–18 (1995)
- [32] Plamondon, R.: A Kinematic Theory of Rapid Human Movements: Part I: Movement Representation and generation. *Biological Cybernetics* 72(4), 295–307 (1995)
- [33] Plamondon, R., Djioua, M.: A Multi-Level Representation Paradigm for Handwriting Stroke Generation. *Human Movement Science* 25(4–5), 586–607 (2006)
- [34] Congedo, G., Dimauro, G., Impedovo, S., Pirlo, G.: A New Methodology for the Measurement of Local Stability in Dynamical Signatures. In: *4th International Workshop on Frontiers in Handwriting Recognition*, pp. 135–144 (1994)
- [35] Impedovo, D., Pirlo, G., Sarcinella, L., Stasolla, E., Trullo, C.A.: Analysis of Stability in Static Signatures using Cosine Similarity. In: *International Conference on Frontiers in Handwriting Recognition*, pp. 231–235 (2012)
- [36] Zimmerman, T.G., Russell, G.F., Heilper, A., Smith, B.A., Hu, J., Markman, D., Graham, J.E., Drews, C.: Retail Applications of Signature Verification. In: *Proceedings of SPIE*, vol. 5404, pp. 206–214 (2004)
- [37] Plamondon, R., Lorette, G.: Automatic Signature Verification and Writer Identification Verification the State of Art. *Pattern Recognition* 22(2), 107–131 (1989)
- [38] Srihari, S.N., Leedham, G.: A Survey of Computer Method in Forensic Document Examination. In: *11th Conf. of the Intl. Graphonomics Society* (2003)
- [39] <http://en.wikipedia.org/wiki/Biometrics>
- [40] Ammar, M., Yoshida, Y., Fukumura, T.: A New Effective Approach for Off-line Verification of Signatures by using Pressure Features. In: *8th Int. Conf. on Pattern Recognition* (1986)

- [41] Leclerc, F., Plamondon, P.R.: Automatic Signature Verification: The State of the Art, 1989-1993. *International Journal of Pattern Recognition and Artificial Intelligence* 8(3), 643–660 (1994)
- [42] Coetzer, J., Herbst, B., Preez, J.D.: Off-line Signature Verification using the Discrete Radon Transform and a Hidden Markov Model. *EURASIP Journal on Applied Signal Processing* 4, 559–571 (2004)
- [43] Nguyen, V., Blumenstein, M., Muthukumarasamy, V., Leedham, G.: Off- line Signature Verification using Enhanced Modified Direction Features in Conjunction with Neural Classifiers and Support Vector Machines. In: *International Conference on Document Analysis and Recognition*, pp. 734–738 (2007)
- [44] Justino, E.J.R., Bortolozzi, F., Sabourin, R.: A Comparison of SVM and HMM Classifiers in the Off-line Signature Verification. *Pattern Recognition Letters* 2005 26(9), 1377–1385 (2005)
- [45] Hanmandlu, M., Yusof, M.H.M., Madasu, V.K.: Off-line Signature Verification and Forgery Detection using Fuzzy Modelling. *Pattern Recognition Letters* 38(3), 341–356 (2005)
- [46] Weiping, H., Xiufen, Y., Kejun, W.: A Survey of Off-line Signature Verification. In: *Proc. International Conference on Intelligent Mechatronics and Automation*, pp. 536–541 (2004)
- [47] Ferrer, M., Alonso, J., Travieso, C.: Off-line Geometric Parameters for Automatic Signature Verification using Fixed-point Arithmetic. *Pattern Analysis and Machine Intelligence* 27(6), 993–997 (2005)
- [48] Huang, K., Yan, H.: Off-line Signature Verification using Structural Feature Correspondence. *Pattern Recognition* 35(11), 2467–2477 (2002)
- [49] Rabasse, C., Guest, R.M., Fairhurst, M.C.: A Method for the Synthesis of Dynamic Biometric Signature Data. In: *Ninth International Conference on Document Analysis and Recognition*, pp. 168–172 (2007)
- [50] Dit, Y.Y., Hong, C., Yimin, X., Susan, G., Ramanujan, K., Takashi, M., Gerhard, R.: SVC 2004: First International Signature Verification Competition. In: *Proceedings of the International Conference on Biometric Authentication*, Hong Kong, July 15-17 (2004)
- [51] Pascual, C.V., Hurtado, A.S., Martinez, E.M., Gaspar, J.M.P.: A New Proposal for Score Normalization in Biometric Signature Recognition Based on Client Threshold Prediction. In: *International Conference on Data Mining*, pp. 1128–1133 (2012)
- [52] Shashi Kumar, D.R., Ravi Kumar, R., Raja, K.B., Chhotaray, R.K., Pattanaik, S.: Biometric Security System Based on Signature Verification Using Neural Networks, pp. 580–583 (2010)
- [53] Maiorana, E., Campisi, P., Neri, A.: Biometric Signature Authentication using Radon Transform-based Watermarking Techniques. In: *Biometrics Symposium*, pp. 1–6 (2007)
- [54] Ratha, N.K., Connell, J.H., Bolle, R.: Secure Data Hiding in Wavelet Compressed Fingerprint Images. In: *ACM Multimedia 2000 Workshops*, pp. 127–130 (2000)
- [55] Maiorana, E., Campisi, P., Neri, A.: Bioconvolving: Cancelable Templates for a Multi-Biometrics Signature Recognition System. In: *International Systems Conference on Digital Object Identifier*, pp. 495–500 (2011)
- [56] Maiorana, E., Campisi, P., Ortega-Garcia, J., Neri, A.: Cancelable Biometrics for HMM-based Signature Recognition. In: *International Conference on Biometrics, Theory, Applications and Systems*, pp. 1–6 (2008)
- [57] Mhatre, Maniroja: Offline Signature Verification Based on Statistical Features. In: *International Conference and Workshop on Emerging Trends in Technology*, Mumbai, India, pp. 59–62

- [58] Maiorana, E., Campisi, P., Fierrez, J., Garcia, J.O., Neri, A.: Cancelable Templates for Sequence-based Biometrics with Application to On-line Signature Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* 40(3), 525–537 (2010)
- [59] Pal, S., Pal, U., Blumenstein, M.: Hindi and English Off-line Signature Identification and Verification. In: *International Conference on Advances in Computing*, pp. 905–910 (2012)
- [60] Pal, S., Alaei, A., Pal, U., Blumenstein, M.: ‘ Multi-Script Off-line Signature Identification. In: *International Conference on Hybrid Intelligent Systems*, pp. 236–240 (2012)
- [61] Pal, S., Alaei, A., Pal, U., Blumenstein, M.: Off-line Signature Verification based on Foreground and Background information. In: *International Conference on Digital Image Computing: Techniques and Applications*, pp. 672–677 (2011)
- [62] Pal, S., Pal, U., Blumenstein, M.: Off-line English and Chinese Signature Identification Using Foreground and Background Features. In: *IJCNN Special Session on Machine Learning for Computer Vision at IEEE World Congress on Computational Intelligence*, pp. 1–7 (2012)
- [63] Pal, S., Pal, U., Blumenstein, M.: A Two-Stage Approach for English and Hindi Off-line Signature Verification. In: Petrosino, A., Maddalena, L., Pala, P. (eds.) *ICIAP 2013*. LNCS, vol. 8158, pp. 140–148. Springer, Heidelberg (2013)
- [64] Pal, S., Pal, U., Blumenstein, M.: Multi-script Off-line Signature Verification: A Two Stage Approach. In: *International Workshop on Automated Forensic Handwriting Analysis, AFHA 2013*, pp. 31–35 (2013)
- [65] Vargas, F., Ferrer, M., Travieso, C.M., Alonso, J.: Offline handwritten signature GPDS-960 Corpus. In: 9th *ICDAR*, pp. 764–768. *IEEE Computer Society* (2007)
- [66] Ferrer, M.A., Alonso, J.B., Travieso, C.M.: Offline Geometric Parameters for Automatic Signature Verification using Fixed-point Arithmetic. *Trans. on IEEE PAMI* 27, 993–997 (2005)
- [67] Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.-J., Vivaracho, C., Escudero, D., Moro, Q.-I.: *MCYT Daseline Corpus: A Bimodal Biometric Database*. *IEEE Proceedings of Visual Image Signal Processing* 150(6) (2003)
- [68] Fierrez-Aguilar, J., Alonso-Hermira, N., Moreno-Marquez, G., Ortega-Garcia, J.: An Off-line Signature Verification System based on Fusion of Local and Global Information. In: Maltoni, D., Jain, A.K. (eds.) *BioAW 2004*. LNCS, vol. 3087, pp. 295–306. Springer, Heidelberg (2004)
- [69] Impedovo, D., Pirlo, G.: Automatic signature verification: The state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38(5), 609–635 (2008)
- [70] Plamondon, R., Lorette, G.: Automatic Signature Verification and Writer Identification—the State of the Art. *Pattern Recognition* 22(2), 107–131 (1989)
- [71] Ortega-Garcia, J., Gonzalez-Rodriguez, J., Simon-Zorita, A., Cruz-Llanas, S.: From Biometrics Technology to Applications Regarding Face, Voice, Signature and Fingerprint Recognition Systems. In: *Biometric Solutions for Authentication* (2002)
- [72] Papamarkos, N., Baltzakis, H.: Off-line Signature Verification using Multiple-Neural Network Classification Structures. In: *13th International Conference on Digital Signal Processing Proceedings*, pp. 727–730 (1997)

- [73] Klement, V., Steinke, K., Naske, R.: The Application of Image Processing and Pattern Recognition Techniques to the Forensic Analysis of Handwriting. In: International Conference on Security through Science Engineering (1980)
- [74] Congedo, G., Dimauro, G., Forte, A.M., Impedovo, S., Pirlo, G.: Selecting Reference Signatures for On-line Signature Verification. In: International Conference on Image Analysis and Processing, pp. 521–526 (1995)
- [75] Lee, J., Yoon, H.S., Soh, J., Chun, B.T., Chung, Y.K.: Using Geometric Extrema for Segment-to-segment Characteristics Comparison in Online Signature Verification. *Pattern Recognition* 37(1), 93–103 (2004)
- [76] Plamondon, R., Lorette, G.: Automatic Signature Verification and Writer Identification—the State of the Art. *Pattern Recognition* 22(2), 7–13 (1989)
- [77] Inamdar, V.S., Rege, P.P., Arya, M.S.: Offline Handwritten Signature based Blind Biometric Watermarking and Authentication Technique using Biorthogonal Wavelet Transform. *International Journal of Computer Applications* 11(1), 0975–8887 (2010)
- [78] Leszczyska, J.P.: On-line Signature Verification using Dynamic Time Warping with Positional Coordinates. In: Proc. SPIE, vol. 6347, pp. 634724-1–634724-08 (2006)
- [79] Huang, K., Yan, H.: Off-line Signature Verification based on Geometric Feature Extraction and Neural Network Classification. *Pattern Recognition* 30(1), 9–171 (1997)
- [80] Garcia-Salicetti, S., Dorizzi, B.: On using the Viterbi Path Along with HMM Likelihood Information for Online Signature Verification. *IEEE Trans. Syst., Man, Cybern. B* 37(5), 1237–1247 (2007)
- [81] Kholmatov, A., Yanikoglu, B.: Identity Authentication using Improved Online Signature Verification Method. *Pattern Recognit. Letters* 26, 2400–2408 (2005)
- [82] Impedovo, D., Pirlo, G., Refice, M.: Handwritten Signature and Speech: Preliminary Experiments on Multiple Source and Classifiers for Personal Identity Verification. In: IWCF, pp. 181–191 (2008)

# Registration of Three Dimensional Human Face Images across Pose and Their Applications in Digital Forensic

Parama Bagchi<sup>1</sup>, Debotosh Bhattacharjee<sup>2</sup>,  
Mita Nasipuri<sup>2</sup>, and Dipak Kumar Basu<sup>2</sup>

<sup>1</sup> Dept. of CSE, MCKV Institute of Engineering, Kolkata-711204, India  
paramabagchi@gmail.com

<sup>2</sup> Dept. of CSE, Jadavpur University, Kolkata-700032, India  
debotosh@ieee.org,  
{mitanasipuri,dipakbasu}@gmail.com

**Abstract.** In digital forensic, three-dimensional face recognition is very challenging problem. The problem, with two-dimensional face recognition, is that, pose variation, illumination changes and expressions tend to reduce the face recognition rate. The problem aggravates in case of pose variations, especially when, the poses are rendered across extreme variations e.g. across 90 degrees. In contrast to two dimensional images, three dimensional images tend to reduce the shortcomings of two dimensional approaches. Since three dimensional images works with depth information, they do not depend on illumination, thus making the facial recognition system more robust. Pose variation affects three dimensional recognition rate hence, for a face to be recognized; it should be perfectly registered in three dimensional framework. In this book chapter, a comparative analysis of registration methods is presented for face recognition across different poses from 0 to 90°. Also, various registration approaches that are able to generalize identity, illumination and can also handle a given set of poses have been discussed in later sections. Also, several approaches used in the field of three dimensional face registration and recognition and their importance in digital forensics have been discussed. The application area of 3D faces recognition, in digital forensic, lies with the fact that, if the face of a person is oriented across a certain angle, he cannot be recognised in that position. So, perfect registration is very necessary to reconstruct the face, to be correctly recognized especially, for identifying subjects required for forensic study. Nowadays, photographic expert members from crime Investigation Bureau are researching for a number of new technologies such as facial recognition software, 3D modelling for crime scenes etc because most CCTV footages have very low picture quality which may not be much helpful for investigations. 3D face recognition system would be beneficial because they would help one to visualize the entire system, in all possible orientations.

**Keywords:** Registration, RBF, SSR Histograms, biometric, digital forensic.

## 1 Introduction

In general, biometrics measures biological characteristics for identification or verification purposes of an individual. The need for biometric in digital forensic appears because IDs and passports can be duplicated; hence more secure systems need to be developed. In biometry, there are two types of biometric methods. One is called behavioural biometrics. It is used for verification purposes. Verification is determining if a person is whom they say they are. This method looks at patterns of how certain activities are performed by an individual. Physical biometrics is the other type used for identification or verification purposes. Identification refers to determine who a person is.

This method is commonly used in criminal investigations. In addition to being used for security systems, authorities have found a number of other applications for facial recognition systems which has proved to be very important biometric in digital forensic.

Authorities were able to reduce duplicate registrations by comparing new facial images to those already in the database. Similar technologies are being used in the United States to prevent people from obtaining fake identification cards and driver's licenses. There are also a number of potential uses for facial recognition that are currently being developed. For example, the technology could be used as a security measure at ATM's, instead of using a bank card or personal identification number, the ATM would capture an image of the face, and compare it to an individual's photo in the bank database to confirm a person's identity. This same concept could also be applied to computers, by using a webcam to capture a digital image of an individual; the face could replace the password as a means to log in.

In the field of digital forensic, human face identification is a challenging field. In cases where physical evidence is available, the remains may be linked to a known person who is missing from the local area. Medical records can also be used to identify the victim. However, in the absence of further physical evidence and the medical records, forensic facial reconstruction is the only means of facilitating victim identification [Jones, 2001]. Facial recognition systems are also beginning to be incorporated into unlocking mobile devices. The android market is working with facial recognition and integrating it into their cell phones. Also, in addition to biometric usages, modern digital cameras often incorporate a facial detection system that allows the camera to focus and measure exposure on the face of the subject, thus guaranteeing a focused portrait of the person being photographed. Some cameras, in addition,, incorporate a smile shutter, or automatically takes a second image if someone closed their eyes during exposure. Because of the limitations of fingerprint recognition systems, nowadays facial recognition systems are finding market penetration as attendance monitoring alternatives because a face is more appropriate biometric than fingerprint recognition because of the fact that fingerprints of individuals can also be forged. The scope of three-dimensional face recognition, over two dimensional face recognition has gained popularity because, a 3D facial recognition and identity management solutions help the military, intelligence and law enforcement, today because of their ability to capture various poses, illumination and occlusions. 3D faces empower law enforcement to swiftly and accurately



identify criminal suspects, even from a low-resolution photo or video surveillance. 3D features enable law enforcement to analyze and compare multiple images precisely for identification of suspects [31].

Various analyses, on the various face recognition algorithms, are being used in mobile phone. Recently a model for tracking facial features on an android phone was developed at the University of Manchester. Scientists at The University of Manchester have developed software for mobile phones that can track your facial features in real-time. Eventually, it will be able to tell who the user is, where they are looking and even how they are feeling. The method is believed to be unrivalled for speed and accuracy and could lead to facial recognition replacing passwords and PIN numbers to log into internet sites from a mobile phone.

There can various types of biometrics commonly used in digital forensic. Some of them are enlisted below:-

- DNA Matching

Chemical Biometric: - The identification of the individual using the analysis of segments from DNA.

- Ear

Visual Biometric: - The identification of the individual using the shape of the ear.

- Eyes - Iris Recognition

Visual Biometric: - The use of the features found in the iris is used to identify an individual.

- Eyes - Retina Recognition

Visual Biometric: - The use of patterns of veins in the back of the eye to accomplish recognition.

- Face Recognition

Visual Biometric: - The analysis of facial features or patterns for the authentication or recognition of an individual's identity. Most face recognition systems either use eigenfaces or local feature analysis.

- Fingerprint Recognition

Visual Biometric: - The use of the ridges and valleys found on the surface tips of a human finger is to identify an individual.

- Finger Geometry Recognition

Visual/Spatial Biometric: - The use of three dimensional geometry of the finger to determine the identity.

- Gait Recognition

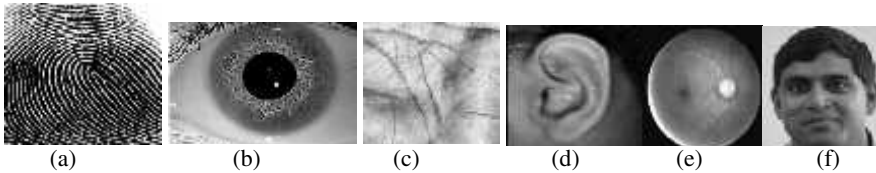
Behavioral Biometric: - The use of an individual's walking style or gait to determine the identity.

- Hand Geometry Recognition

Visual/Spatial Biometric: - The use of the geometric features of the hand such as the lengths of fingers and the width of the hand to identify an individual.

- Auditory Biometric: - The use of the voice as a method of determining the identity of a speaker for access control.

Fig-1 shows the common biometric traits used in digital forensics [1].



**Fig. 1.** Commonly used biometrics: (a) fingerprint (b) iris (c) palm (d) ear (e) retina and (f) face in digital forensic

## 2 Importance of Three Dimensional Face Registration and Recognition in Digital Forensic

The human face plays a very pivotal role in digital forensic science. Using the human face as a key to security, biometric face recognition technology has received significant attention in the past several years due to its potential for a wide variety of applications. The primary advantage of face recognition, as compared with other biometrics systems using fingerprint/palmprint and iris recognition is that, face recognition has distinct advantages because of its non contact process. Face images can be captured from a distance without touching the person being identified, and the identification does not require interacting with the person. Fig-2 shows, how a face recognition system works for a biometric system.



**Fig. 2.** A face recognition system used in Digital Forensic Systems

In the literature, many two dimensional face recognition systems in biometrics have been developed. Fig-3 shows the steps that are used for a face recognition system, starting from image acquisition to face recognition.



**Fig. 3.** Steps of a face recognition system

As a biometric, facial recognition [2] is a form of computer vision that uses faces attempting to identify a person or verify a person's claimed identity. Though two dimensional face recognition systems are simple to implement, they have several disadvantages. Three-dimensional face recognition represents an improvement over two dimensional face recognition in some respects. Recognition of faces from still images is a difficult problem, because the illumination, pose and expression changes in the images create great statistical differences and the identity of the face itself becomes shadowed by these factors. Humans are very capable of this modality, precisely because they learn to deal with these variations. Three-dimensional face recognition has the potential to overcome feature localization, pose and illumination problems, and it can be used in conjunction with two dimensional systems.

Three dimensional faces recognition varies a lot with respect to pose variations because of which they are to be registered. Registration is the process of aligning two three-dimensional datasets to a common framework. In case the facial alignment is incorrect, the recognition rate of three-dimensional faces would diminish. In the next section, we review the current research on three dimensional face registration and recognition. We focus on different representations of three dimensional information, and the fusion of different source of information. Another new technique in facial recognition uses the visual details of the skin, as captured in standard digital or scanned images. We conclude by a discussion of the future of three-dimensional face recognition.

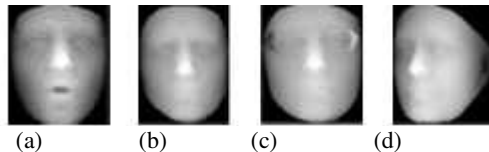
The benefits of facial recognition are that it is not intrusive, can be done from a distance even without the user being aware they are being scanned. What sets apart facial recognition from other biometric techniques is that it can be used for surveillance purposes; as in searching for wanted criminals, suspected terrorists, and missing children. Facial recognition can be done from far away so with no contact with the subject, so they are unaware they are being scanned. Facial recognition is most beneficial to use for facial authentication than for identification purposes, as it is too easy for someone to alter their face, features with a disguise or mask, etc.

### 3 A Review of the Various Three Dimensional Registration and Recognition Techniques

As discussed earlier, if a face recognition system is to be set up for digital forensic, for effective recognition of the system accurate registration between two three dimensional faces is necessary. For complete recognition the various steps which are required for three dimensional image processing are enlisted below:-

#### 3.1 Data Acquisition

Normally three dimensional files are obtained by sensors. Three dimensional information needs to be pre-processed after acquisition. Depending on the type of sensor, there might be holes and spikes in the range data. Eyes and hair will not reflect the light appropriately, and the structured light approaches will have trouble accurately registering those portions. Illumination still effects the three dimensional acquisition, unless accurate laser scanners are used [2, 3]. Normally when the image is acquired, it is projected into a two-dimensional plane which is called the range image, more specifically a 2.5D range image as shown in Fig-4.



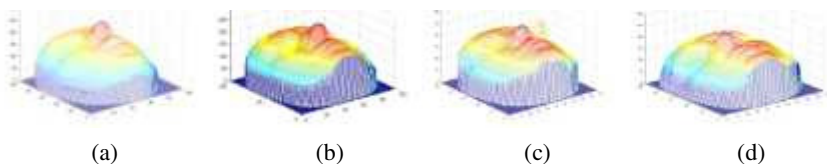
**Fig. 4.** Range images from the Bosphorus Database for a person in frontal-pose (a), sad-face (b) occluded face with glasses(c) and face oriented across y-axis (d)

#### 3.2 Three Dimensional Pre-processing

Surface smoothing[4] refers to the fact that noisy spikes and other various deviations are sometimes caused on the three dimensional face images. So some types of smoothing techniques are to be applied. In this present technique, the concept of two-dimensional weighted median filtering technique has been extended to three dimensional face images. The present technique performs filtering of three dimensional dataset using the weighted median implementation of the mesh median filtering. The weighted median filter is a modification of the simple median filter.

**Weighted median filtering:-** Weighted Median (WM) filters are the filters that have the robustness and edge preserving capability of the classical median filter and resemble linear FIR filters. In addition, weighted median filters belong to the broad class of nonlinear filters called stack filters. This enables the use of the tools developed for the latter class in characterizing and analyzing the behaviour of weighted median filters in noise attenuation capacity. Applications of weighted median filters include idempotent weighted median filters for speech processing,

adaptive weighted median and optimal weighted median filters for image and image sequence restoration. After pre-processing, the range data corresponding to Fig-4 would look as in Fig-5.



**Fig. 5.** Range images from the Bosphorus Database after smoothing for a person in frontal-pose (a), sad-face (b), occluded face with glasses(c) and face oriented across y-axis (d) corresponding to Fig-4

### 3.3 Feature Selection

In order to register two three dimensional images there are many prevailing techniques of feature selection which are necessary for complete registration. In Table-1, the different feature detection methods which are necessary for three dimensional registration are enlisted [7].

### 3.4 Three Dimensional Face Registration

It has already been discussed that, for an effective recognition, perfect registration is an essential factor in the field of three- dimensional face recognition for digital biometric. Facial expressions as well as pose changes; both are essential factors for the degradation of three-dimensional face recognition. In the Section below, brief discussions of each of the three dimensional registration techniques have been mentioned. Now, for effective registration across pose, the face should be correctly aligned in the frontal position. In the Section below, we hereby discuss the various methodologies for registering the 3D range image across varying poses, and we shall also discuss how effectively each method handles various 3D range images across different poses.

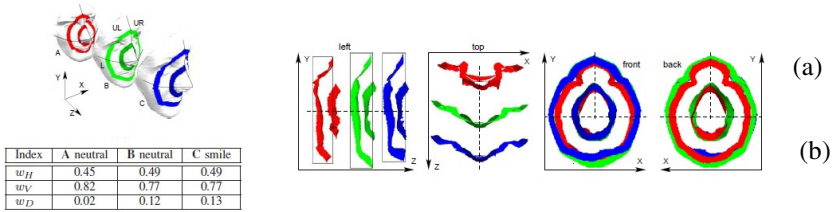
**a) Facial expression based approaches for pose variant three-dimensional face recognition based on landmarks:** - In this method, features were extracted at the landmark regions and face matching was performed according to Hierarchical Graph Matching, with graph nodes positioned at the landmarks. In [9], multiple face regions are originated by intersecting three-dimensional face scans with spheres of increasing radius centred on the middle point between the nose tip and the nasion. In [8], multiple overlapping regions around the nose are segmented, and the scores of ICP matching on these regions are combined together. A further improvement of the approach has been proposed in [9] by considering a multi instance enrolment of gallery scans with multiple expressions (experiments are provided using up to five

**Table 1.** A Comparative Study of Various Feature Detection Techniques

CATEGORY OF METHODS	DESCRIPTORS	SHAPE FEATURE	ADVANTAGES	LIMITATIONS	REFS
<i>Global feature-based</i>	Cord and angle histograms	Histograms of the length	Only considering surface normals	a. Simplifying triangles to their centres b. Intolerant to impact of shape	[17][18][19]
<i>Distribution feature-based</i>	Shape distributions	Collection of shape functions	a. Fast, simple, useful to discriminate three-dimensional shapes b. Randomization of the surface sampling process improves the estimation c. Histogram accuracy can be controlled with sampling density.	a. Shape functions are not adequate to describe the three dimensional shape. Better to be pre-classification.	[21]
<i>Spatial map based</i>	Shape histograms	Accumulation of the surface points in the bins based on a nearest-neighbour rule	a. Intuitive b. make use of Quadratic form distance functions to take into account the distances between histogram bins	a. Voxelization is needed prior to descriptor extraction b. Need further reduction of dimensionality of feature vectors.	[22]
	Radial cosine transform	Coefficients of radial cosine functions	a. Easy to calculate b. Uses small number of features	a. Retrieval results are worse than commonly used methods.	[23]
	Three dimensional Fourier transform	Three dimensional Fourier transform	a. Associate with a spatial alignment procedure to be geometric in-variant; b. Satisfy the storage and computational complexity requirements c. Completely independent of the mesh topology	Only the canonical three- dimensional of octahedron-based partition is rotation invariant	[20][28]
	Shape spectrum	Spherical functions on concentric spheres	a. Consider the internal structure of a model by using functions on concentric spheres; b. Rotation invariant by using CPCA.	a)Alignment to principal axes is needed; b)Three-dimensional leads to problems with outliers	[7]

scans per individual). Accordingly, up to 140 ICP region matches are required to compute the similarity between a probe scan and the scans representing an enrolled individual. Robustness to no neutral facial expressions is improved at the cost of greater computational complexity, thus making these approaches more suited to face verification than identification.

In this work, face partitioning was done using equal width isogeodesic stripes. The stripes provide an effective representation of three dimensional faces that, permits distinguishing facial differences due to, different facial traits of different individuals. The method using isogeodesic stripes has been shown in Fig 6(a), and the partitioning of the various features has been shown in Fig 6(b).



**Fig. 6.** (a) Sample face models B and C: Models of the same individual with neutral and smiling expression; A: Model of different individual with the neutral expression. (b) Projections of the pairs of face stripes on the coordinate planes.

**b) A robust three dimensional face matching in the presence of nonrigid deformation and pose changes:** In the work on three dimensional face registration and recognition [10], a proposed work was performed in the presence of nonrigid deformation and pose changes in the test scan. A hierarchical surface sampling scheme was used to augment fiducial landmarks for analyzing three-dimensional facial surfaces across expression. The fiducial landmarks needed during expression learning were manually extracted here. Additional landmarks (74 points) in facial surface regions with a little texture are automatically extracted using the geodesic based approach. Three dimensional deformation learned from a small number of subjects is transferred to the three dimensional neutral models in the gallery. To generate three dimensional nonneutral models, the corresponding deformation is synthesized in the three dimensional neutral model. Two types of deformable models have been built, expression specific and expression generic. The matching is performed by analyzing and fitting the deformable model to a given test scan, which is formulated as a minimization of a cost function. The block diagram for the approach used is shown in Fig-7.

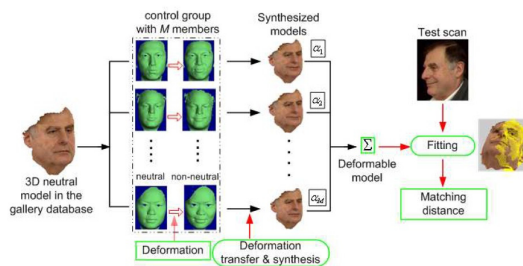


Figure 1. Deformation modeling for 3D face matching.

**Fig. 7.** Deformation modelling for three-dimensional faces matching

The deformation learned from the control group is transferred to the three-dimensional neutral model. Each subject in the control group provides its own deformation transform. The method used above uses the following parameters:-

Two types of transformations are applied to a three dimensional deformable model.

1. The first one is a rigid transformation due to the head pose changes, which can be represented by a rotation matrix and a translation vector.
2. The second one is the nonrigid deformation, modelled by the weights fitting the deformable model to a given test scan, which is formulated as an optimization problem to minimize the cost function. Human faces share a common geometric topology, which can be represented by the facial landmarks. A fiducial set of 9 landmarks are extracted 8(i.e., two inner eye corners, two outside eye corners, two mouth corners, nasion, nose tip, and subnasal)were used to model the framework. Based on the layer of landmarks:-

- (1) The faces are registered by matching the nonneutral scan with the neutral scan to estimate the displacement vector of landmarks due to the expression change
- (2) Establish a mapping from the landmark set of the neutral scan to that of the three-dimensional neutral model
- (3) Use the mapping to transfer the landmarks in the non-neutral scan to the three-dimensional neutral model.
- (4) Apply to other vertices in the three dimensional neutral model to move them to the new positions caused by the expression.

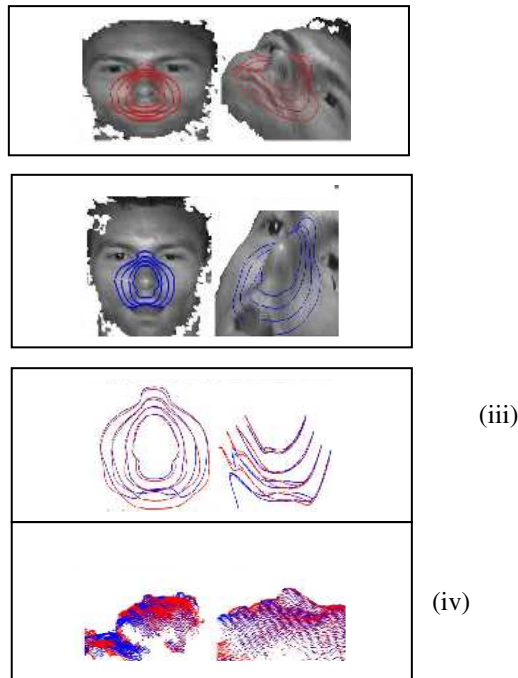


**Fig. 8.** Deformation modelling for three-dimensional faces matching. To match the 2.5D test scan (b) to a three dimensional neutral face model (a) in the gallery database.

**c) An RBF-based approach to map noisy three dimensional range images to pose aligned or poses normalized depth maps:-** In this approach, pairwise aligned normalized depth maps from noisy three dimensional range images in unrestricted poses were developed. The system so deployed in three dimensional pose alignment applications consisted of the following four stages:-

- (i) Data filtering
- (ii) Nose tip identification and sub vertex localization
- (iii) Computation of the face orientation
- (iv) Generation of either a pose aligned, or a pose normalized depth map.





**Fig. 9.** The influence of mouth closed (red)/open (blue) on isoradius contours. (i) Mouth closed. (ii) Mouth open. Note that isoradius contours fall under the texture map in the mouth area. (iii) Isoradius contours after alignment: front view and profile view (associated with d, right). (iv) Aligned range images

Here an implicit radial basis function (RBF) model of the facial surface and this is employed within all four stages of the process. For example, in stage (ii), construction of novel invariant features was based on sampling this RBF over a set of concentric spheres to give a spherically sampled RBF (SSR) shape histogram. In stage (iii), a second novel descriptor, called an isoradius contour curvature signal, was defined, which allowed rotational alignment to be determined using a simple process of 1D correlation. The system was tested both on the University of York (UOY) three dimensional face dataset and the Face Recognition Grand Challenge (FRGC) three dimensional data. For a more challenging data, the SSR descriptors significantly outperforms the three variants of spin images, successfully identifying nose vertices at a rate of 99.6%. Nose localization performance on the higher quality FRGC data, which has only small pose variations, is 99.9%. Fig-9 shows a block diagram of the approach.

**(d) A facial symmetry based approach to handle pose variations in three-dimensional face recognition:** - This method to handle pose variations in three-dimensional face recognition domain, performed comparisons among interpose

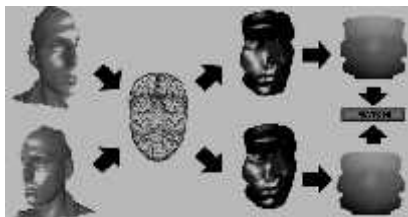
scans using a wavelet based biometric signature. It is suitable for real life applications as it only requires half of the face to be visible to the sensor. In this method, no user intervention was required, and the method was completely automatic. No subject cooperation was required, and the method could handle large pose variations. It can be applied to large databases: It has a reasonable computational cost with excellent scalability. In this method, initial registration was done using AFM (Annotated Face Model) using detected landmark allows pose estimation of each facial scan. Then, the AFM is fitted to the facial scan using a subdivision based deformable model framework that is extended to allow symmetric fitting. The symmetric fitting solves the problem of the missing data. The following steps are given below:-

- Step 1. Pre-processing: Standard pre-processing techniques are used to filter the raw data
- Step 2. Three dimensional Landmark Detection: A robust landmark detector is used for pose estimation (to determine if it is a frontal, left, or right scan).
- Step 3. Registration: The raw data are registered to the AFM using a two stage approach.
- Step 4. Symmetric Deformable Model Fitting: The AFM is fitted to the data using facial symmetry. The fitted model is then converted to a geometry image and a standard image.
- Step 5. Wavelet Analysis: A wavelet transform is applied on the geometry and normal images and the wavelet coefficients are stored as a biometric signature.

The parameters for the landmark localization, and registration processes, are enlisted as follows:-

1. Extract candidate landmarks from the Shape Index map.
2. Classify candidate landmarks by matching them with the corresponding Spin image templates.
3. Create feasible combinations of five landmarks from the candidate landmark points.
4. Compute the rigid transformation that best aligns the combinations of five candidate landmarks.
5. Fuse accepted combinations of five landmarks (left and right) in complete landmark sets of eight landmarks.
6. Compute the rigid transformation that best aligns the combinations of eight landmarks.
7. Sort consistent, complete landmark sets in descending order according to a distance metric.
8. Select the best combination of landmarks based on the distance metric
9. Obtain the corresponding rigid transformation for registration.

Fig-10 shows the 3D AFM model, where only half of the face is used for registration purpose. The approach method is very suitable for real-world biometric applications were proposed.



**Fig. 10.** Interpose matching using the proposed method (left to right): Opposite side facial scans with extensive missing data, generic AFM, deformed AFM for each scan (facial symmetry used), and extracted geometry images.

**e) A feature extraction approach based on depth images to handle poses variations in any axis (i.e. X, Y and Z):-** A novel work was also done which extracts features i.e. the nose-tip based on the intensity based values of the depth image. A very significant work was done on the three dimensional face images which concentrate on the use of smoothing by weighted median filters for analysis of feature detection. The proposed work could be divided into the following stages:-

Step 1. Facial Image Acquisition:-Here a three dimensional range image is acquired.

Step 2.Preprocessing of the range image:- Here thresholding of the range image is done using Otsu's thresholding algorithm.

Step 3. Feature selection using depth intensity concept:-Next nose tip is detected using a maximum intensity technique across any different poses.

Step 4.Alignment of models:-In the last and final step angle of elevations is found out across any axis.

Step 5.Registration:-Finally the range image is registered in the frontal pose and results compared to frontal scans

A three dimensional range image may be oriented across the X, Y and Z axes. So, in this case, a significant work has been done by the authors in [25] of calculating the pose angle across X, Y and Z axes and then registering the three dimensional range images across the axis. In Table-2, a discussion is presented regarding the various three dimensional registration techniques and how their impact would affect the field of biometrics in digital forensic and how important they are in the field of digital forensic. The importance of 3D faces across pose lies with the fact that, owing to the various pose orientations, it is quite obvious that, for correct recognition different methodologies are used for face recognition. In the following Table, the methodologies have been described and in what way and for which poses the methods, could be applied have also been discussed. So, if subjects of those form had been taken up in digital forensics, he following methods could be used for effective recognition.

**Table 2.** A Comparative Analysis of Different three dimensional Registration Techniques

Sl	Name of the Technique	Databases Used	Poses Considered and Corresponding Angles	Advantage of the three-dimensional Registration Methods in Biometric
1.	Face Registration By Hierarchical Graph Matching	FRGC Database(With expressions)	No pose but faces with expressions were considered.	Here, facial differences could be distinguished due to different facial traits of different individuals from differences induced by facial expressions of the same individual. The FRGC database consisted of 50,000 3D faces
2.	Deformation Modeling for three-dimensional Face Registration and Matching	FRGC Database(With Pose and Expression Both)	Poses oriented across Y-Axis:- 0° to 45°	This method is said to improve matching accuracy in the presence of expression variations along with large pose changes with 97% accuracy. The FRGC database consisted of 50,000 3D faces.
3.	RBF based Registration approach	UoY face dataset, FRGC dataset	Poses with minor variations oriented across X, Y and Z axes.	For the more challenging data, the SSR descriptors significantly outperform three variants of spin images, successfully identifying nose vertices at a rate of 99.6%. Nose localization performance on the higher quality FRGC data, which has only small pose variations, is 99.9%. The UoY database currently consists of over 5000 3D face models of approximately 350 people (15 models each). The FRGC database consisted of 50,000 3D faces.
4.	Facial symmetry based approach to handle pose variations in three-dimensional face	FRGC Dataset	Poses oriented across Y axes but with variations up to 40°.	The approach method is very suitable for real world biometric applications because it uses only half of a face. The FRGC database consisted of 50,000 3D faces.
5	A feature extraction approach based on depth images to handle poses variations	FRAV3D Database	Poses with minor variations oriented across X, Y and Z axes for poses ranging from 18 to 40 degrees.	Here, in case of all the images which are aligned across X, Y and Z axes, at first the angle across which the image is rotated has been determined which proves the novelty of the algorithm. After testing the proposed method on 472 images from the FRAV3D database, the method, correctly registers 358 images thus giving a performance rate of only 75.84%.

**3.5 Face Recognition:** A facial recognition device is one that views an image or video of a person and compares it to one that is in the database. It does this by comparing structure, shape and proportions of the face; distance between the eyes, nose, mouth and jaw; upper outlines of the eye sockets; the sides of the mouth; location of the nose and eyes; and the area surrounding the cheek bones. Upon enrolment in a facial recognition program, several images are taken of the subject at different angles and with different facial expressions. At time of verification and identification the subject stands in front of the camera for a few seconds, and then the image is compared to those that have been previously recorded. Some security measures have been put into place to prevent the subject from using a picture or mask being scanned in a facial recognition program. When the user is being scanned, they may be asked to blink, smile or nod their head.

Another security feature would be the use of facial thermograph to record the heat in the face. The main facial recognition methods are feature analysis, neural network, and eigenfaces, automatic face processing. Some facial recognition software

algorithms identify faces by extracting features from an image of a subject's face. Other algorithms normalize a gallery of face images and then compress the face data, only saving the data in the image that can be used for facial recognition. A probe image is then compared with the face data. A fairly new method on the market is three-dimensional facial recognition. This method uses 3D sensors to capture information about the shape of the face. This information is then used to identify distinctive features on the face, such as the contour of eye sockets, nose and chin. The advantages of 3D facial recognition are that it is not affected by changes in lighting, and it can identify a face from a variety of angles, including a profile view.

## 4 Conclusion and Future Scope

In addition to video surveillance and criminal identification, facial recognition systems are being used to reconstruct facial images of skulls from skeletons and to age pictures of victims missing in cold case investigations. In this book chapter, we have emphasized the concept of 3D face registration and recognition with respect to pose. Because the problem of face alignment is a major issue, many recognition system manufacturers are now developing software programs for their equipment that will construct a 3D presentation of a person.

Alignment of facial images for comparisons to known images presents another challenge. New software is being developed to produce three-dimensional images that can be manipulated for proper alignment.

Over the past decade, there has been a large volume of scientific research on facial recognition. In this discussion, on three dimensional face registration and its importance in digital forensic, we hope to have emphasized the key points of the need for three dimensional registration and its utmost importance in the field of forensic science. There are some other factors like illumination, facial occlusions which cause the degradation of a facial recognition system. Various steps involved in the reconstruction process are by now automated; however, the key to complete automation of the process is landmark identification, and this was the motivation for this research. The goal of this research was to take a step forward towards automation of the forensic facial reconstruction process by developing a mechanism which for 3D registration and recognition. Despite recent advances in the area, facial recognition in a surveillance system is often technically difficult. The main reasons are difficulties in finding the face by the system. These difficulties arise from people moving, wearing hats or sunglasses, and not facing the camera. However, even if the face is found, identification might be difficult because of the lighting (too bright or too dark), making features difficult to recognize.

However, the present works have some limitations. We have to find out some robust methodologies for discarding outliers in case of very noisy images. Also, as a part of our future work, we shall develop a more robust method of registration for registering extreme poses because for unknown persons with extreme poses correct measures for registration need to be found out which would be of valuable need to digital forensic and biometric field. Automation of forensic facial registration and reconstruction process is a challenging task, and several researchers are working towards automating the process.

**Acknowledgement.** This work has been supported partially by a grant from DeiTY, MCIT, Govt. of India.

## References

1. Jain, A.K., Kumar, A.: Biometrics of Next Generation: An Overview. In: Proc. of Second Generation Biometrics. Springer (2010)
2. Wood Jr., J.D., Horn, C., Gtaune, J., Thomas, A.: Biometrics A Look at Facial Recognition. Documented Reading published in (2003)
3. Bowyer, K., Chang, K., Flynn, P.: A survey of multi-modal two-dimensional + three-dimensional face recognition. Technical Report, Notre Dame Department of Computer Science and Engineering (2004)
4. Bagchi, P., Bhattacharjee, D., Nasipuri, M., Basu, D.K.: A novel approach for nose tip detection using smoothing by weighted median filtering applied to three dimensional face images in variant poses. In: Proc. of Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21-23, p. 272. IEEE, Periyar University (2012)
5. Bagchi, P., Bhattacharjee, D., Nasipuri, M., Basu, D.K.: A Novel Approach in detecting pose orientation of a three dimensional face required for face registration. In: Proc. of International Conference on Intelligent Infrastructure, Science City, Kolkata, December 1-2, pp. 1–2 (2012)
6. Xu, C., Wang, Y., Tan, T., Quan, L.: Automatic three dimensional face recognition combining global geometric features with local shape variation information. In: Proc. AFGR, pp. 308–313 (2004)
7. Zhang, L., Fonseca, M.D., Ferreira, A.: Survey on three dimensional shape descriptors. In: Proceedings of SPIE Conference on Nonlinear Image Processing and Pattern (2007)
8. Lee, Y., Park, K., Shim, J., Yi, T.: Three dimensional face recognition using statistical multiple features for the local depth information. In: Proc. ICVI (2003)
9. Lu, X., Colbry, D., Jain, A.K.: Three dimensional Model Based Face Recognition. In: Proc. ICPR (2004)
10. Akgul, C.B.: Three dimensional Shape Descriptors and Similarity Learning. Thesis
11. Saupe, D., Vranic, D.V.: Three dimensional model retrieval with spherical harmonics and moments. In: Radig, B., Florczyk, S. (eds.) DAGM 2001. LNCS, vol. 2191, p. 392. Springer, Heidelberg (2001)
12. Vrani, D.V., Saupe, D.: Description of three dimensional-shape using a complex function on the sphere. In: Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2002), Lausanne, Switzerland, pp. 177–180 (August 2002)
13. Funkhouser, T., Min, P., Kazhdan, M., Chen, J., Halderman, A., Dobkin, D., Jacobs, D.: A search engine for three dimensional models. Proc. of ACM Trans. Graph. 83–105 (2003)
14. Kazhdan, M., Funkhouser, T., Rusinkiewicz, S.: Rotation invariant spherical harmonic representation of three dimensional shape descriptors. In: Proc. of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing (SGP 2003), Aire-la-Ville, Switzerland, pp. 156–164. Eurographics Association (2003)
15. Podolak, J., Shilane, P., Golovinskiy, A., Rusinkiewicz, S., Funkhouser, T.: A planar reflective symmetry transform for three dimensional shapes. In: Proc. of ACM SIGGRAPH (2006)
16. Paquet, E., Murching, A., Naveen, T., Tabatabai, A., Rioux, M.: Description of shape information for two-dimensional and three dimensional objects. Proc. of Signal Processing: Image Communication 16, 103–122 (2000)

17. Paquet, E., Rioux, M.: Nefertiti, A query by content software for three dimensional models databases management. In: Proc. of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling (NRC 1997), Washington, DC, USA, p. 345. IEEE Computer Society (1997)
18. Paquet, E., Rioux, M.: Nefertiti, A tool for three dimensional shape databases management. Proc of SAE Transactions: Journal of Aerospace 108, 387–393 (2000)
19. Ankerst, M., Kastenmüller, G., Kriegel, H.-P., Seidl, T.: Three dimensional shape histograms for similarity search and classification in spatial databases. In: Güting, R.H., Papadias, D., Lochovsky, F.H. (eds.) SSD 1999. LNCS, vol. 1651, pp. 207–226. Springer, Heidelberg (1999)
20. Duta Gaci, H., Sankur, B., Yemez, Y.: Transform-based methods for indexing and retrieval of three dimensional objects. Three Dimensionalim, 188–195 (2005)
21. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. PAMI 24(4), 509–522 (2002)
22. Kortgen, M., Park, G.-J., Nonvoting, M., Klein, R.: Three dimensional shape matching with three dimensional shape contexts. In: Proc. of The 7th Central European Seminar on Computer Graphics (April 2003)
23. Horn, B.K.P.: Extended Gaussian images. Proc. of the IEEE 72, 1671–1686 (1984)
24. Kang, S.B., Ikeuchi, K.: The complex EGI: A new representation for three- dimensional pose determination. Proc. of IEEE Trans. Pattern Anal. and Mach. Intell. 15(7), 707–721 (1993)
25. Koendering, J.: Solid shape. In: Proc of. The MIT Press (1990)
26. Bagchi, P., Bhattacharjee, D., Nasipuri, M., Basu, D.K.: A method for nose-tip based three dimensional face registration using maximum intensity algorithm. In: Proc of Proceedings IC3A, JIS College of Engineering, January 11-12 (2013)
27. Zaharia, T., Preteux, F.: Shape-based retrieval of three dimensional mesh models. In: Proc. of the IEEE International Conference on Multimedia and Expo (ICME 2002), Lausanne, Switzerland (August 2002)
28. Zaharia, T., Preteux, F.: Three dimensional shape-based retrieval within the MPEG-7 frame work. In: Proceedings SPIE Conference on Nonlinear Image Processing and Pattern Analysis XII, San Jose, CA, vol. 4304, pp. 133–145 (January 25, 2013)
29. <http://www.manchester.ac.uk/aboutus/news/display/?id=6289>
30. <http://animetrics.com/animetrics-introduces-advanced-investigative-facial-recognition-solution-for-law-enforcement/>

# Computational Methods for the Analysis of Footwear Impression Evidence\*

Sargur N. Srihari and Yi Tang

Department of Computer Science and Engineering, Center of Excellence for Document, The State University of New York SUNY, Buffalo, New York, USA  
srihari@buffalo.edu

**Abstract.** Impressions of footwear are commonly found in crime scenes. Yet they are not routinely used as evidence due to: (i) the wide variability and quality of impressions, and (ii) the large number of footwear outsole designs which makes their manual comparison time-consuming and difficult. Computational methods hold the promise of better use of footwear evidence in investigations and also in providing assistance for court testimony. This paper begins with a comprehensive survey of existing methods, followed by identifying several gaps in technology. They include methods to improve image quality, computing features for comparison, measuring the degree of similarity, retrieval of closest prints from a database and determining the degree of uncertainty in identification. New algorithms for each of these problems are proposed. An end-to-end system is proposed where : (i) the print is represented by an attribute relational graph of straight edges and ellipses, (ii) a distance measure based on the earth-mover distance, (iii) clustering to speed-up database retrieval, and (iv) uncertainty evaluation based on likelihoods. Retrieval performance of the proposed design with real crime scene images is evaluated and compared to that of previous methods. Suggestions for further work and implications to the justice system are given.

**Keywords:** Footwear, Impression evidence, Computational forensics, Image similarity, Crime scene images.

## 1 Introduction

Marks made on floors, carpets and other surfaces by the sole of footwear, known as footwear impressions, are the most commonly found type of evidence in crime scenes. Outside soles of footwear, known as outsoles, contain patterns that are designed by the footwear manufacturer. The design is both for functionality, i.e., gripping the walking surface, and for aesthetics, i.e., pleasing appearance. The patterns can change in distinctive ways over time depending on length of wear and walk characteristics such as gait and pressure. Impressions are created when

---

\* This work was supported by the Office of Justice Programs, US Department of Justice on NIJ Award 2007-DN-BX-K135. The opinions expressed are those of the authors and not of the DoJ.



footwear is pressed or stamped against a surface such as floor or furniture, in which process, the characteristics of the outsole are transferred to the surface. Impressions can contain three-dimensional information, e.g., on snow, wet dirt or at the beach, but more often contain only two-dimensional patterns, e.g., on a floor or carpet. They are said to be present more often and found more frequently than fingerprints. Evidence provided by a positively identified mark can be as strong as evidence provided by other types of impression evidence such as fingerprints, tool marks, and typewritten impressions [1].

Footwear impression patterns can be useful for either identifying the sole type or the individual sole that made the impression. If the mark is identified as having been made by an individual outsole, to the the exclusion of all others, then it is referred to as *individualization*. It is based on *individualizing characteristics*, which are random marks the sole has acquired during its life. Individualizing characteristics are unique to the particular footwear that is the source of the impression. They are attributable to shoe sole defects such as nicks, scratches, cuts, punctures, tears, embedded air bubbles caused by manufacturing imperfections, and ragged holes [2]. A combination of position, configuration, and orientation of each defect, which are the result of events that occurred in its life, are unique to each shoe. A defect position is characterized relative to: print perimeter, particular tread elements or portions of patterns, or other defects. A defect shape is characterized by its length, width, and other shape measures. The rotational orientation of the defect helps differentiate from other similarly shaped defects.

A broader type of identification is *classification* to determine the specific sole type, e.g., brand, based on *class characteristics*, which are features of the sole type. Detail retained may be insufficient to uniquely identify an individual shoe but is still very valuable. Since a wide variety of footwear is available on the market, with most having distinctive outsole patterns, any specific model will be owned by a very small fraction of the general population, although the same outsole pattern can be found on several different footwear brands and models. If the outsole pattern can be determined from its mark, then this can significantly narrow the search for a particular suspect. Class characteristics are useful for discriminating between different sole types. They capture the geometry of the pattern. Since there are a large number of sole types, they can be used to narrow-down the possibilities. Determining sole type can be regarded as a problem of image retrieval where the query is the print of unknown type and the database consists of all known prints whose impressions can be obtained using a chemical surface. Individualizing and class characteristics together enable determining whether a crime scene print matches a known.

Although ubiquitous, the poor quality and wide variability of impressions as well as the large number of manufactured outsole patterns makes their analysis and courtroom presentation difficult. Even in Europe, where footwear impression evidence is more commonly used, it is not used as frequently as it could be. For example, only 500 of 14,000 recovered prints in the Netherlands were identified [3]. This is because footwear impressions are usually highly degraded, prints are inherently complex and databases are too large for manual comparison. There

is variability in the quality of footwear impressions because of the variety of surfaces on which the impressions are made.

The rest of the paper is organized as follows. After a review of current practice in the US (Section 2), and the published computational literature (Section 3), algorithms for several subproblems are discussed in Section 4: image processing to improve image quality, extraction of features for class characterization, methods for measuring the similarity of prints, computing features for individualization, and quantifying opinion. Implications of the methods to practice as well as future work that needs to be done are indicated in Section 5.

## 2 Current Practice

The forensic examiner collects and preserves footwear and tire tread impression evidence, makes examinations, comparisons, and analyses in order to: (i) include, identify, or eliminate a particular footwear, or type of outsole, as the source of an impression, (ii) determine the brand or manufacturer of the outsole or footwear, (iii) link scenes of crime, and (iii) write reports and provide testimony as needed. The photograph of the impression or of the lifted impression or cast can be subsequently scanned and a digital image produced. Forensic analysis requires comparison of this image against other images such as: (i) marks made by footwear currently and previously available on the market and (ii) marks found at other crime scenes.

An image of a footwear impression can be obtained using photography, gel, or electrostatic lifting or by making a cast when the impression is in soil. Subsequently, in the forensic laboratory, the image is compared with prints and impressions of known footwear. In computerized identification, known prints (collected with care to capture of all possible impression information) are scanned, processed and indexed into a database, with the objective of retrieving the most likely matching prints.

Difficulty in identification is due to poor quality images and differences in environment between impression and knowns. While digital image enhancement, e.g., contextual thresholding, can enhance impression quality, debris, shadows and artifacts are difficult to filter out. Thus it is useful to segment the image into useable (impressed by footwear) and discardable regions (impressed by other artifacts such as debris).

European research has focused on tasks with important practical differences from the needs of US examiners. Impressions from scenes, assembled from several locations, are searched to find matches with crime scene impressions. Usable impressions are present in 30% of all burglaries [4], e.g., a study of several jurisdictions in Switzerland revealed that 35% of crime scenes had usable footwear impressions, and 30% of all burglaries provide usable impressions[5]. Timely identification allows linking of crime scenes— since most crimes are committed by repeat offenders, several offenses are common in the same day, and offenders rarely discard footwear between crimes[6]. Since manual identification is laborious there exists a real need for automated methods.

Most crimes investigated in the US are homicides and assaults, not burglaries. In such cases, particularly homicides, it is unlikely that the same impression will appear in another case. Here the classification task of determining brand, style, size, gender etc., is of importance. Through classification, even if the person could not be identified, the search could be narrowed down to a smaller set of suspects. Forensic examiners of footwear and tire impression evidence are a community of about 200 professionals in the U. S. Guidelines for the profession are given by the Scientific Working Group on Footwear and Tire Tread Evidence (SWGTTREAD). Footwear prints constitute about 80-90% of the case-work of the tread examiner who deals with both footwear and tire-marks.

The final step in footwear impression evidence analysis is to state the result of comparison for presenting forensic evidence in the courtroom. In order to establish a uniform ground for interpreting footwear impression evidence, the ENFSI (European Network of Forensic Institutes) footwear impression and tool mark working group has proposed a 5-point conclusion scale ranging from identification, very strong (strong) support, moderately strong support, limited support.

### 3 Existing Software and Algorithms

Several software tools for processing and comparison of footwear impressions have been described in the literature. We provide here a summary of such methods as a backdrop for the algorithms we propose in Section 4. The earliest were semi-automatic methods of manually annotated footwear print descriptions using a codebook of shape primitives, e.g., wavy patterns, geometric shapes and logos [7,8]. The query print is also encoded in a similar manner. The most popular such systems today are SOLEMATE and SICAR [9,10]. These systems rely on manually encoding shoe-prints using a codebook of shapes and geometric primitives, such as wavy patterns, zig-zags, circles, triangles, and the query footwear impression requires it to be encoded in a similar manner. The process is laborious, time-consuming and can be the source of poor performance as the same pattern can be encoded differently by different users.

Although automatic classification of footwear prints is not yet practical, there are several published methods. A summary of the published retrieval methods and their performance is given in Table 1. Cumulative match score (CMS) is defined as follows. The identification process assumes a closed test, i.e., the true match is in the database. The input is compared to each entry in the database and the similarity scores are numerically ranked in descending order. If any of the top  $r = 1, 5, 10$  similarity scores corresponds to the input it is considered as a correct match. The percentage of times one of those  $r$  similarity scores is the correct match for all inputs is referred to as the CMS.

In early work, Mikkonen and Astikainen (1994) [20] proposed a classification system in which codes based on basic shapes are used as a pattern descriptor. Geradts and Keijzer (1996) [3] described an automatic classification for outsole designs using Fourier features. The approach employs shapes generated from footwear prints using image morphology operators. Spatial positioning and

**Table 1.** Survey of algorithms for automatic footwear print retrieval

Authors	Features	Performance (Cumulative Match Score)									Limitations	Dataset
		Full Prints			Partials			Crime Scene				
		@1 %	@5 %	@10 %	@1 %	@5 %	@10 %	@1 %	@5 %	@10 %		
deChazal, Flynn, et.al.(2005) [11]	Power spectral density (PSD)	64	87	90	50	70	77	-	-	-	No Scaling invariance	475 prints from For. Sci. Lab., Ireland
Zhang, Allinson (2005)[12]	Edge direction, FT histogram	85	95	97	-	-	-	-	-	-	No partials	512 prints
Pavlou, Allinson (2006)[13]	SIFT	86	90	93	85	90	92	-	-	-	No SoCs	368 prints of Forensic Sci. Serv., UK
Crookes, Bouridane, Su, Gueham (2007)[14]	Local Image Features (LIF)	100	100	100	100	100	100	-	-	-	Synthesized SoCs	500 clean prints, 50 degraded
Crookes, Bouridane, Su, Gueham (2007)[14]	Phase Only Correlation (POC)	100	100	100	100	100	100	-	-	-	No rotational invariance	100 clean prints, 64 synthetic
Gueham, Bouridane, Crookes (2008) [15]	POC	-	-	-	-	-	96	-	-	-	Tested with 200 prints	
Patil, Kulkarni (2009)[16]	Gabor transform	100	100	100	100	100	100	-	-	-	No SoCs (features rely on pixel intensities)	1400 clean full/partial & some synthetic noisy prints
Dardi, Cervelli, Carrato (2009)[17]	Texture	-	-	-	-	-	-	10	40	73	Tested with 87 known prints and 30 SoCs	87 known and 30 real SoC ENSFI
Tang, Srihari (2010)[18,19]	Shape Att. Relational Graph (ARG)	100	100	100	100	100	100	70	90	92	Slow speed (compensated by clustering)	1400 degraded, 1000 known & 50 real SoC

frequencies of shapes are used for classification with a neural network. No performance measures are reported. Alexander et al. (1999) [4] presented a fractal pattern matching technique with mean square noise error as a matching criteria to match the collected impression against database prints.

Fourier descriptors, which are invariant to translation and rotation, have also been used for classification of full and partial prints [21,11]. First and fifth rank classification are 65% and 87% on full-prints, and 55% and 78% for partials. The approach shows that although footwear prints are processed globally they are encoded in terms of the local information evident in the print. In [12] pattern edge information is employed for classification. After image de-noising and smoothing operations, extracted edge directions are grouped into a quantized set of 72 bins at five degree intervals. This generates an edge direction histogram for each pattern which after applying a Discrete FT provides a description with scale, translational and rotational invariance. The approach deals well with variations,

however query examples originate from the learning set and no performance is given for partial prints.

de Chazal et al. (2005) [11] proposed a fully automated shoe print classification system which uses power spectral density (PSD) of the print as a pattern descriptor. Here, PSD is invariant to translation and rotation of an image, crucial information of the print is preserved by removing the low and high frequency components and 2D correlation coefficient is used as similarity measure. Zhang and Allinson (2005) [12] proposed an automated shoe print retrieval system in which edge direction histogram is used to represent the shapes in shoes. The features consist of 1-D discrete Fourier Transform (FT) on the normalized edge direction histogram and the Euclidean distance is used as similarity measure.

Feature-point based methods, such as SIFT (Scale invariant feature transform) [22], have demonstrated good performance in general content-based image retrieval due to invariance with respect to scale, rotation and translation. However, they may be inappropriate for footwear impressions. This is partly because, as local extrema in the scale space, SIFT key points may not be preserved both among different shoes of the same class and through the life-time of a shoe. This problem is further complicated by the extremely poor quality and incompleteness of crime scene footwear impressions. Pavlou and Allinson (2006) [13] presented classification results where maximally stable extremal region (MSER) feature detectors are encoded with SIFT descriptors as features after which a Gaussian feature similarity matrix and Gaussian proximity matrix are used as the similarity measure. In some crime scenes, only partial shoe-prints (termed as “half prints” and “quarter prints”) are available. Partial matching has to focus on how to fully make use of regions available, with the accuracy of matching algorithms decreasing with print size.

Ghouti et al. (2006) [23] describe a so-called ShoeHash approach for classification where directional filter banks (DFB) are used to capture local/global details of shoe-prints with energy dominant blocks used as feature vector and normalized Euclidean-distance similarity. Su et al. (2007) [24] proposed a shoe-print retrieval system based on topological and pattern spectra, where a pattern spectrum is constructed using the area measure of granulometry, the topological spectrum constructed using the Euler number and a normalized hybrid measure of both used for matching. Crookes et al. (2007) [14] described two ways to classify shoe-prints: (i) in the spatial domain, modification of existing techniques: Harris-Laplace detectors and SIFT descriptors is proposed; the Harris corner detector is used to find local features; Laplace based automatic scale selection is used to decide the final local features and a nearest neighbor similarity measure, and (ii) in the transform domain, phase-only correlation (POC) is used to match shoe-prints. Gueham et al. (2008) [15] evaluated the performance of Optimum Trade-off Synthetic Discriminant Function (OTSDF) filter and unconstrained OTSDF filter in classifying partial shoe-prints.

As an exercise in data mining, Sun et. al. [25] clustered shoe outsoles using color (RGB) information as features where the number of clusters  $k$  was varied from 2 to 7 and the clustering results of  $k$ -means and expectation maximization

were compared; the results are of limited use since RGB information of outsole photographs are absent in impression evidence. Algarni and Hamiane (2009) [26] proposed a retrieval system using Hu's moment invariants as features and compared similarity measures: Euclidean, city block, Canberra and correlation distance. Xiao and Shi (2008) [27] presented matching using PSD and Zernike moments. Jing et al. (2009) [28] presented a new feature, directionality. Here, features extracted from co-occurrence matrix, Fourier transform and directional mask are matched using sum-of-absolute-difference. Nibouche et al. (2009) [29] proposed a solution for matching rotated partial prints. Harris points encoded with SIFT descriptors are used as features and they are matched using random sample consensus (RANSAC).

Dardi et al. (2009) [17] described a texture based retrieval system. A Mahalanobis map is used to capture texture and then matched using a correlation co-efficient measure. In subsequent work [30,31] they offer a cumulative match score comparison between Mahalanobis, [11] and [15]. Wang et al. (2009) [32] presented a wavelet and fuzzy neural network approach. Patil and Kulkarni (2009) [16] proposed using the Gabor transform to extract multi-resolution features and then Euclidean distance for matching. Rotation is estimated by the Radon transform and compensated by rotating in the opposite direction.

While footwear impression image analysis methods have been described in many papers, there are many gaps in the technology, e.g., among the many image processing and feature extraction algorithms it is not clear as to which ones are best suited for the task of retrieving reference images from a database (in response to a real crime scene query). More generally, methods are needed to: enhance image quality, represent patterns commonly found in footwear prints (that are also useful for comparison), determine the degree of similarity between evidence and known, retrieve closest matches in a reference data set, map comparison results to an opinion scale, and combine multiple scene images from the same source.

## 4 Proposed Methods and Algorithms

We describe here methods and algorithms for the following tasks:

1. Data sets for the design and evaluation of algorithms (Section 4.1).
2. Enhancing the quality of crime scene images for further processing (Section 4.2)
3. Representing footwear outsole patterns by extracting: (i) class characteristics and (ii) individualizing characteristics (Sections 4.3-4.5).
4. Similarity measures between patterns for use in comparison, retrieval and individualization (Section 4.6).
5. Search algorithms, including performance metrics and clustering of reference patterns. (Section 4.7).
6. Characterizing uncertainty of match between evidence and known (Section 4.8).

#### 4.1 Data Sets

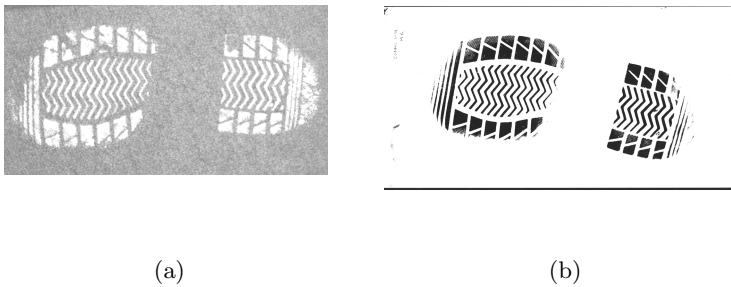
The development and evaluation of algorithms for any pattern analysis and recognition task critically depends upon the availability of data sets. They are used for training machine learning algorithms and for testing their performance. The data sets should ideally be representative of the population since the methods themselves are based on statistical analysis of the data. Three types of commonly used footwear print data sets are: digital images of outsoles provided by manufacturers, simulated crime scene images, and real crime scene images. Examples of such data sets are given below.

**Photographs of Outsoles.** Footwear manufacturers usually make available images of outsoles and uppers on commercial websites. A web crawler can visit a given set of vendor websites, and recover such images. An example of the types of images available is given in Fig. 1. About 10,000 such images were downloaded for the purpose of design and evaluation of algorithms.



**Fig. 1.** Digital images of footwear outsoles and uppers available on the web. The particular model shown is called “Nike Air Force 1” which is most often encountered in U. S. crime scenes.

**Simulated Scene Images.** The process of recovery of prints in a crime scene is described in [1]. To create simulated crime scene prints, people are asked to step on powder and then onto a carpet to create a simulated crime scene print. Then the picture of the print is taken with a forensic scale near the print. The resolution of the images is calculated using the scale in the images and then scaled to 100 dpi. The prints are also captured on chemical paper to create the reference print. A chemical print is the known print which is obtained by a person stamping on a chemical pad and then on chemical paper, which would leave clear print on the paper. The chemical prints are converted into digital camera images of resolution 100dpi examples of which are shown in Fig. 2. Since the simulated crime scene prints tend to be of relatively high quality this leads to over-optimistic results in verification and identification.



**Fig. 2.** Simulated crime scene and reference images: (a) print on carpet with powder, and (b) print on sheet of chemical paper

**Crime Scene and Reference Images.** Statistical models are best constructed from actual crime scene images and the reference data sets used with them. However these are generally hard to find. Some examples from a database of 350 crime images are shown in Fig. 3 together with the ground-truth in Fig. 5. Reference prints can be obtained by taking impressions of footwear outsoles provided by footwear vendors—some examples from a set of 5,000 reference prints are shown in Figure 4.

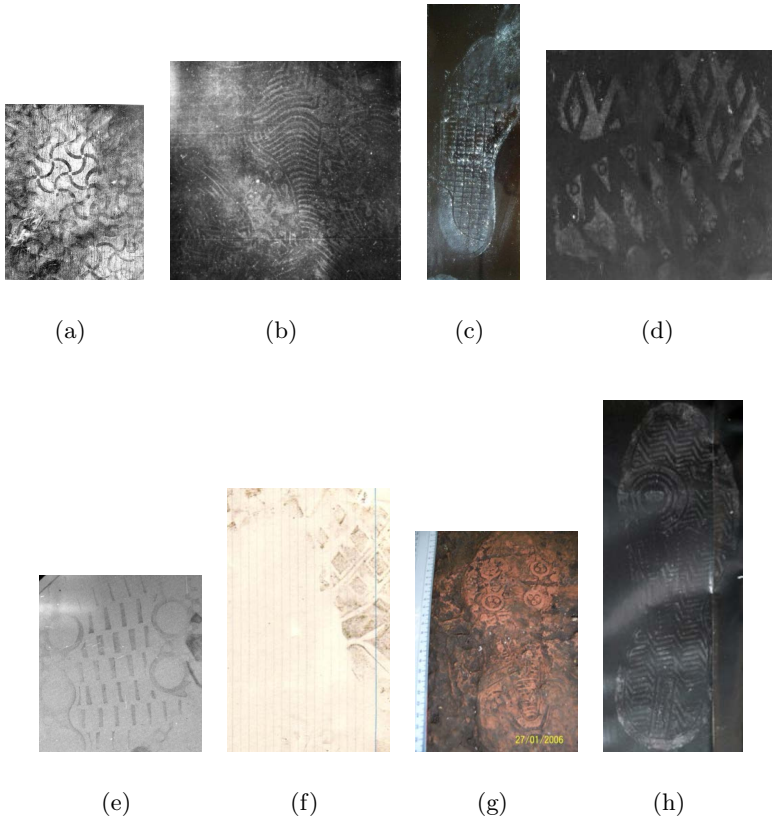
There are multiple prints from the same scene, e.g., in the first set 194 scene images are from 176 crime scenes and 144 scene images in the second are from 126 crimes. Each of the 50 scene images in the first dataset came from a different crime scene. Among these there are multiple shoe prints such as two partial shoe prints from the same crime scene, same marks taken at different illumination, same marks taken at different angles/orientation etc.

The resolution of reference images varies from 72 dpi to 150 dpi. Scene image resolution varies from 72 dpi to 240 dpi. The scene image dataset contains an equal number of color and gray-scale images. Only 3% of the reference images are direct photographs of the outsole of brand new shoes. The reference images can be broke down as follows. 97% are gray scale images. they are actually prints. 3% are color images, which are direct photographs of the outsole of the shoes on the market. Very few (less than 0.1%) are binary images.

## 4.2 Enhancing Image Quality

The matching of crime scene impressions to known prints largely depends on the quality of the extracted image from the crime scene impression. Thus the first step in dealing with both crime scene prints and database prints is that of processing them in a way that makes further processing more effective and/or efficient. Two approaches are: image labeling and edge detection. In image labeling, different pixels or regions of the image are labeled as foreground (impression) or background; it can be done using either thresholding or pixel classification.

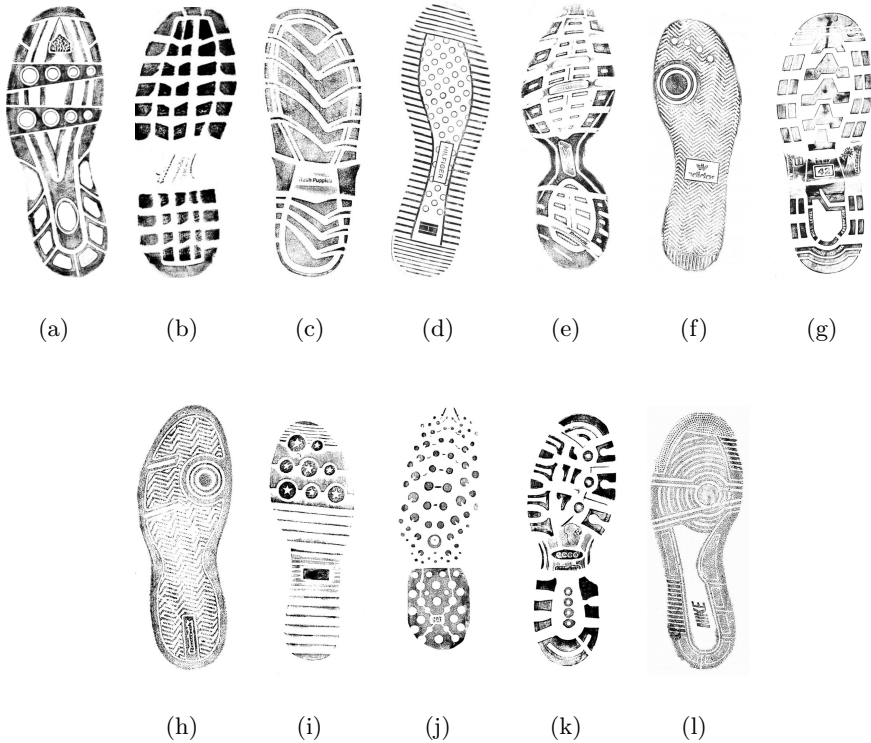




**Fig. 3.** Some crime scene images

**Thresholding.** One simple method of labeling images as foreground/background is global thresholding. A threshold value for the gray-scale is selected and all pixels with an intensity lower than this value are marked as background and all pixels with higher values are marked as foreground. Different strategies for determining the global thresholding value exist. A simplistic method, for example, models the intensities as a histogram with the assumption of two main intensity peaks (foreground and background), selecting a middle point as the threshold. A more sophisticated method is Otsu thresholding [33] which is based on a threshold which minimizes weighted within class variance. Another method is a neural network, e.g., one with two layers with four input neurons, three middle layer neurons and one sigmoidal output.

*Adaptive Thresholding.* A drawback of global thresholding is inability to cope with images that have a variety of intensities. An impression on carpet, for example, is often difficult to threshold since when the background is completely



**Fig. 4.** Some reference images (knowns)

below/above the chosen threshold value, large portions of the print will also be missing. A solution is adaptive thresholding. Instead of selecting a single threshold value for the entire image, it is dynamically determined throughout the image. This can cope with larger changes in background, such as variations in background material (carpet, flooring, etc.) and lighting. Such images often lack the separation of peaks necessary to use global thresholding. Smaller sub-images are much more likely to be more uniform than the image overall. It selects the threshold value for each individual pixel based on the local neighborhood's range of pixel intensities. For some  $n$  pixels around a given pixel, the thresholding value is calculated via mean, median, mean-C, etc. and used to determine whether a single pixel is part of the foreground or background, with different selections of sampling giving different results. After tuning the method to shoe-prints, this method gives high quality results at reasonable resolution. Some sample images are shown in Figure 6.

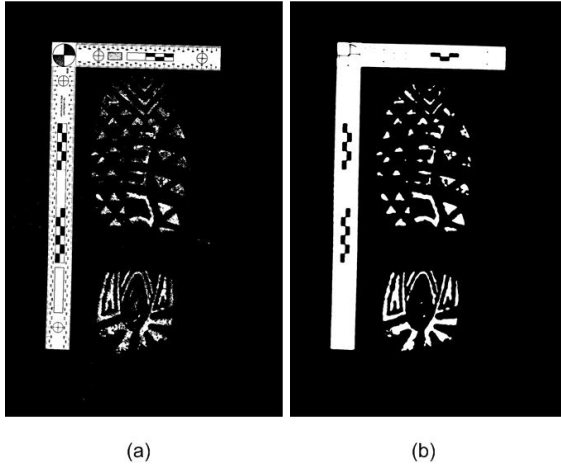
**CRF Classification.** While thresholding algorithms take into account only the value of the pixel or at most the surrounding pixels in making a decision,

Answers for "Crimes and Matches (144 images from 126 crimes Jan-Dec'07)" Test Data						
Test Image No	Brand	Model(s)	ImageID	Group Brand	Group Model(s)	Group ImageID
1	31-01-2007-01	K Swiss	Rinzler, Rinzler_Jewel, Rinzler SP	55907		
2	31-01-2007-02					
3	9174 jhm2	Adidas	Tholicke	22282		
4	8W90020-01	Lacoste	Selano, Kema	53151		
5	DSC_0010	Umbro	Kingston	62834		
6	DSC_0043_1-1	Puma	Race Cat J, Rep Cat Mid J, Speed Cat	15412		
7	FF		Charlie	450	Beta, Wilcox, Yerka	141
8	FOOTPRINT-355-01	Nike	Air Max Huarache 2	42280		
9	FOOTPRINT-355-02					
10	FOOTPRINT-361-01	Nike	Braun Canvas, Blazer Suede	27636	Pape Jeans	Action Leather Shoe 15877
11	FOOTPRINT-361-02				Umbro	Isol, Iso, ISP 7376
12	FOOTPRINT-367-01	Ascot	LK2049V	45562	Kowas	Golner, Colpat 54494
13	FOOTPRINT-367-02					
14	FOOTPRINT-370-01	Adidas	360 Shell	22065		
15	FOOTPRINT-374-01	Giorgio	Looser Boot	28847	Rugged Trail	Boot 41203
16	FOOTPRINT-374-02				Wrangler	Crew 28738
17	FOOTPRINT-374-03					
18	FOOTPRINT-394-1	Vans	Lancer Lani	44596		
19	FOOTPRINT-394-2					
20	FOOTPRINT-396-01	Nike	Air Force I, Air Force Low	42251		
21	FOOTPRINT-397-01	Reebok	Classic Leather Mustang	23422		
22	FOOTPRINT-397-02					
23	FOOTPRINT-397-03					
24	FOOTPRINT-398-01					
25	FOOTPRINT-401-01	Blox	Blaze, Houston	26825	Path	Yogi, Yogi Low Skate 31703
26	FOOTPRINT-401-02					
27	FOOTPRINT-401-03					
28	FOOTPRINT-401-04					
29	FOOTPRINT-401-05					
30	FOOTPRINT-402-01	Lacoste	Score, Score V, Score Strap	53147		
31	FOOTPRINT-402-02	Nike	Air Force I, Air Force Low	42251	Ellesse	Baham, Bologna Mid 75317
32	FOOTPRINT-404-01	uzq	Hero, Gothic, Shield (Zrocs)	58998		
33	FOOTPRINT-404-02					
34	FOOTPRINT-404-03					
35	FOOTPRINT-404-04					
36	FOOTPRINT-405-01	GrnSport	Contractor, Director, Combat	15205	Adlan	Boloi, Boltimore 30522
37	FOOTPRINT-407-01	PDQ	Ingleson	51241	MX2	Pizon 71727
38	FOOTPRINT-407-02					
39	FOOTPRINT-407-03					
40	FOOTPRINT-408-01	H-Tec	Miami	39071	H-Tec	Twister, Excalbur, Brooklands 1382
41	FOOTPRINT-415-01	Columbia	Blackrock, Thunderscout	19411	Pymmes	LP'16 39143
42	FOOTPRINT-422-01	Joe Boxer	Skate 47460	38373		
43	FOOTPRINT-422-02					
44	FOOTPRINT-423-01	Nike	Air Max Deluxe E	8150	Nike	Air Sentry Plus, Air Max Doro 37550
45	FOOTPRINT-427-01	Wilson	Prostaff 1000, Prostaff 710	45614	Wilson	Drawback 45615
46	FOOTPRINT-427-02					
47	FOOTPRINT-427-03					
48	FOOTPRINT-428-01	Globe	Motta, Flux, Falcon	21012		
49	FOOTPRINT-430-01	K-Swiss	Olsen, Farias	51185		
50	FOOTPRINT-434-01	Frank Wright	Weller	1753		
51	FOOTPRINT-435-01	Thom McAn		84608	40962	
52	FOOTPRINT-435-02					
53	FOOTPRINT-438-01	Lee		2046	22801	

Fig. 5. Ground Truth associated with crime scene images

the information contained in other areas of the print can be useful, in inferring whether a pixel belongs to the foreground or background. The contextual information from other regions can be incorporated using probabilistic models known as *conditional random fields (CRFs)* [34]. CRFs are partially directed probabilistic graphical models [35] which belong to a class of machine learning models known as discriminative models, as opposed to generative models which need a full joint distribution to be estimated before a decision can be made. CRFs have been successfully used in image segmentation problems including documents containing handwriting [36]. The model exploits the inherent long range dependencies that exist in the images and hence is more robust than approaches using neural networks and other binarization algorithms. Here we describe the application of the CRF model to labelling pixels in footwear print images.

Our task is to learn a mapping from image  $\mathbf{x}$  to labels  $y$ . Each  $y$  is a member of a set of possible image labels  $\mathcal{Y} = \{\text{Impression, Background}\}$ . The input image  $\mathbf{x}$  is segmented into  $m$  "patches"  $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ . The patch size is chosen to be small enough for high resolution and big enough to extract enough features.



**Fig. 6.** Adaptive thresholding results: (a) crime scene image (b) enhanced image using adaptive thresholding

We choose non-overlapping patches,  $3 \times 3$  pixels. A CRF is used to label each patch using the labels of the neighboring patches.

The probabilistic CRF model is as follows. Using the Hammersley-Clifford theorem  $p(y|\mathbf{x}) = \frac{1}{Z} \prod_i \phi_i(D_i)$  where the  $D_i$  are cliques of nodes in the graph with potentials  $\phi_i$ . Here node variables correspond to patches and labels. Assuming only up to pairwise clique potentials to be non-zero, the joint distribution over the labels  $y = \{y_1, y_2, \dots, y_m\}$  can be written as

$$p(y|\mathbf{x}) = \frac{1}{Z} \exp \left( \sum_j A_j(y_j, \mathbf{x}) + \sum_{(j,k) \in E} I_{jk}(y_j, y_k, \mathbf{x}) \right) \quad (1)$$

where  $Z$  is a normalizing constant known as the partition function and  $A_i$  and  $I_{ij}$  are the unary and pairwise potentials and  $E$  are edges in the graph. Thus we can define the conditional probabilistic model

$$P(y|\mathbf{x}, \theta) = \frac{e^{\psi(y, \mathbf{x}; \theta)}}{\sum_{y'} e^{\psi(y', \mathbf{x}; \theta)}} \quad (2)$$

where  $\theta$  consists of the model parameters and  $\psi(y, \mathbf{x}; \theta) \in \mathcal{R}$  is a potential function defined as :

$$\psi(y, \mathbf{x}; \theta) = \sum_{j=1}^m \left( A(j, y_j, \mathbf{x}; \theta^s) + \sum_{(j,k) \in E} I(j, k, y_j, y_k, \mathbf{x}; \theta^t) \right) \quad (3)$$

The first term in (3) is called the state term, sometimes called the *association potential* as mentioned in [37], and it associates the characteristics of that patch with its corresponding label.  $\theta^s$  are called the state parameters for the

CRF model. Analogous to it, the second term in (3) called the *interaction potential*, captures the neighbor/contextual dependencies by associating pair wise interaction of the neighboring labels and the observed data.  $\theta^t$  are called the transition parameters of the CRF model.  $E$  is a set of edges that identify the neighbors of a patch; a 24-neighborhood model was used.  $\theta$  comprises of the state parameters,  $\theta^s$  and the transition parameters,  $\theta^t$ .

The association potential can be modeled as  $A(j, y_j, \mathbf{x}; \theta^s) = \sum_i (f_i^{s2} \cdot \theta_{ij}^{s2})$  where  $f_i^{s2}$  is the  $i^{th}$  state feature for that patch and  $\theta_{ij}^{s2}$  is a state parameter. The state features used will be described shortly. In order to introduce a non-linear decision boundary, the state features,  $f_i^{s2}$  are obtained by transforming the input features  $f_i^{s1}$  by the *tanh* function to give the transformed state feature  $f_i^{s2} = \tanh(\sum_l (f_l^{s1}(j, y_j, \mathbf{x}) \cdot \theta_{il}^{s1}))$  where  $f_l^{s1}$  is the  $l^{th}$  state feature extracted for that patch; the transformed features are analogous to the outputs at the hidden layer of a neural network. The state parameters  $\theta^s$  are a union of the two sets of parameters  $\theta^{s1}$  and  $\theta^{s2}$ . The interaction potential  $I(\cdot)$  is an inner product between the transition parameters  $\theta^t$  and the transition features  $f^t$  is as follows:  $I(j, k, y_j, y_k, \mathbf{x}; \theta^t) = \sum_i (f_i^t(j, k, y_j, y_k, \mathbf{x}) \cdot \theta_{ijk}^t)$ .

*Parameter Estimation.* There are numerous ways to estimate the parameters of this CRF model [38]. In order to avoid the computation of the partition function the parameters are learnt by maximizing the likelihood of the data. Here we use conjugate gradient to maximize the likelihood. The maximum likelihood estimate of the parameters,  $\theta$ , based on a data set of size  $M$  is given by

$$\theta_{ML} = \arg \max_{\theta} \prod_{i=1}^M P(y_i | y_{N_i}, \mathbf{x}, \theta) \quad (4)$$

where  $P(y_i | y_{N_i}, \mathbf{x}, \theta)$ , which is the probability of the label  $y_i$  for a particular patch  $i$  given the labels of its neighbors,  $y_{N_i}$ , is

$$P(y | \mathbf{x}, \theta) = \frac{e^{\psi(y, \mathbf{x}; \theta)}}{\sum_a e^{\psi(y_i=a, \mathbf{x}; \theta)}} \quad (5)$$

where  $\psi(y_i, \mathbf{x}; \theta)$  is defined by (3). Note that (4) has an additional  $y_{N_i}$  in the conditioning set. This makes the factorization into products feasible as the set of neighbors for the patch from the minimal Markov blanket. It is also important to note that the resulting product only gives a pseudo-likelihood and not the true likelihood. The estimation of parameters which maximize the true likelihood may be very expensive and intractable for the problem at hand.

Combining (4) and (5), the log-likelihood is

$$\mathcal{L}(\theta) = \sum_{i=1}^M \left( \psi(y_i = a, \mathbf{x}; \theta) - \log \sum_a e^{\psi(y_i=a, \mathbf{x}; \theta)} \right). \quad (6)$$

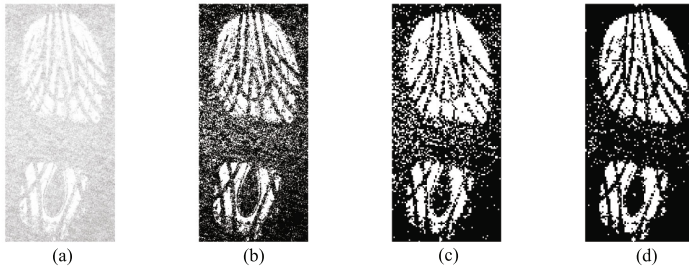
The parameters are estimated by maximizing the log-likelihood function in (6) using gradient descent.

*Features for CRF.* Since features depend on whether the print is powder on a carpet, mud on a table etc, a general definition of the texture of footwear-prints is difficult. Thus an interactive design is used where the user provides the texture samples of the foreground and background from the image. The sample size is fixed to be  $15 \times 15$  which is big enough to extract information and small enough to cover the print region. There could be one or more samples of foreground and background. The feature vector of these samples are normalized image histograms. There are four state features, the first two of which are derived from the probability distribution of gray levels in the patch: (i) entropy of the patch defined as  $E(P) = -\sum_i^n p(x_i) * \log(p(x_i))$ , and (ii) standard deviation of the patch  $STD(P) = \sqrt{\sum_i^n (x_i - \mu)^2}$ . The other two state features, which are based on cosine similarity between normalized image histogram vectors of two patches defined as  $CS(P_1, P_2) = \frac{P_1 * P_2}{|P_1| |P_2|}$ , are: (iii) the cosine similarity between the patch and the foreground sample feature vectors and (iv) the cosine similarity between the patch and the background sample feature vectors. The transition feature is the cosine similarity between the current patch and the surrounding 24 patches.

*Performance.* Pixel labeling performance of several different algorithms are shown in Fig. 7. It includes an example input image and results from each of three methods: Otsu thresholding, a neural network and a CRF. Both the neural network and CRF models used the same feature set other than the transition feature. The input images were converted from RGB jpeg format to grayscale, with a resolution of 100 dpi, before processing. Overall performance on a data set of 45 images (11 hand-truthed prints yielding 320,000  $3 \times 3$  patches for training and 34 images for testing), measured in terms of precision, recall and F-measure, are given in Fig. 7 (e). Precision, P, is defined as the percentage of the extracted pixels which are correctly labeled as foreground(shoe-print). Recall, R, is the percentage of the foreground successfully extracted. F-measure is the equally weighted harmonic mean of precision and recall i.e.,  $F = 2PR / (P + R)$ . Performance of Otsu thresholding is poor if either the contrast between the foreground and the background is less or the background is inhomogeneous. The neural network performs a little better by exploiting the texture samples that the user provided. CRF tends to outperform both by exploiting the dependency between the current patch and its neighborhood, i. e., if a patch belongs to foreground but is ambiguous, the evidence given by its neighborhood patches helps in deciding its polarity.

**Edge Detection.** Rather than labeling pixels in the gray-scale image to convert to a binary image, an alternative is to detect sharp discontinuities, or edges in the input image, as the starting point. Edge detection has a firm basis in biological vision and has been studied extensively. Edges in the image can be used to detect more global geometrical patterns as described in Section 4.4.

Among various edge detectors the Canny edge detector [39] has been shown to have many useful properties. It is considered to be the most powerful edge



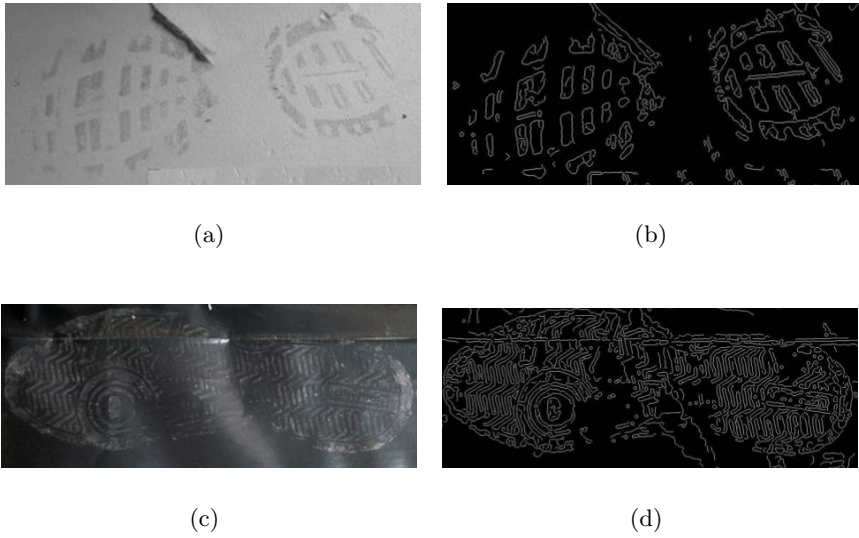
Method	Precision	Recall	F-measure
Otsu	40.97	89.64	56.24
Neural Network	58.01	80.97	59.53
CRF	52.12	90.43	66.12

(e) Summary Results

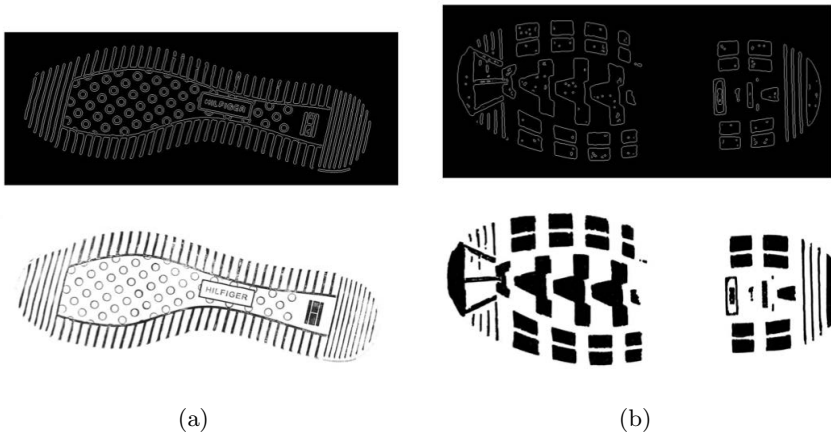
**Fig. 7.** Results of three image pixel labeling methods: (a) an input crime scene test image, (b) result obtained by applying Otsu thresholding, (c) result of neural network thresholding, and (d) result of CRF labeling. Summary of results with 34 test images are tabulated in (e) whose columns correspond to retrieval performance metrics (precision, recall and F-measure percentages)

detector since it uses a multi-stage algorithm consisting of noise reduction, gradient calculation, non-maximal suppression and edge linking with hysteresis thresholding. The detected edges preserve the most important geometric features on shoe outsoles, such as straight line segments, circles, ellipses. The results of applying the Canny edge operator to crime scene images is shown in Fig. 8. Results with some database images are shown in Fig. 9.

Prior to edge detection, morphological operations are performed on database images [40]. The morphological operations are: dilation, erosion and filling holes in the binary image. The result is a more exact region boundary that improves the quality of edge detection. Morphological operations play a vital role in fetching the exact contours of the different shapes like line, ellipse and circle. We perform morphological operations (dilation and erosion) to make the interior region of the boundary uniform and then extract the boundary using Canny edge detection. Since the interior region is uniform, canny edge detector does not detect any edges inside the boundary and it improves the quality of edge detection. Specifically, each database shoe-print is processed in the following order: Edge detection  $\rightarrow$  Dilation  $\rightarrow$  Erosion  $\rightarrow$  Flood fill  $\rightarrow$  Complement. This procedure is illustrated using a sample print in the Fig. 10(a-f). As shown in Fig. 10(g), the edge image of the enhanced print has much better quality for feature extraction.



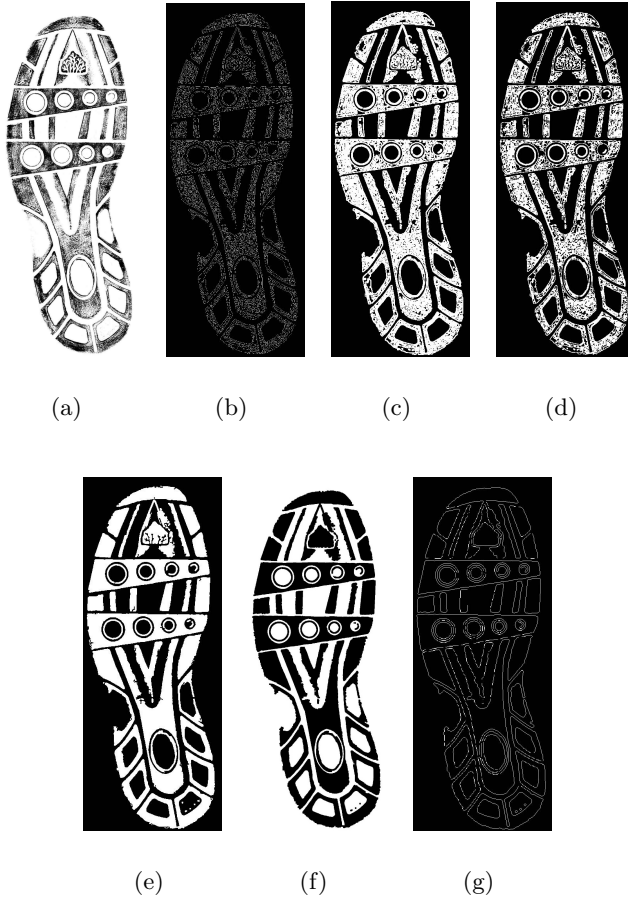
**Fig. 8.** Results of applying edge detection to crime scene images. Two pairs of images are shown corresponding to input and edge image: (a,b), (c,d).



**Fig. 9.** Results of edge detection on two reference images

Dilation and erosion make the interior region of the boundary uniform and then extract the boundary using edge detection. Since the interior region is uniform the edge detector does not detect any edges inside the boundary. Edge detection showing the intermediate results of morphological operations is shown in Figure 11. Database Prints are subject to the sequence: Edge Detection, Morphological Operation and Edge Detection. Crime Scene Prints are subjected to



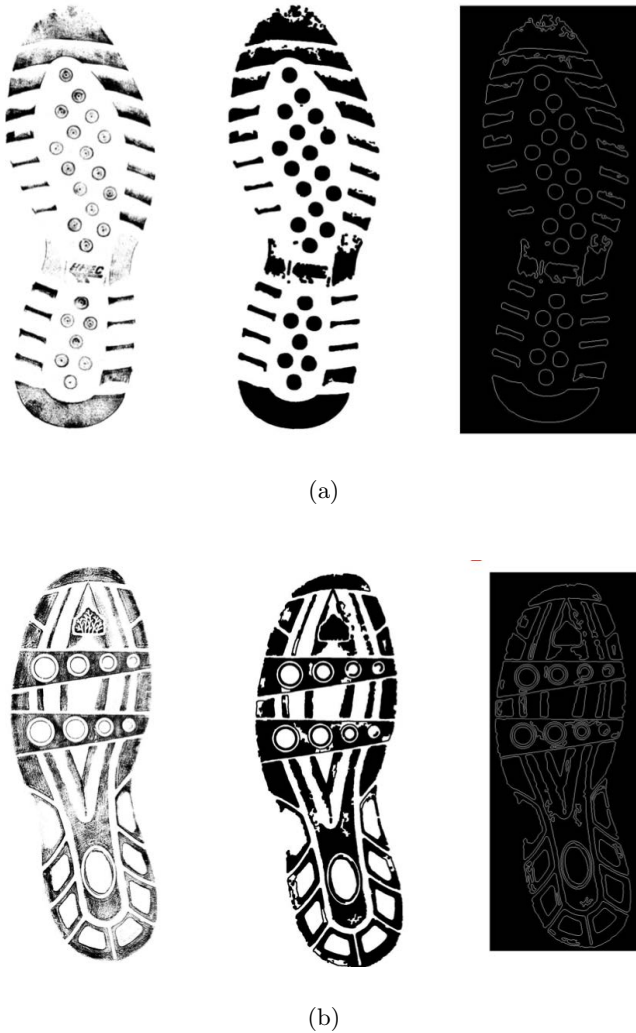


**Fig. 10.** Morphological operations for image enhancement: (a) input, (b) edge image, (c) after dilation, (d) after erosion, (e) after flood fill, (f) after complement, which is the final output, and (g) edge image of enhanced print

only Edge Detection. For crime scene prints, because of their poor quality, we directly extract features from the edge image of original image. It takes 4-5 seconds to process one image on a desktop computer.

### 4.3 Characteristics of Outsole Patterns

Discriminating characteristics of outsole patterns and footwear impressions can be classified into two categories: those acquired during the manufacturing process and those acquired from wear. Manufacturing features are those that come from the manufacturing process, which include design patterns and defects. Acquired



**Fig. 11.** Results of edge detection showing intermediate morphological operations on two data base images

features refer to attributes that develop during the lifetime of the footwear, such as wear pattern and damage.

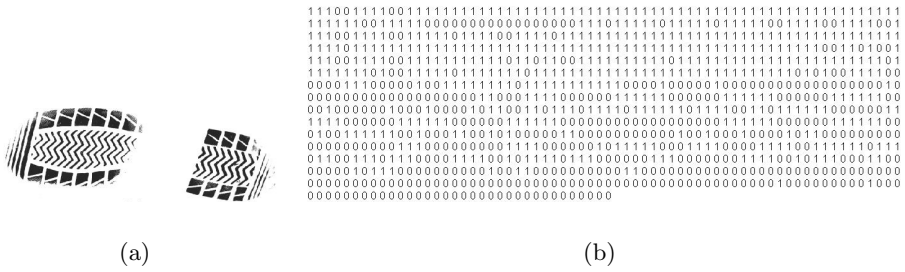
As in any pattern comparison task, the first step for matching a query print against a reference print is a representation in terms of characteristics. The ideal representation would allow discrimination between different outsoles but also be invariant to various transformations such as rotation, translation, distortion and noise. Once a set of characteristics are determined there is also a need for a suitable measure of similarity between feature sets.

Color, texture and shapes of primitive elements are commonly used to recognize objects in computer vision [41]. However color is absent here since acquired impression prints are gray-scale images. Textures are sensitive to acquisition methods and susceptible to wear while shapes are resistant to wear and present over a long period of time. Shape features are also robust against occlusion and incompleteness, i.e., the wear or variation of a local region on the outsole will be less likely to affect shape features in other regions.

We discuss here three different types of characteristics for representing outsole patterns. Associated with each is a similarity measure for comparison of two inputs. The methods are GSC, SIFT and geometrical patterns represented as an attribute relational graph.

**GSC.** In the field of document analysis, the central task is that of recognizing two-dimensional patterns such as characters. Many different features have been developed and one that we have used with success for handwriting recognition and writer verification are the GSC (gradient, structural, concavity) features. The GSC features are based on detecting local, intermediate and global features (see Fig. 12) [42]. The basic unit of an image is the pixel and we are interested in its relationships to neighbors at different ranges from local to global. In a sense, we are taking a multi-resolution approach to feature generation. GSC features are generated at three ranges: local, intermediate and global. In the basic approach the feature vector consists of 512 bits corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) features. Each of these three sets of features rely on dividing the scanned image into a  $4 \times 4$  region. Gradient features capture the frequency of the direction of the gradient, as obtained by convolving the image with a Sobel edge operator, in each of 12 directions and then thresholding the resultant values to yield a 192-bit vector. The structural features capture, in the gradient image, the presence of corners, diagonal lines, and vertical and horizontal lines, as determined by 12 rules. Concavity features capture, in the binary image, major topological and geometrical features including direction of bays, presence of holes, and large vertical and horizontal patterns. The input shoe-print is represented as two  $4 \times 4$  regions or a fixed-dimensional (1028-bit) binary feature vector. The similarity between two GSC feature vectors is computed using a correlation measure.

**SIFT.** In the field of computer vision a popular algorithm for detecting key features of three-dimensional objects in digital images is the scale invariant feature transform (SIFT) [43]. The objective of SIFT is to extract and describe invariant features from images that can be used to perform matching between different views of an object in a scene. Four major steps of the algorithm are: scale-space extrema detection, key point localization, orientation assignment and key-point descriptor construction. The scale space is constructed by convolving the input image with a Gaussian function and resampling the smoothed image. Maxima and minima are determined by comparing each pixel in the pyramid to its 26 neighbors (in a  $3 \times 3$  cube). These maxima and minima in the scale space are called



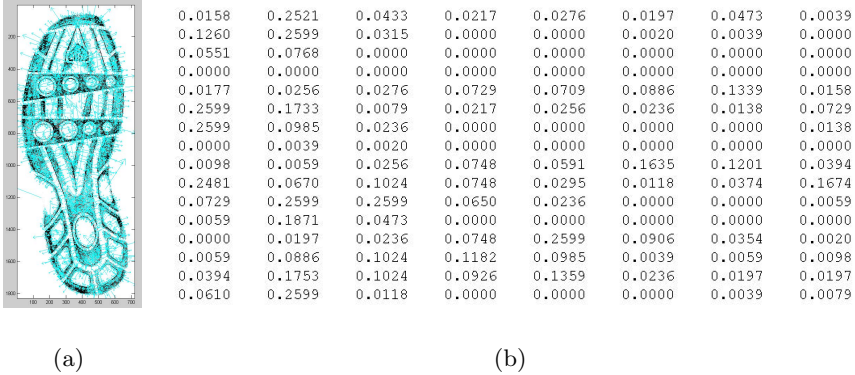
**Fig. 12.** Representation of an outsole pattern using features (GSC) designed for two-dimensional shapes in document analysis: (a) input impression which is characterized by (b) a 1,024-dimensional GSC binary feature vector

as key points, which are in turn described by a 128-dimensional vector: a normalized description of gradient histogram of the region around that key-point. The number of key points detected by the SIFT algorithm varies from image to image. Key-points of a shoe-print image are shown in Fig. 13(a) where there are 15,499 key-points. One such key-point descriptor is shown in Fig. 13(b). The similarity between two descriptors is computed using the Euclidean distance between two 128-dimensional vectors and the similarity between two images is the number of key-points that match. SIFT is commonly used in content-based image retrieval and is said to be used in Google image search.

**Performance with GSC and SIFT.** GSC features, which are designed for two-dimensional patterns, are very fast and work well with complete shoe-prints but break-down when prints are partial; a fix can be made by detecting whether the print is partial. SIFT features, which are designed for three-dimensional objects, work better than GSC for partial prints, particularly since they were designed to handle occlusion in scenes. SIFT is invariant to transformations of scale, rotation and translation of shoe-prints [22]. However, due to local extrema in the scale space, SIFT key-points are not preserved both among different shoes of the same class and throughout the lifetime of a single shoe. A representation based on geometrical patterns in a graph works significantly better than SIFT in retrieval as described next (see Fig. 34).

#### 4.4 Geometrical Patterns

Patterns of outsoles usually contain small geometrical patterns involving short straight line segments, circles and ellipses. An analysis of 5,034 outsole prints revealed that 67% have only line segments (some examples are shown in Fig. 14, where the line segments have a minimum length of 25 pixels), 1.5% have only circles (Fig. 15), 0.004% have only ellipses (Fig. 16), and 24% are combinations of lines, circles and ellipses. The principal combination of shapes are lines-circles



**Fig. 13.** Representation of an outsole pattern using features (SIFT) designed for three-dimensional objects in computer vision: (a) input image annotated by key-points where each blue arrow shows key-point orientation extracted by the SIFT algorithm, and (b) descriptors for one key-point

which constitute 16% (Fig. 17), lines-ellipses constitute 6% (Fig. 18), circles-ellipses-0.1% (Fig. 19) and lines-circles-ellipses-0.7% (Fig. 20). Texture patterns (Fig. 21) constitute the remaining 8%. The complete distribution is given in Table 2. This analysis shows that the three basic shapes are present in 92% of outsole prints. Furthermore, patterns other than circles and ellipses can be approximated by piecewise lines.

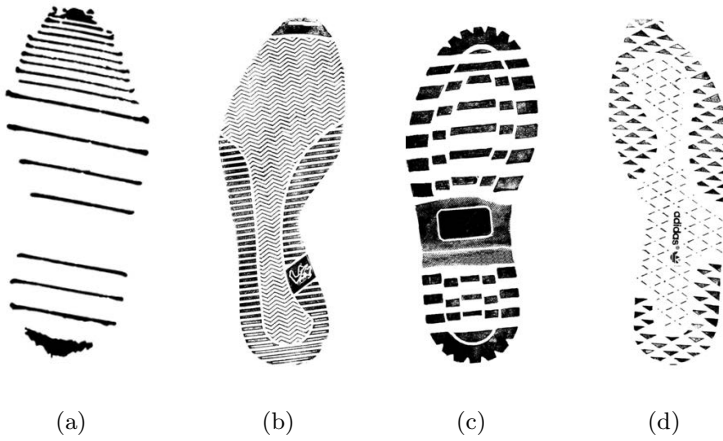
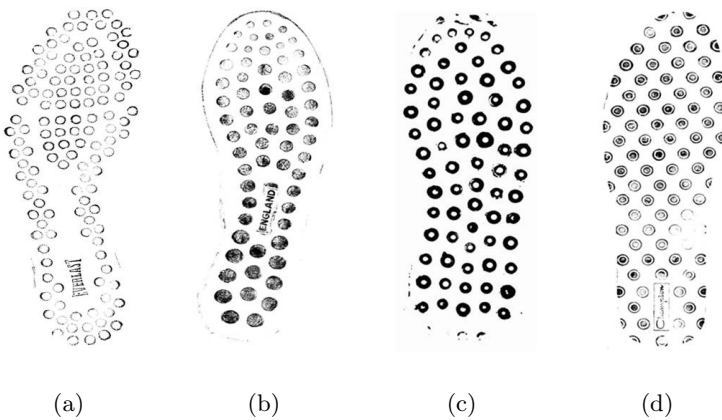
When projected on to a plane, most man-made objects can be represented as combinations of straight line and ellipse segments. Mathematically, straight line segments and circles are special cases of ellipses. An ellipse with zero eccentricity is a circle and an ellipse with eccentricity of 1 is a straight line; where the eccentricity of an ellipse is defined as  $\sqrt{1 - (b/a)^2}$  where  $a$  and  $b$  are the lengths of the semi-major and semi-minor axes.

While an ellipse detector alone can capture 92% of the primitive shapes, we choose to use specialized detectors for straight lines and circles since they are more efficient. The feature extraction approach is to detect the presence, location and size of three basic shapes: *straight line segments*, *circles/arcs* and *ellipses*. Since all three are geometrical shapes with simple parametric representations, they are ideal for the application of a robust method of detecting shapes.

The Hough transform[44] is a method to automatically detect basic geometrical patterns in noisy images. It detects features of a parametric form in an image by mapping foreground pixels into parameter space, which is characterized by an  $n$  dimensional accumulator array, where  $n$  is the number of parameters necessary to describe the shape of interest. Each significant pixel from the shape of interest would cast a vote in the same cell of an accumulator array, hence all pixels of a shape gets accumulated as a peak. The number of peaks corresponds to the number of shapes of interest in the image. Originally designed for

**Table 2.** Distribution of geometric patterns in a database of footwear outsole prints

Fundamental Patterns	No. of Prints
Line segments	3397
Lines & Circles	812
Lines & Ellipses	285
Only Circles/Arcs	73
Lines, Circles & Ellipses	37
Only Ellipses	15
Circles & Ellipses	5
Texture	410
<b>Total - 5034 prints</b>	

**Fig. 14.** Footwear outsole patterns containing line segments only**Fig. 15.** Footwear outsole patterns containing circles only

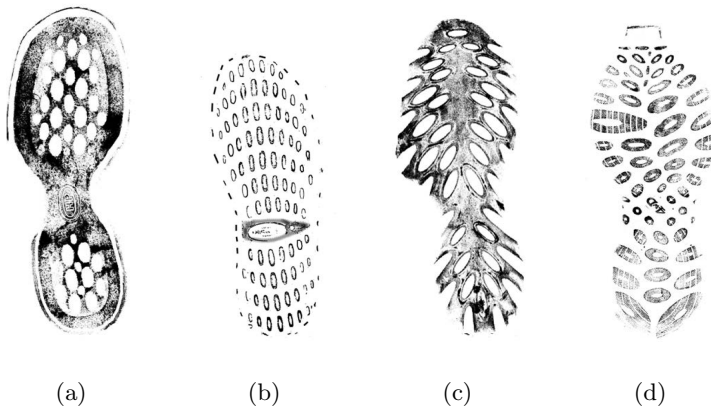


Fig. 16. Footwear outsole patterns containing ellipses only

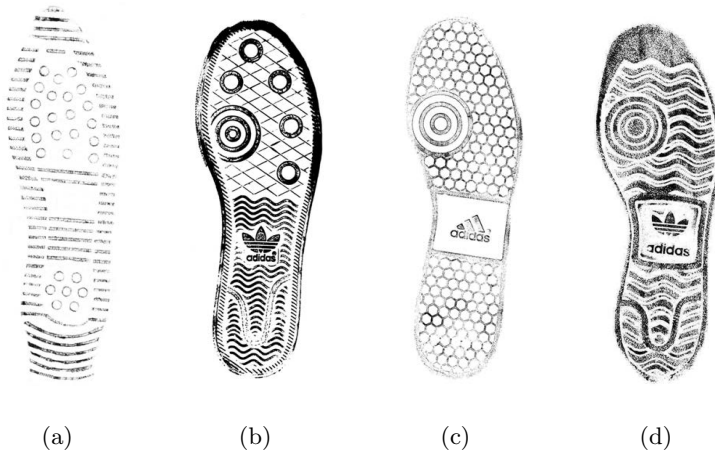
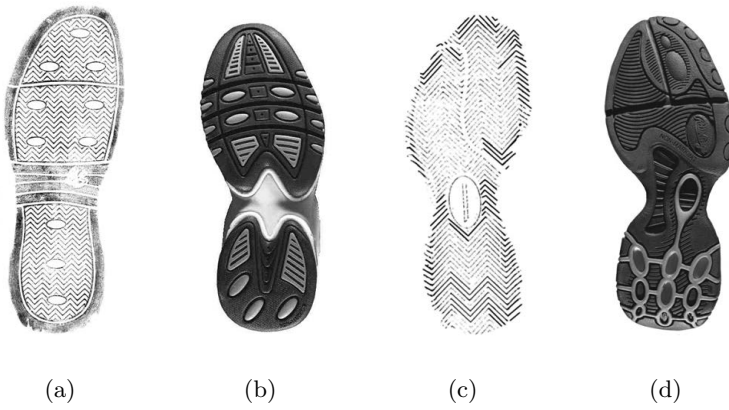
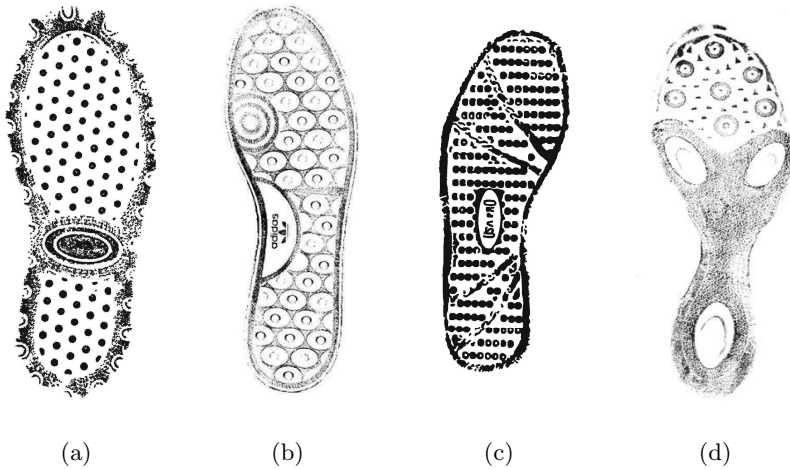


Fig. 17. Footwear outsole patterns containing lines and circles

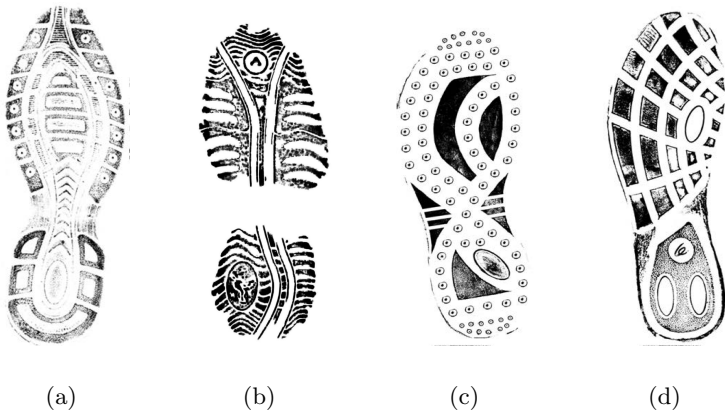


**Fig. 18.** Footwear outsole patterns containing lines and ellipses

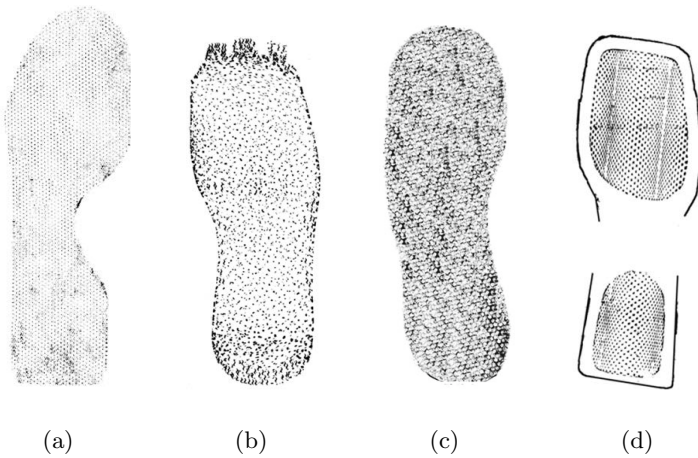


**Fig. 19.** Footwear outsole patterns containing circles and ellipses





**Fig. 20.** Footwear outsole patterns containing lines, circles and ellipses



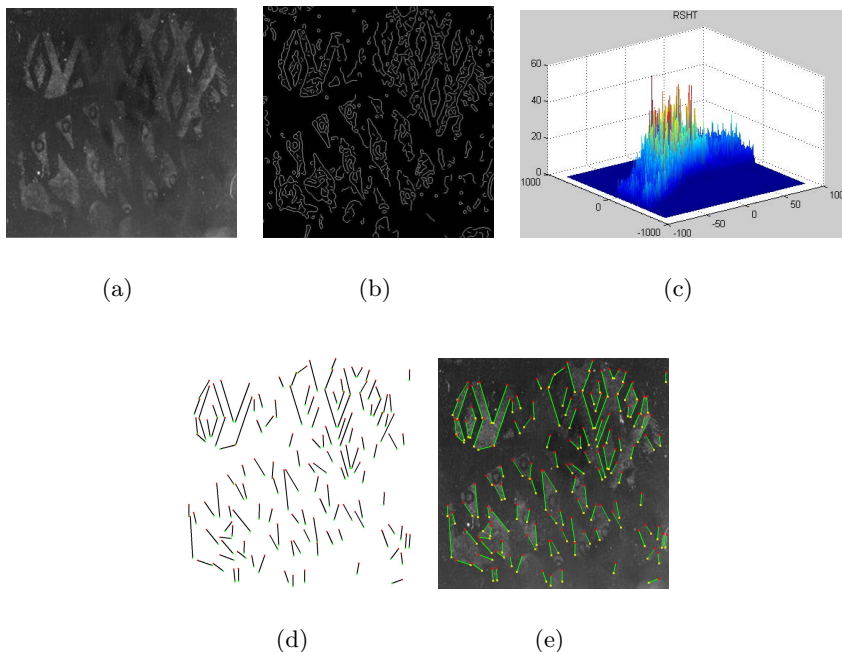
**Fig. 21.** Footwear outsole patterns containing texture only

detecting straight lines in cloud chamber photographs and later generalized to circles and ellipses, the Hough transform has found success in many applications such as detecting cancerous nodules in radiological images and structure of textual lines in document images[45].

1. *Line Segments*: Using the polar coordinate system, a straight line can be represented by two parameters  $r$  and  $\theta$ . The Hough transform maps each pixel in the Cartesian  $x$ - $y$  plane to a 2-dimensional accumulator array using the transformations defined by  $x = r\cos\theta$  and  $y = r\sin\theta$ . The values of  $r$  and  $\theta$  at which the accumulator elements peak represent the presence of straight lines.
2. *Circles*: It involves building a 3-dimension accumulator array corresponding the center coordinates and the radius. Gradient orientation is used to limit the generation of spurious votes. Further, spatial constraints are used to identify spurious circles. Gradient orientation is used to limit the generation of spurious votes[46]. Further, spatial constraints are used to eliminate spurious circles.
3. *Ellipses*: In a Cartesian plane, an ellipse can be described by its centre  $(p, q)$ , length of the semi-major axis  $a$ , length of the semi-minor axis  $b$  and the angle  $\theta$  between the major axis and the  $x$ -axis. Thus five parameters  $(p, q, a, b, \theta)$  are required to uniquely describe an ellipse[47]. These five parameters demand a five-dimensional accumulator which is computationally expensive but the Randomized Hough transform (RHT) [48] for ellipse detection is more efficient.

We describe next algorithms for lines and ellipses based on the Hough transform; since the circle is a special case of the ellipse the same algorithm can be used.

**Line Detection.** The Standard Hough Transform (SHT) to detect lines consists of three steps: transform and accumulation, peak selection, and line verification. However, most scenes have complex geometric structures. The number of line segments in a scene image of moderate size (say  $1000 \times 1000$ ) can be several hundred. Each set of collinear points votes for a peak in accumulator. Detecting all the true peaks accurately while suppressing spurious ones is difficult. In addition, short line segments are easily missed, which may be useful for discriminating similar print structures. Since the standard Hough transform (SHT) cannot be applied directly, an iterative procedure is used to remove interference in peak selection, using a verification criterion. First, connected components are labeled in the edge image. For each component, the Hough transform is applied and peaks are detected. When a peak is identified and the line segments are extracted, pixels contributing to those line segments are eliminated from the edge image, and an updated accumulator is obtained by applying SHT on the modified edge image. The process of extracting straight line segments in a crime scene impression is shown in Figures 22.



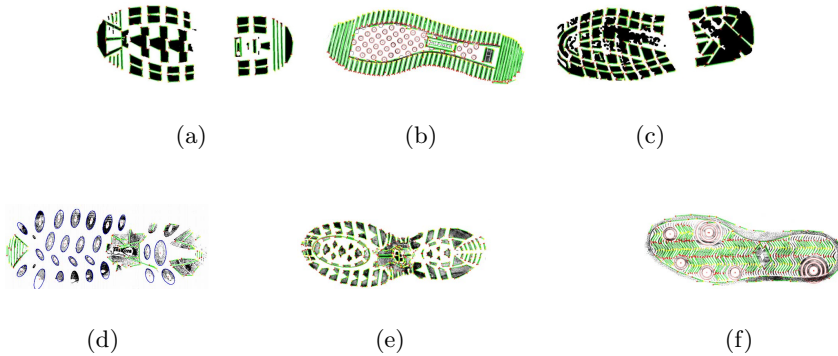
**Fig. 22.** Detecting line segments using the restricted straight line Hough transform: (a) input crime scene image, (b) edge detected image (c) accumulator histogram, (d) detected line segments, and (e) line segments overlaid on the input image

**Ellipse Detection.** The ellipse is a fundamental shape in both natural and man-made objects and hence frequently encountered in images. Existing ellipse detection algorithms, viz., randomized Hough transform (RHT) and multi-population genetic algorithm (MPGA), have disadvantages. The RHT performs poorly with multiple ellipses and MPGA has a high false positive for complex images. The proposed algorithm selects random points using constraints of smoothness, distance and curvature. In the process of sampling, parameters of potential ellipses are progressively learnt to improve parameter accuracy. New probabilistic fitness measures are used to verify ellipses extracted: ellipse quality based on the Ramanujan approximation and completeness. Experiments on synthetic and real images show performance better than RHT and MPGA in detecting multiple, deformed, full or partial ellipses in the presence of noise and interference. A detailed description of the algorithm is given in [19].

Results of extracting circles and ellipses in data base prints are shown in 23.

#### 4.5 Graph Representation

Structural representations have long been used in computer vision to represent complex objects and scenes for image matching [49]. Graph representations have



**Fig. 23.** Shapes detected in reference images: lines, circles and ellipses are shown in green, red and blue respectively

a great advantage over feature vectors because of they can explicitly model the relationship between different parts and feature points [50].

After detecting their presence, the impression image is decomposed into a set of primitives. To obtain a structural representation of these primitives, an *attributed relational graph* (*ARG*) [51,52] is built. An *ARG* is a directed graph that can be represented as a 3-tuple  $(V, E, A)$  where  $V$  is the set of vertices, also called nodes,  $E$  is the set of relations (edges) and  $A$  is the set of attributes. Each edge describes the spatial relationship between a pair of nodes. The attributes include node attributes (unary) and edge attributes (binary).

There are three types of nodes, corresponding to lines (L), circles (C) and ellipses (E), and nine types of edges: line-to-line (L2L), line-to-circle (L2C), line-to-ellipse (L2E), circle-to-circle (C2C), circle-to-ellipse (C2E), ellipse-to-ellipse (E2E), circle-to-line (C2L), ellipse-to-line (E2L) and ellipse-to-circle (E2C). Attributes of nodes and edges should be defined such that they are scale/rotation invariant, and capture spatial relationships such as distance, relative position, relative dimension and orientation.

Three attributes are defined for nodes which represent the basic shapes detected.

1. *Quality* is the ratio of the number of points on the boundary of the shape (perimeter pixels) to the perimeter of the shape.
2. *Completeness* is the standard deviation of the angles of all on-perimeter pixels with respect to the center of circle/ellipse, *std*, normalized as  $std/180$ . If a wide range of angles are present, implying that most of the shape is represented, there will be more angles represented and this value is high, while a partial figure will have smaller diversity of angles and this value will be low. While the range of angles is 0 to 360 for circles and ellipses, for a straight line there are only two angles with respect to the center, 0 and 180.

3. *Eccentricity* is the degree of elongation, defined as the square root of 1 minus square of ratio of minor to major axes. For a circle eccentricity is 0 and for a straight line eccentricity is 1.

Edge attributes are dependent upon the pair of shapes they connect. They use the relative position definitions between lines, circles and ellipses. Some attributes are normalized to the range [0,1] using the sigmoid function. A complete list of node and edge attributes is given in Figure 24.

So as to handle missing nodes or incorrectly detected nodes, which may arise due to noise, occlusion and incompleteness, a *fully-connected graph* is used. If for the sake of computational efficiency we consider only local relationships, as is often done in Markov models, it would lead to poor results since the only image components discernible in a print may be those at the extremities.

This means that there is a directed edge from each node to all nodes including itself; a node is connected to itself because we can use a general formula for computing the cost between two graphs. Thus in a directed graph with  $N$  nodes there will be  $N + 2(N(N - 1)/2) = N^2$  edges. The number of attributes at each edge depends on the types of nodes it connects.

The ARG for a scene image is shown in Fig. 25; the values of node and edge attributes for a portion of the subgraph with four nodes are given in Table 3.

**Table 3.** Node and Edge Attributes for four-node subgraph shown in Figure 25(d)

Nodes and Edges	Attributes
Node 1	[0.0000, 0.7468, 0.5699]
Node 2	N/A
Node 3	N/A
Node 4	N/A
$E_{11}$	[0.5000, 0.0000, 0.0000]
$E_{12}$	[0.4285, 0.1976, 0.5989]
$E_{13}$	[0.4593, 0.1976, 0.3195]
$E_{14}$	[0.4809, 0.1387, 0.2316]
$E_{21}$	[0.5715, 0.1976, 0.5989]
$E_{22}$	[0.0000, 0.5000, 0.0000, 0.0000, 0.0200]
$E_{23}$	[0.0000, 0.5312, 0.0584, 0.0000, 0.0200]
$E_{24}$	[0.0323, 0.5527, 0.0609, 0.0146, 0.0626]
$E_{31}$	[0.5407, 0.1976, 0.3195]
$E_{32}$	[0.0000, 0.4688, 0.0584, 0.0000, 0.0200]
$E_{33}$	[0.0000, 0.5000, 0.0000, 0.0000, 0.0200]
$E_{34}$	[0.0324, 0.5217, 0.0091, 0.0090, 0.1018]
$E_{41}$	[0.5191, 0.1387, 0.2316]
$E_{42}$	[0.0323, 0.4473, 0.0609, 0.0085, 0.0901]
$E_{43}$	[0.0324, 0.4783, 0.0091, 0.0091, 0.0903]
$E_{44}$	[0.0000, 0.5000, 0.0000, 0.0000, 0.0200]

## 4.6 Graph Similarity

Central to both retrieval and identification is a method for computing similarity between images. Equivalently, the inverse of similarity is a distance measure. The choice of similarity or distance measure is important since it influences the retrieval result, uncertainty of match, and the quality of clusters in partitioning the database for efficiency.

L2L				L2E			
Att	Definition	Normalization	Weight	Att	Definition	Normalization	Weight
$N-a$	$\frac{ L1.\theta - L2.\theta }{180}$	-	0.4472	$N-rs$	$\frac{L.len}{L.len + E.ER}$	-	0.5
$N-rs$	$\frac{L1.len}{L1.len + L2.len}$	-	0.4472	$rd$	$\frac{dist(L.m, E.cen)}{L.len + E.ER}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.5
$rd$	$\frac{dist(L1.m, L2.m)}{L1.len + L2.len}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.4472	$rp1$	$\frac{\min(OA, OB)}{\max(OA, OB)}$	$\frac{rp_1 + 1}{2}$	0.5
$pd$	$\frac{dist(L1.m, L2)}{L1.len + L2.len}$	$\frac{pd}{\sqrt{1 + pd^2}}$	0.4472	$rp2$	$\frac{\min( OA ,  OB )}{\max( OA ,  OB )}$	$\frac{rp_2 + 1}{2}$	0.5
$rp1$	$\frac{\min(OA, OB)}{\max(OA, OB)}$	$\frac{rp_1 + 1}{2}$	0.4472	$N-ro$	$\frac{ L.\theta - E.\theta }{180}$	-	0.5
$rp2$	$\frac{\min( OA ,  OB )}{\max( OA ,  OB )}$	$\frac{rp_2 + 1}{2}$	0.4472	C2L			
C2E				$N-rs$	$\frac{C.r}{C.r + E.ER}$	-	0.5774
$N-rs$	$\frac{C.r}{C.r + E.ER}$	-	0.5774	$rd$	$\frac{dist(C.cen, L)}{C.r}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.5774
$rd$	$\frac{dist(C.cen, E.cen)}{C.r + E.ER}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.5774	$rp$	$\frac{\min(S1, S2)}{\max(S1, S2)}$	$\frac{rp + 1}{2}$	0.5774
$rp$	$\frac{\min(OA, OB)}{\max(OA, OB)}$	$\frac{rp + 1}{2}$	0.5774	L2C			
E2E				$N-rs$	$\frac{L.len}{C.r + L.len}$	-	0.5774
$e\_ratio$	$\frac{E1.e}{E1.e + E2.e}$	-	0.2236	$rd$	$\frac{dist(C.cen, L)}{C.r}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.5774
$f(1e)$	$\frac{E1.e - E2.e + 1}{2}$	-	0.2236	$rp$	$\frac{\min(S1, S2)}{\max(S1, S2)}$	$\frac{rp + 1}{2}$	0.5774
$rd$	$\frac{dist(E1.cen, E2.cen)}{E1.ER - E2.ER}$	$\frac{rd}{\sqrt{1 + rd^2}}$	0.4472	C2C			
$N-rs$	$\frac{E1.ER}{E1.ER + E2.ER}$	-	0.4472	$N-rs$	$\frac{C1.r}{C1.r + C2.r}$	-	0.7071
$N-ro$	$\frac{ E1.\theta - E2.\theta }{90}$	-	0.4472	$rd_1$	$\frac{dist(C1.cen, C2.cen)}{C1.r + C2.r}$	$\frac{rd_1}{\sqrt{1 + rd_1^2}}$	0.7071
$rp$	$rp(E1.major-axis, E2.major-axis)$	$\frac{rp + 1}{2}$	0.4472	$rd_2$	$\frac{dist(C1.cen, C2.cen)}{C1.r - C2.r}$	$\frac{rd_2}{\sqrt{1 + rd_2^2}}$	0.0

### Node Attributes

Node	Attributes	Definition
Circle	Completeness	Standard deviation of the angle that all on-circle pixels make with respect to the center.

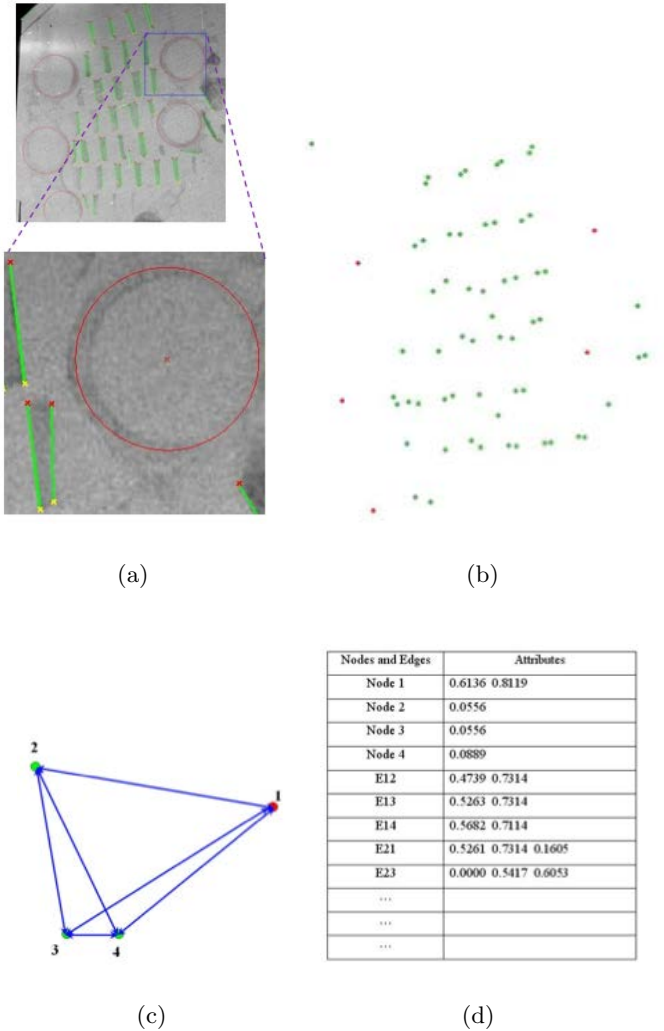
Node	Attributes	Definition
Circle	Quality	Number of pixels on circle Circumference of circle

Node	Attributes	Definition
Ellipse	Eccentricity	$\sqrt{1 - \frac{b^2}{a^2}}$

### Symbols and its Definition

L	Line segment	r	Radius	e	eccentricity	C	Circle	len	Length	att	attributes	max	maximum
E	Ellipse	m	Mid-point	cen	Centre	pd	Perpendicular distance	ER	$\sqrt{a \cdot b}$	N	Normalized		absolute
dist	Euclidean distance	rp	Relative position	rd	Relative distance	rs	Relative size	$\theta$	orientation	a, b	Semi-major axis and semi-minor axis of the ellipse respectively.	p, q	Center of the ellipse

**Fig. 24.** Definitions of node and edge attributes in attribute relational graph where nodes correspond to geometrical shapes



**Fig. 25.** Attribute Relational Graph: (a) circles and straight lines in scene image with magnification of a portion showing three straight lines and a circle, (b) centers of all straight lines and circles, (c) graph for the two straight lines and circle, and (d) attributes of nodes and edges

Image retrieval applications typically employ histogram (or probability density) distance measures. Bin-by-bin distance measures such as Euclidean distance (or its generalization known as the Minkowski distance) and Kullback-Leibler divergence are perceptually unsatisfactory. Earth Mover’s Distance (EMD), a cross bin distance metric is popular in content-based image retrieval [53]. Advantages of EMD are that it allows partial matches, ability to efficiently handle

high-dimensional feature spaces and closeness to perceptual similarity when applied to image histograms.

**Earth Mover's Distance.** EMD evaluates the least amount of work that is needed to transform one distribution into the other. Consider the evaluation of the distance between two signatures (histograms)  $P_1 = \{P_{1i} | 1 \leq i \leq n_1\}$  and  $P_2 = \{P_{2j} | 1 \leq j \leq n_2\}$ . The bins  $[P_{1i}]$  have corresponding weights  $\mathbf{w}_1 = [w_{1i}]$  and similarly  $[P_{2j}]$  have weights  $\mathbf{w}_2 = [w_{2j}]$ . The ground distance matrix  $\mathbf{C} = [c_{ij}]$  specifies *ground distance* between all pairs of bins,  $c_{ij}$ . The flow matrix  $\mathbf{F} = [f_{ij}]$ , where  $f_{ij}$  is the amount of "supplies" transferred from bin  $P_{1i}$  to bin  $P_{2j}$ . The goal is to find proper values of  $\mathbf{F}$  in order to *minimize* the overall work given by

$$WORK(\mathbf{w}_1, \mathbf{w}_2, \mathbf{C}) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} f_{ij} \quad (7)$$

which is subject to the following constraints:

$$f_{ij} \geq 0, \quad \forall 1 \leq i \leq n_1, 1 \leq j \leq n_2, \quad (8)$$

$$\sum_{j=1}^{n_2} f_{ij} \leq w_{1i}, \quad \forall 1 \leq i \leq n_1, \quad (9)$$

$$\sum_{i=1}^{n_1} f_{ij} \leq w_{2j}, \quad \forall 1 \leq j \leq n_2, \quad (10)$$

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij} = \min\left(\sum_{i=1}^{n_1} w_{1i}, \sum_{j=1}^{n_2} w_{2j}\right). \quad (11)$$

Constraint 8 allows moving "supplies" from  $P_1$  to  $P_2$  and not vice versa. Constraint 9 limits the amount of "supplies" that can be sent by the bins in  $P_1$  to their weights. Constraint 10 limits the bins in  $P_2$  to receive no more "supplies" than their weights. Constraint 11 forces to move the maximum amount of "supplies" possible. This amount is referred to as the total flow in the transportation problem.

This is a linear programming problem which is solved efficiently by the transportation simplex algorithm [54]. Once the flow matrix  $\mathbf{F}$  is found, the Earth Mover's Distance is defined as the overall work normalized by the total flow

$$EMD(P_1, P_2) = \frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} c_{ij} f_{ij}}{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f_{ij}}. \quad (12)$$

The computation of EMD assumes that there exists a proper distance measure to compute ground distance matrix  $\mathbf{C}$ , where the element  $c_{ij}$  is the unit distance between a pair of bins  $P_{1i}$  and  $P_{2j}$ , i.e. the work required to move one unit of "supplies" from the source bin  $P_{1i}$  to the destination bin  $P_{2j}$ . It is straightforward to define such a distance between histogram bins because of their strict relative order.



**Modification of EMD for Footwear Outsole Patterns.** Robust ARG matching requires an assignment algorithm that yields not only a correspondence between two sets of vertices but also the similarity between them. In EMD, the bins are replaced by vertices and relations between them. Both vertices (nodes) and relations (edges) have attributes associated with them. The vertices also have associated weights with them, which are useful in performing assignment. However, when matching two ARGs, the ground distance between two vertices depends not only on the two vertices themselves, but also is related to their incident edges. Therefore, computing the ground distance between two vertices, involves a combinatorial optimization procedure to establish correspondence as consistently as possible between the *attributed trees* rooted at vertices. Hence, *direct application of the basic EMD algorithm cannot solve the ARG matching problem* and it needs to be augmented with a method for computing the ground distance matrix between all pairs of nodes.

Nested structure of EMD has been used to achieve robust ARG matching in computer vision [55]. However, it does *not* work well when two graphs to be matched have multiple attributes of different scales, and the difference in each attribute between two ARGs contribute *unequally* to the resulting overall distance. In this case, we need to apply appropriate weights on different attributes to balance their contributions to the overall distance, so that the difference in one feature/attribute will not dominate the overall distance. This step is essential as crime scene marks are created in an uncontrolled environment and they are highly degraded and partial, too. The weights for different attributes can be learnt using sensitivity analysis. First, we elaborate how learned weights are incorporated into EMD, followed by how to learn the weight vector.

A completely connected ARG is formally defined as  $P = (V, R, n)$  where  $V = \{V_i | 1 \leq i \leq n\}$  is the set of nodes and  $R = \{R_{ij} | 1 \leq i, j \leq n\}$  is the set of relations between nodes. Each node has a weight and an attribute vector,  $V_i = (w_i, \mathbf{v}_i)$  and each relation  $R_{ij}$  has an attribute vector  $\mathbf{r}_{ij}$ . Let ARG of 1<sup>st</sup> and 2<sup>nd</sup> footwear prints be  $FP_1 = (V_1, R_1, n_1)$  and  $FP_2 = (V_2, R_2, n_2)$  respectively. To compute the Footwear print distance (FPD) between  $FP_1$  and  $FP_2$ , an appropriate mapping  $M$  between the two sets of nodes is needed. The cost or ground distance matrix is  $\mathbf{C} = [c_{ij}]$  where  $c_{ij} = c(V_{1i}, V_{2j} | V_{1i} \in V_1, V_{2j} \in V_2)$ . The unit cost or distance between  $V_{1i}$  and  $V_{2j}$  is evaluated based on the similarity of the spatial configurations at the two nodes, which is explained later in this section.

By providing *identical* weights for all nodes the nested structure of EMD can handle the case of subgraph matching, i.e.,

$$w_{1i} = w_{2j} = \frac{1}{\max(n_1, n_2)}, 1 \leq i \leq n_1, 1 \leq j \leq n_2. \quad (13)$$

Unlike EMD, a node of  $FP_1$  can transfer its weight to only one node of  $FP_2$ . This is known as *uniqueness constraint*. To enforce one-to-one correspondence, each node  $i$  in the first ARG can match only one node  $j$  in the second ARG

or left unmatched, i.e.  $f_{ij}$  may take the value of either  $\frac{1}{\max(n_1, n_2)}$  or 0,  $\forall i \in \{1, \dots, n_1\}, j \in \{1, \dots, n_2\}$ . Therefore, we rewrite Eq. 12 as

$$\text{FPD}(FP_1, FP_2) = \frac{\frac{1}{\max(n_1, n_2)} \sum_{\{(i,j)|f_{ij}>0\}} c_{ij}}{\sum_{\{(i,j)|f_{ij}>0\}} f_{ij}} \quad (14)$$

The total number of correspondence pairs between the two ARGs is  $\min(n_1, n_2)$  so the total amount of flow transferred from  $FP_1$  to  $FP_2$  is  $\frac{\min(n_1, n_2)}{\max(n_1, n_2)}$ . Substituting this term for the denominator in Eq. 14 we get,

$$\text{FPD}(FP_1, FP_2) = \frac{\sum_{\{(i,j)|f_{ij}>0\}} c_{ij}}{\min(n_1, n_2)} \quad (15)$$

*Cost Determination between Two Nodes.* For a given pair of nodes in two graphs, say  $V_{1i}$  and  $V_{2j}$ , how one node is different from the other depends not only on the nodes, but also on how they relate to their respective neighbors in terms of distance, orientation, position etc. This means that the distance  $c_{ij}$  between the two nodes should be evaluated based on the distance between an attributed relational sub-graph rooted at  $V_{1i}$  and attributed relational sub-graph rooted at  $V_{2j}$ . Each attributed relational sub-graph is an *Attributed Tree* (AT) [56]. ARG & Attributed tree for two sample prints are shown in Fig. 26. This leads to a *nested structure of ARG matching*, which consists of inner and outer steps. For the outer step, the unit cost or distance between  $V_{1i}$  and  $V_{2j}$ , is defined as

$$c(V_{1i}, V_{2j}) = \text{EMD}(AT_{V_{1i}}, AT_{V_{2j}}), \quad (16)$$

where  $AT_{V_{1i}}$  and  $AT_{V_{2j}}$  are attributed trees rooted at  $V_{1i}$  and  $V_{2j}$  in the two ARGs. The tree  $AT_{V_{1i}}$  consists of the root vertex  $V_{1i}$  and its connection to the rest of the  $n_1 - 1$  vertices.

To calculate the distance between the two trees  $AT_{V_{1i}}$  and  $AT_{V_{2j}}$  using EMD framework, we build the inner cost matrix  $\hat{C} = [c_{ij}^{\hat{}}]$  whose elements correspond to pairwise node-to-node ( $V_{1\hat{i}}$  to  $V_{2\hat{j}}$ ) distances *in the two trees*. The *inner cost* between  $V_{1\hat{i}}$  and  $V_{2\hat{j}}$  takes into account not only the unary attributes of the nodes but also their edges attributes and is calculated by

$$c(V_{1\hat{i}}, V_{2\hat{j}}) = \alpha d_E(\mathbf{v}_{1\hat{i}}, \mathbf{v}_{2\hat{j}}) + (1 - \alpha) d_E(\mathbf{Q} * \mathbf{r}_{1\hat{i}\hat{i}}, \mathbf{Q} * \mathbf{r}_{2\hat{j}\hat{j}}) \quad (17)$$

where  $\alpha$  is a weight co-efficient in the interval  $[0, 1]$ ,  $d_E$  is the Euclidean distance,  $\mathbf{r}_{1\hat{i}\hat{i}}$  is the attribute vector of the edge between  $V_{1i}$  and  $V_{1\hat{i}}$ ,  $\mathbf{Q}$  is the weight vector and the operator ‘\*’ denotes the element-wise product between two vectors. Parameter  $\alpha$  reflects the relative importance of the difference of node attributes and the difference of edge attributes in the evaluation of inner cost between two nodes, and is set to 0.5 assuming *equal importance*. Weight vector  $\mathbf{Q}$  for all edge attributes is derived using sensitivity analysis.

Nodes  $V_{1i}$  and  $V_{2j}$  may have one of three possible labels: ‘L’, ‘C’ and ‘E’ corresponding to lines, circles, or ellipses respectively. Thus there are 9 combinations of labels for  $(V_{1i}, V_{2j})$ . A line cannot match with a circle or an ellipse regardless of their attributes and neighbors; while a circle and ellipse can match to

some degree. Thus the unit matching cost for *non-matching* label pairs is  $c('L', 'C') = c('L', 'E') = 1$ . For other label pairs, the node-to-node inner costs are determined using Eq. 17.

*Computing the Weight Vector Using Sensitivity Analysis.* The distance between ARGs has different sensitivities for different attributes. The weight vector  $\mathbf{Q}$  in Eq. 17 takes care of the differences in sensitivities. For large  $n_1$ ,  $\frac{2n_1}{(2n_1-1)} \approx 1$ , thus we have  $Q_k \approx \frac{1}{\sqrt{m}}$ . When  $n_1 = 2$ ,  $Q_k = \frac{4}{3\sqrt{m}}$ . This indicates that we can determine the weights  $\{Q_k, k = \{1, \dots, m\}\}$  by first deriving the value of  $Q_k$  in the case of 2-nodes, then multiplying it by  $\frac{3}{4}$ . The contribution of *each* edge attribute for all pairs of nodes to distance can be calculated as  $\frac{\frac{2n_1}{(2n_1-1)*\sqrt{n}} * 1 * n_1 (n_1-1) * \alpha}{n_1^2} = \frac{n_1-1}{(2n_1-1)\sqrt{n}}$ .

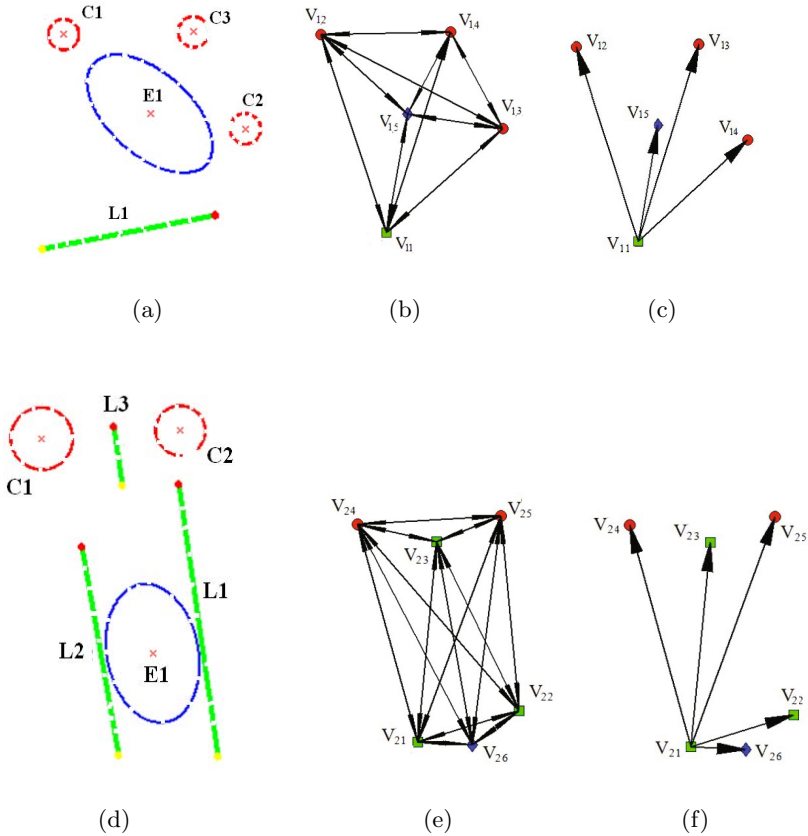
**Examples.** An example of distance computation with two simple prints and their graphs is shown in Fig. 26. Print  $P_1$  has five imperfect elements: three circles, an ellipse and a straight line, its ARG has five nodes  $\{V_{11}, ..V_{15}\}$ . Print  $P_2$  has six imperfect elements: two circles, one ellipse and three straight line segments, its ARG has six nodes  $\{V_{21}, ..V_{26}\}$ . Thus the number of edges in their ARGs are  $2 \times \binom{5}{2} = 20$  and  $2 \times \binom{6}{2} = 30$  respectively.

The process of similarity computation in a more realistic scenario involving actual footwear prints is shown in Figure 27. In this case the distance evaluates to a much smaller value of 0.0835 indicating a finer degree of match.

Sensitivity analysis [57] is a system validation technique which can be used to determine robustness of the distance measure when the inputs are slightly disturbed. Its application here is to determine as to how sensitive the distance measure is to changes with respect to attributes. Plots of distance with respect to each of the attributes is obtained. A linear change is consistent with human perception whereas nonlinear behavior needs justification for its suitability. This analysis showed linear correlation with most attributes.

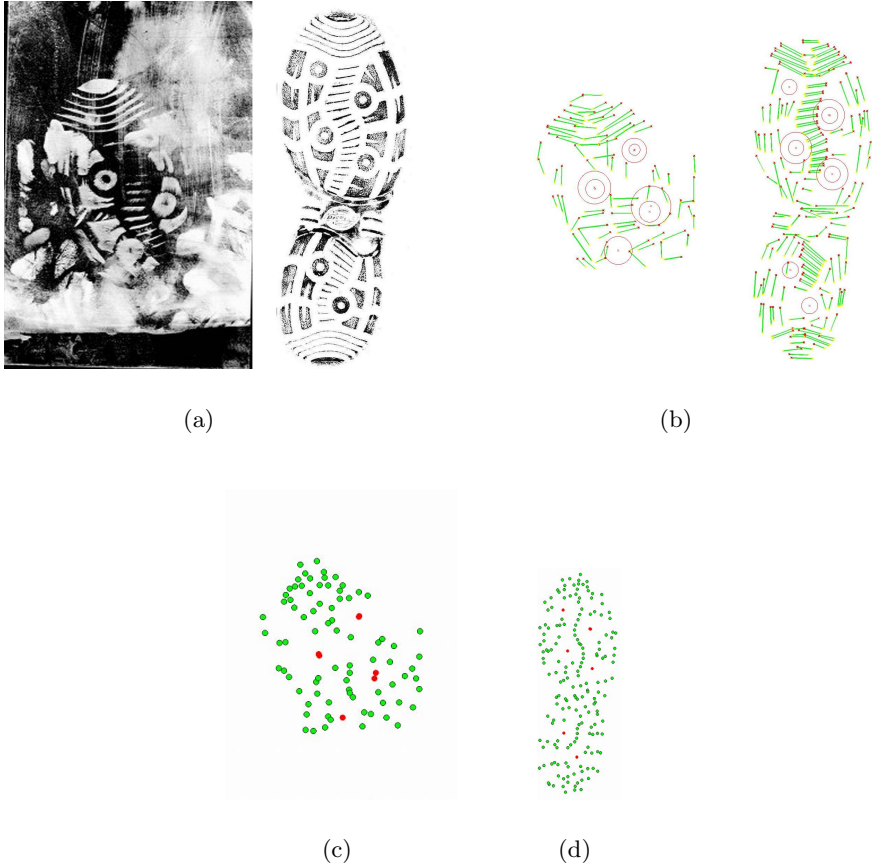
## 4.7 Search Algorithms

A functional block-diagram of an end-to-end system to retrieve closest matches to a query crime scene image in a database of reference images is shown in Figure 28. Image enhancement operations such as edge detection or contextually based image pixel labeling are performed on both the input and the known images. Next, a feature representation is constructed for the image either by extracting them from the entire image or by detecting local patterns in outsoles. The design should attempt to integrate several levels of analysis: (i) global shoe properties: heavily worn or brand new, shape, size etc., (ii) detailed and distinctive local features should be utilized to increase the discriminative power in order to confirm a match. Each level requires a different variety of image analysis techniques from robust geometric and texture feature detectors to detailed correlation of distinctive minutiae and their spatial arrangement.



**Fig. 26.** Distance computation between two simple prints: (a) print  $P_1$  with five primitive elements, (b) attributed relational graph of  $P_1$  with vertices  $V_{11}..V_{15}$ , (c) attributed tree rooted at  $V_{11}$ , (d) print  $P_2$  with six elements, (e) attributed relational graph of  $P_2$  with vertices  $V_{21}..V_{26}$  and (f) attributed tree rooted at  $V_{21}$ . Nodes represented by squares, circles and diamonds represent lines, circles and ellipses respectively. Using attributes of nodes and edges as defined in Figure 24 the distance evaluates to 0.5674.

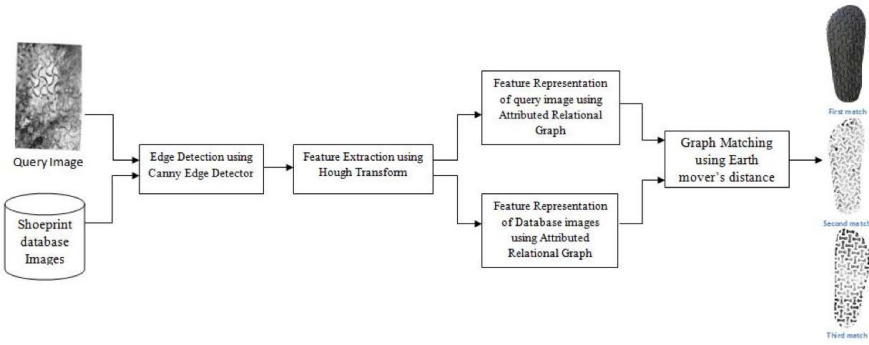
A similarity measure appropriate to the feature description is used in the comparison of two images. In the design shown a graph representation of the characteristic features is used— where each node denotes a single geometrical primitive, such as a circle, an ellipse, a line segment, with attributes describing unary features of this primitive; each attributed edge between a pair of nodes represents spatial relationships between them. Thus the problem of image retrieval and matching is converted to an attributed graph matching problem. It involves establishing correspondence between the nodes of the two graphs. Retrieving the most similar prints to an impression can be made faster by clustering the database prints beforehand.



**Fig. 27.** Example of similarity computation between a crime scene image and a known outsole pattern: (a) input image and pattern, (b) geometric primitives detected in both, and (c,d) corresponding ARGs, where only nodes are shown for clarity. The distance between the two ARGs is 0.0835.

**Reference Pattern Clustering.** The computational complexity of distance computation for two ARGs with  $n_1$  and  $n_2$  nodes is  $O(n_1^2 n_2^2 \max(n_1, n_2))$ . Since the computation is intensive, it is necessary to use approximate methods to speed-up the retrieval process. One approach is to eliminate several edge evaluations. Another is to cluster the reference images so that not all comparisons need to be made.

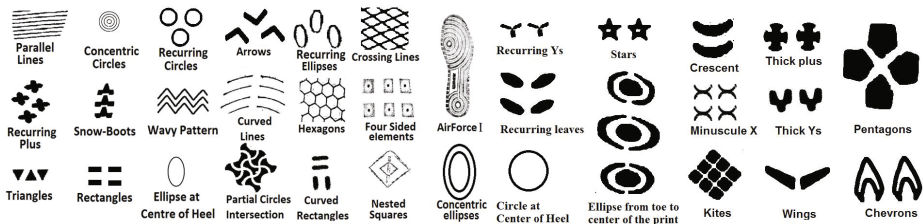
Clustering algorithms can be generally divided into partition-based, density-based and hierarchical based methods [58]. Algorithms like  $k$ -means, hierarchical clustering, and expectation maximization requires similarity matrix consisting of pair-wise distance between every footwear prints in dataset. Building similarity matrix is computationally expensive for a large dataset. Further, the ARG



**Fig. 28.** Functional architecture for matching a query image with a database of soleprint reference patterns

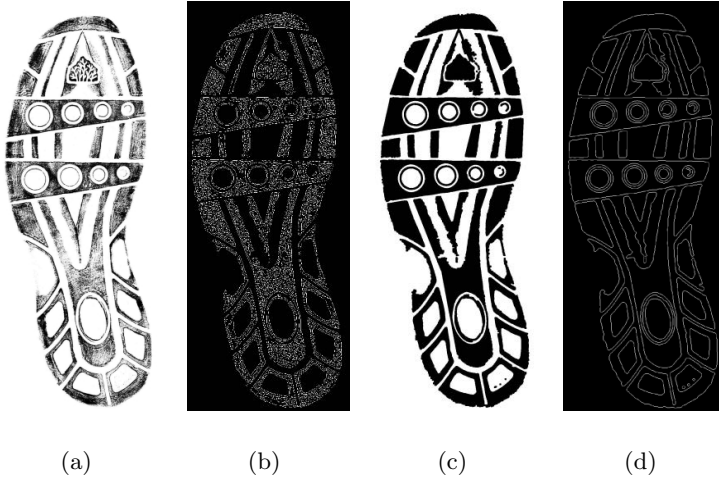
representing a footwear print has 200-300 nodes on average and nodes can vary considerably in terms of relative size, position etc. This makes the feature space very *sparse* and therefore similar footwear prints tend to stay close to each other and dissimilar ones stay apart. Hence, to cluster the entire dataset we use *recurring patterns* as *fixed* cluster centers [18].

Footwear outsoles typically contain recurring patterns such as waves and concentric circles [5,59]. Each such pattern can represent a group of similar patterns. Each pattern is simple and its graph structure has a small number of nodes. Further, the ARG representing a footwear print has 200-300 nodes on average and nodes can vary considerably in terms of relative size, position etc. This makes the feature space very *sparse* and therefore similar footwear prints tend to stay close to each other and dissimilar ones stay apart. Hence, to cluster a huge dataset *recurring patterns* can be used as cluster representatives, which serve as *initial seed clusters* [60]. From visual inspection of 2,660 prints, 33 recurring patterns were determined and used as cluster representatives (see Figure 29). For each reference image, its distance to each pattern is computed and then assigned it to the nearest cluster representative. These cluster representatives are similar to cluster means in *k*-means algorithm but these "means" are fixed. Efficiency is achieved by exploiting sparseness of the feature space.



**Fig. 29.** Recurring patterns in outsole prints that are used as canonical cluster centers

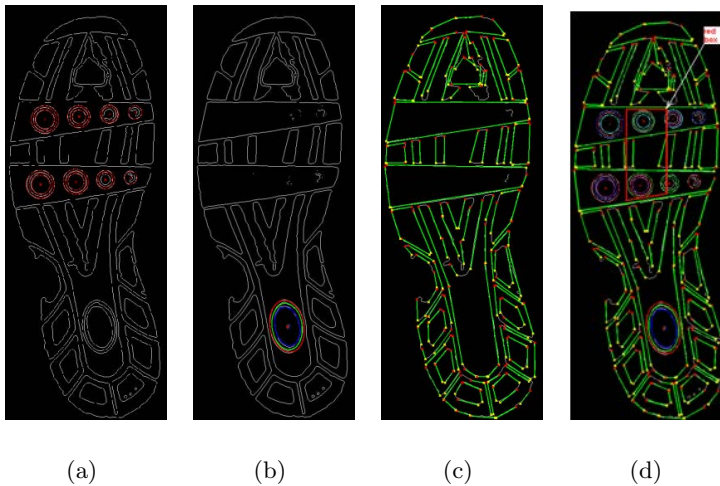
*Clustering Step 1.* The first step in feature extraction is to perform morphological operations such as dilation and erosion (Figure 30). This makes the interior region of the boundary *uniform* and hence the edge detector [61] does not detect any edges inside the boundary. This helps to enhance the quality of the edge image.



**Fig. 30.** Illustration of step 1 of clustering, where morphological operations are performed on reference patterns: (a) an input grey-scale image, (b) edge image of (a), (c) result of morphological operation on (a), (d) edge image of (c)

*Clustering Step 2.* The simple Hough transform (SHT) is used to detect circles in footwear prints. Pixels of detected circles are removed from the edge image and fed as input for ellipse detection using the randomized Hough transform (RHT). Pixels of detected ellipses are removed from the edge image and the output is fed as input for line detection. Features are extracted in the order: circle, ellipse and line (Figure 31). This is because circles are degenerated ellipses and arbitrary shapes in footwear print are approximated by piecewise lines.

*Clustering Step 3.* For each detected feature, node attributes of completeness and quality of circle, eccentricity of the ellipse etc. are computed. Further, edge attributes like relative distance and position between nodes are calculated and finally an ARG is constructed (Figure 32). The distance between each reference print and every cluster representative is calculated. Then each print is assigned to the nearest representative, for which the distance is below threshold  $T$ . If distance between a print and cluster representatives are greater than  $T$ , then the print remains as a single cluster.

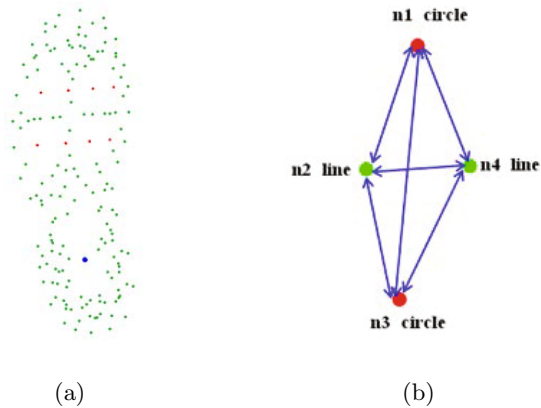


**Fig. 31.** Illustration of step 2 of clustering where the Hough transform is used to extract features. The sequence of operations is circle→ellipse→line. Detected features are: (a) circles. (b) ellipses. (c) line segments. (d) all features. Red box indicates a small region in the footwear print.

*Clustering Performance.* With  $T = 0.15$ , and 1,000 outsole patterns, the clustering algorithm assigned 550 patterns to one of 20 clusters whereas the remaining 450 were unique enough to be singleton clusters. Sample clusters based on the canonical patterns of Fig. 29 are shown in Fig. 33. Clustering accuracy is measured by the  $F$ -measure of retrieval, which is the weighted harmonic mean of precision and recall (Figure 34(a)). An advantage of using cluster centers is significant reduction in computation. For a database of 1000 prints, there are 499,500 pairwise distances. With clustering based on  $k$  recurring patterns as seed,  $1000 \times k$  distance computations are needed; with  $k = 20$ , computation is reduced by 96%. This efficiency is achieved without compromising the accuracy or recall rate.

**Retrieval Performance.** Evaluation metrics for retrieval performance are the *cumulative match characteristic (CMC)* and speed. The CMC answers the question [11] “what is the probability of finding a match in the first  $n$  percent of database images?” The *cumulative match score* is the proportion of times when the correct reference print is in the first  $n$  percent of the sorted database. This metric can be used even when there is a single match in the database. For a dataset of 50 crime scene prints, used as query, and 1066 reference patterns, containing meta data such as brand and model, the CMC curve *before* clustering is shown in Fig.34(b). The CMC curve after clustering, where the query is matched against the cluster representatives to find the closest cluster and then






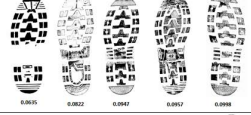






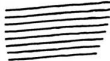



**Fig. 32.** Illustration of step 3 of clustering where a graph is constructed: (a) nodes in ARG of footwear pattern in Figure 30(a) with edges omitted due to complete connectivity, (b) subgraph for region enclosed within red box of Figure 31(d)

against each pattern in the cluster to retrieve the top  $n$  matches, does not show significant degradation. From the CMC curve, the top 0.1% of database patterns will contain the correct match with probability of 0.43. Tests with crime scene marks have an error of 0.08% – the confidence interval for the sample size is the interval [0.03%, 0.18%].

The CMC of ARG-EMD is much better than that of the SIFT feature descriptor [43] also shown in Figure 34(b). SIFT, which is commonly used in image retrieval including Google's similar image search, performs only slightly better than randomly selecting a reference pattern. While SIFT features are not preserved among different outsoles of the same class and through wear lifetime, ARG-EMD extracts durable geometric features (of lines, ellipses, and their relationships) and demonstrates invariance to scale and rotation. ARG-EMD has additional desirable properties: allows partial matching in a natural way, is robust to the change of the relational structure, and consistent with perceptual similarity (as can be seen in the two examples of Figure 35).

*Speed.* The time for processing a scene and reference image depends on the number of nodes in each graph. If the average time to compute one distance is 30 seconds then for a single query and 1,000 database entries, it takes 20-30 minutes. In a large reference database, the efficiency(speed) of retrieving a query print becomes important. Effective indexing techniques should be designed to enter standard reference patterns. Speed can be improved by: (i) reducing the number of nodes by merging two detected lines which are associated with a single straight boundary (to be done), (ii) using pre-filtering to enhance the speed performance e.g., computing the Euclidean distance between global feature vectors of the query print and each database print, and ignoring those database prints too far

Cluster Name	Cluster Representative	Total number of prints in the cluster	Sample prints
Wavy Pattern		145	
Snow Boots Pattern		5	
Concentric Circles		20	
Recurring Plus		6	
Partial Circles Interleaves		4	
Parallel Lines		25	

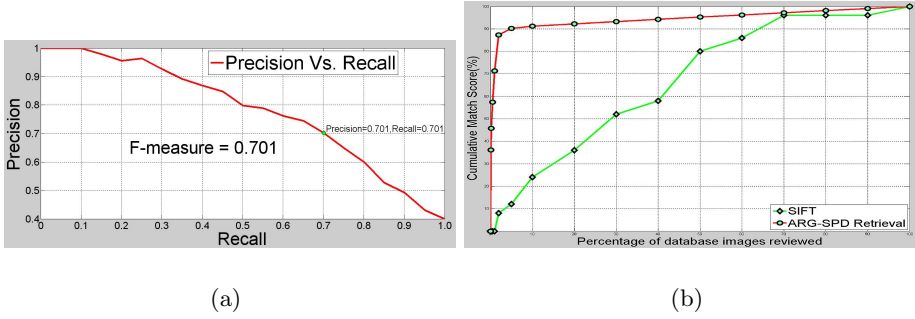
**Fig. 33.** Sample clusters of reference patterns based on using the canonical patterns in Fig. 29

from the query to be a potential match, (iii) relaxing full connectivity in graph by triangulation, and (iv) other improvements. In terms of performance, with 50 scene images the average time was 120 minutes before clustering and 8 minutes after with no significant retrieval degradation.

#### 4.8 Quantifying Uncertainty of Match

In reporting the results of a comparison between the evidence and known an expression of the uncertainty involved is useful. This opinion can be expressed in probabilistic terms using statistical methods for computing the strength of evidence [62]. A rule for converting likelihood ratios into scales has also been suggested [63].

For evidence interpretation, three different approaches have been stated: “Classical”, “Likelihood Ratio” and “Full Bayes’ Rule”. The likelihood ratio approach [64] is widely accepted among various forensic investigations as it provides a transparent, consistent and logical framework to discriminate among competing



**Fig. 34.** Retrieval performance: (a) precision-recall curve for circles only, and (b) cumulative match characteristic, which gives the probability of correct match in the top  $n\%$  of ranked database, of ARG-EMD (based on circles, ellipses and straight lines) compared with that of SIFT

hypotheses. In the Full Bayes’ Rule approach, the posterior probability of a set of hypotheses given the existing evidence is determined. Although this method has been a very common practice of forensic document examiners in central European countries, it has been said that there is no creditable justifications for its validity and appropriateness[65].

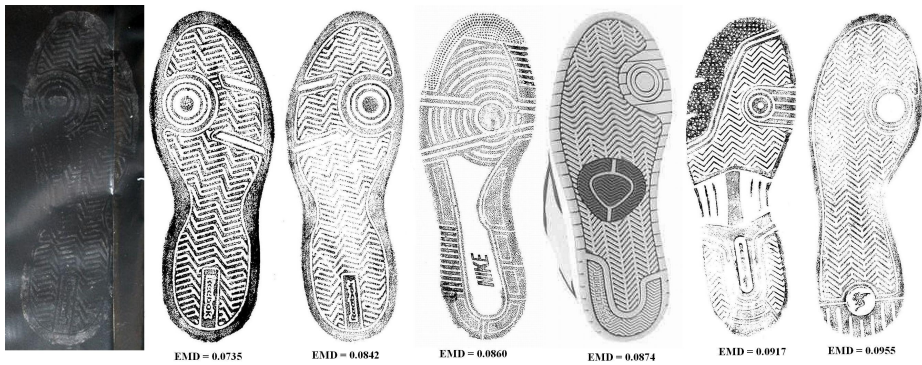
The likelihood ratio is the ratio of the two probabilities of the evidence given two competing hypotheses:  $h^0$  – the crime scene print is created by the same footwear as the known print and  $h^1$  – the crime scene print is not from the known. This ratio can be expressed as:  $LR = \frac{P(E|h^0,I)}{P(E|h^1,I)}$ . where  $E$  is the evidence given by the crime scene mark, and  $I$  is all the background information relevant to this case. This approach can be decomposed into the following three steps: (i) estimate the within-class and between-class shoe-print variability, (ii) compute the LR for the evidence, and (iii) convert the LR into a verbal scale.

**Degradation Model.** In order to obtain a probabilistic measure it is necessary to characterize within-class and between-class variabilities. *Within-class variability* measures the variance of features of multiple prints from the same outsole. To be able to simulate different variations caused by wears, incompleteness and the change of medium and illumination, we can apply image degradation models multiple times on each database image to produce a set of degraded outsole prints.

**Approximating the Likelihood Ratio.** Direct computation of the likelihood ratio is infeasible due to the large number of possible variations of the same and different distributions. Since uncertainty is a function of similarity of the characteristics (both class and individualizing) as well as the rarity of the characteristics [66] the following methods based on (i) distance and (ii) distance and rarity can be used.



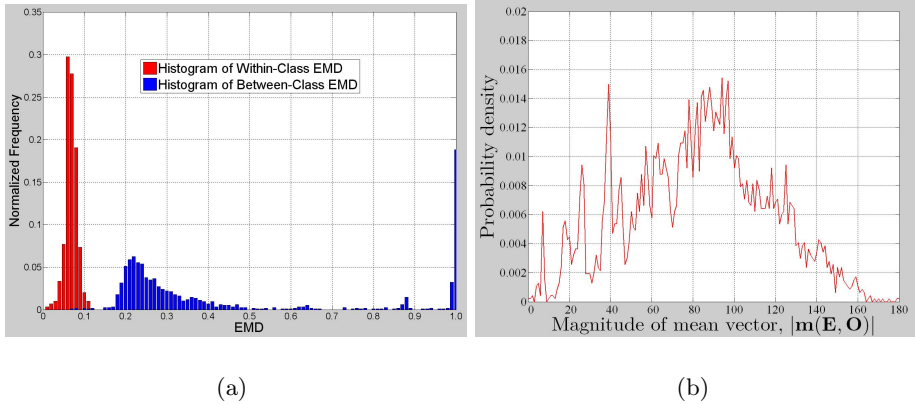
(a)



(b)

**Fig. 35.** Results of automatic retrieval with two queries shown as the left-most images followed on the right by the top database entries retrieved. It can be seen that the top choices are similar to human perception.

*Distance Method.* A matching algorithm can be applied to calculate the distance between each pair of within-class prints. It is then possible to build a probability distribution of within-class distance. *Between-class variability* measures the variance of features of multiple prints that are from different classes. In a similar way the within-class variability can be modeled. Given a distance between the crime scene mark and a test mark made by the suspect's shoe, we can compute the likelihood of the observed distance  $d$  given the hypothesis that the two marks are from the same source, as well as the likelihood of the distance given the hypothesis that the two marks are from different sources. The ratio of these two likelihoods is then calculated to get  $LR_D = \frac{P(d|h^0)}{P(d|h^1)}$ .



**Fig. 36.** Likelihood Ratio methods: (a) *distance method* is based on histograms of intra- and inter-class distance, and (b) *distance and rarity method* uses distribution of magnitude of mean vector  $\mathbf{m}(\mathbf{E}, \mathbf{O})$  (built with 5,289 pairs of samples)

Using footwear impressions together with ground truth, histograms for  $P(d|h^i)$  ( $i = 0, 1$ ) are built as shown in Fig. 36; in this example 1,060 degraded footwear prints with ground truth were used. Modeling  $P(d|h^0)$  by a Gaussian  $\mathcal{N}(d|\mu_0, \sigma_0^2)$ , and  $P(d|h^1)$  by a mixture of Gaussians the likelihood ratio is computed. The distribution of LR, determined from a learning set, can be used to convert the LR value into an opinion scale.

*Distance and Rarity Method.* The distance method provides a severe approximation to the true likelihood ratio by going from a high-dimensional feature space to a one-dimensional distance space. A better approach, as shown in [66], is to estimate *LR* as the product of two factors, one based on *difference* and the other on *rarity*:

$$LR_{DR} = P(\mathbf{d}(\mathbf{o}, \mathbf{e})|h^0) * \frac{1}{P(\mathbf{m}(\mathbf{o}, \mathbf{e}))}, \tag{18}$$

where  $\mathbf{d}(\mathbf{o}, \mathbf{e})$  is the *difference* between object vector  $\mathbf{o}$  and evidence vector  $\mathbf{e}$ , and  $\mathbf{m}(\mathbf{o}, \mathbf{e})$  is the *mean* of  $\mathbf{o}$  and  $\mathbf{e}$ .

When features are extended from vectors to graphs associated with feature sets  $\mathbf{E}$  and  $\mathbf{O}$ , correspondence between them is not apparent, and the number of corresponding elements (features) in  $\mathbf{E}$  and  $\mathbf{O}$  may be different. Instead of defining a distribution on the set difference  $\mathbf{E} - \mathbf{O}$  we can use the distribution of the distance  $d(\mathbf{E}, \mathbf{O})$ . Next we discuss an approximation to the rarity term.

In computing the distance between two ARGs, we determine corresponding nodes between them, This *induces* two sub-graphs in the two original ARGs that have an equal number of nodes and edges. We construct a feature vector from each of the two sub-graphs. The scalar distance (if computed by Euclidean distance) is the magnitude of the vector difference. By analogy, we can use

the distribution of the *magnitude* of the mean vector as a substitute for the distribution of the mean vector itself, i.e.

$$LR_{DR} = \frac{P(\mathbf{d}(\mathbf{E}, \mathbf{O})|h^0)}{P(\mathbf{m}(\mathbf{E}, \mathbf{O}))} \approx \frac{P(d(\mathbf{E}, \mathbf{O})|h^0)}{P(|\mathbf{m}(\mathbf{E}, \mathbf{O})|)}. \quad (19)$$

This approximation has intuitive appeal: two graphs with more matched features will have a greater value of  $|\mathbf{m}(\mathbf{E}, \mathbf{O})|$  than with fewer matched features. By mapping the distribution of mean vector  $\mathbf{m}(\mathbf{E}, \mathbf{O})$  of *varied* length to the distribution of its magnitude, we have overcome the difficulty of defining the distribution of the difference  $\mathbf{E} - \mathbf{O}$ , avoided normalization and made both numerator and denominator have the same dimension. This mapping does give a reasonable approximation of the original rarity. In experiments the FPD was computed between all pairs of prints in the training data set (1,060 prints) yielding an average error rate of 4.5% with the distance method and 2.5% with the distance and rarity method.

## 5 Summary and Conclusions

While footwear impressions are commonly found in crime scenes, they are not often used in either the investigative or prosecutorial phases of criminal justice due to many practical difficulties. Reliable automated tools should enable more use of footwear impression evidence. A review of methods of footwear print examination reveals the need for computational solutions for several tasks: enhancing the quality of crime scene images, representing outsole patterns so as to be useful in comparison, evaluating similarity between evidence and known, implementation of algorithms to retrieve closest matches in a reference database, performance evaluation metrics and quantifying uncertainty of opinion. Data sets useful in developing methods are: (i) simulated prints (crime scene prints obtained by stepping on talcum powder and then on carpet, and known prints by stepping on chemically treated paper), (ii) photographs of outsoles retrieved by a web crawler from shoe-vendor websites, and (iii) actual crime scene prints and corresponding known prints. Since results with simulated images tend to be over-optimistic results should focus on real crime scene prints.

For extracting foreground pixels from crime scene images, a method based on utilizing statistical dependencies between nearby pixels (one based on CRFs) is better than thresholding methods. For representing the geometrical patterns commonly found in outsole prints, a structural method performs better than simple two-dimensional (GSC) and three-dimensional (SIFT) representations. The structural method is based on detecting component geometric shapes, principally ellipses of different eccentricities. The relationships between these elements in the print is then modeled as a graph whose nodes represent primitive elements (together with defining attributes relating to parameters such as radius as well as quality in the image) and whose edges represent spatial relationships (also attributed with a list of characteristics). Given two patterns represented as graphs, their similarity is determined by using a graph distance measure, one

related to measuring histogram distance and the Wasserstein metric. It characterizes similarity by a number ranging from 0 to 1.

The retrieval task is to find the closest match to a crime scene print in a local/national database so as to determine footwear brand and model. This process is made faster if database prints are grouped into clusters of similar patterns. For this an ARG is constructed for each known print, where each node is a primitive feature and each edge represents a spatial relationship between nodes. The distance between ARGs is used as similarity measure. This distance is computed between each known print and a pre-determined set of canonical patterns to form clusters. By clustering known images into cognitively similar patterns, higher efficiency is achieved in retrieval. The following topics of further research can be identified:

1. Statistical machine learning approaches can be used effectively in several phases such as enhancement of the crime scene image similarity computation, and drawing a conclusion,
2. A standardized database of crime scene marks would allow researchers to develop and benchmark the performance of their algorithms and systems.
3. Robustness and sensitivity of the similarity measures needs to be further studied, e.g., different sizes of the query image, increased number of footwear models, etc.
4. The use of the similarity metrics in the computation of likelihoods using both class characterizing and individualizing features need to be studied so as to provide uncertainty measures in comparison.

## References

1. Bodziak, W.: Footwear Impression Evidence Detection, Recovery and Examination, 2nd edn. CRC Press (2000)
2. Stone, R.S.: Footwear examinations: Mathematical probabilities of theoretical individual characteristics. *Journal of Forensic Identification* 56, 577–599 (2006)
3. Geradts, Z., Keijzer, J.: The image-database REBEZO for shoeprints with developments on automatic classification of shoe outsole designs. *Forensic Science International* 82, 21–31 (1996)
4. Alexander, A., Bouridane, A., Crookes, D.: Automatic classification and recognition of shoeprints. In: Proc. Seventh International Conference Image Processing and Its Applications, vol. 2, pp. 638–641 (1999)
5. Girod, A.: Computerized classification of the shoeprints of burglar's shoes. *Forensic Science International* 1, 59–65 (1982)
6. Bouridane, A., Alexander, A., Nibouche, M., Crookes, D.: Application of fractals to the detection and classification of shoeprints. In: Proceedings International Conference Image Processing, vol. 1, pp. 474–477 (2000)
7. Sawyer, N.: SHOE-FIT: A computerised shoe print database. In: Proc. European Convention on Security and Detection (1995)
8. Ashley, W.: What shoe was that? the use of computerised image database to assist in identification. *Forensic Science Int.* 82(1), 7–20 (1996)
9. Foster, Freeman: Solemate (2010), <http://fosterfreeman.com>

10. Bouridane, A.: *Imaging for Forensics and Security: From Theory to Practice*, 1st edn. Springer (2009)
11. de Chazal, P., Flynn, J., Reilly, R.B.: Automated processing of shoeprint images based on the Fourier transform for use in forensic science. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 341–350 (2005)
12. Zhang, L., Allinson, N.: Automatic shoeprint retrieval system for use in forensic investigations. In: *UK Workshop on Computational Intelligence* (2005)
13. Pavlou, M., Allinson, N.M.: Automatic extraction and classification of footwear patterns. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) *IDEAL 2006*. LNCS, vol. 4224, pp. 721–728. Springer, Heidelberg (2006)
14. Crookes, D., Bouridane, A., Su, H., Gueham, M.: Following the footsteps of others: Techniques for automatic shoeprint classification. In: *Second NASA/ESA Conference on Adaptive Hardware and Systems*, pp. 67–74 (2007)
15. Gueham, M., Bouridane, A., Crookes, D.: Automatic classification of partial shoeprints using advanced correlation filters for use in forensic science. In: *International Conference on Pattern Recognition*, pp. 1–4 (2008)
16. Patil, P.M., Kulkarni, J.V.: Rotation and intensity invariant shoeprint matching using gabor transform with application to forensic science. *Pattern Recognition* 42, 1308–1317 (2009)
17. Dardi, F., Cervelli, F., Carrato, S.: A texture based shoe retrieval system for shoe marks of real crime scenes. In: Foggia, P., Sansone, C., Vento, M. (eds.) *ICIAP 2009*. LNCS, vol. 5716, pp. 384–393. Springer, Heidelberg (2009)
18. Tang, Y., Srihari, S.N., Kasiviswanthan, H.: Similarity and clustering of footwear prints. In: *IEEE Symposium on Foundations and Practice of Data Mining (GrC 2010)*. IEEE Computer Society Press (2010)
19. Tang, Y., Srihari, S.N.: Ellipse detection using sampling constraints. In: *Proc. IEEE Int. Conf. Image Proc.*, IEEE Computer Society Press (2011)
20. Mikkonen, S., Astikainen, T.: Database classification system for shoe sole patterns - identification of partial footwear impression found at a scene of crime. *Journal of Forensic Science* 39(5), 1227–1236 (1994)
21. Huynh, C., de Chazal, P., McErlean, D., Reilly, R., Hannigan, T., Fleud, L.: Automatic classification of shoeprints for use in forensic science based on the Fourier transform. In: *Proc. 2003 International Conference Image Processing*, vol. 3, pp. 569–572 (2003)
22. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. on Pattern Analysis and Machine Intel.* 27, 1615–1630 (2005)
23. Ghouti, L., Bouridane, A., Crookes, D.: Classification of shoeprint images using directional filter banks. In: *International Conference on Visual Information Engineering*, pp. 167–173 (2006)
24. Su, H., Crookes, D., Bouridane, A.: Thresholding of noisy shoeprint images based on pixel context. *Pattern Recognition Letters* 28(2), 301–307 (2007)
25. Sun, W., Taniar, D., Torabi, T.: Image mining: A case for clustering shoe prints. *International Journal of Information Technology and Web Engineering* 3, 70–84 (2008)
26. AlGarni, G., Hamiane, M.: A novel technique for automatic shoeprint image retrieval. *Forensic Science International* 181, 10–14 (2008)
27. Xiao, R., Shi, P.-f.: Computerized matching of shoeprints based on sole pattern. In: Srihari, S.N., Franke, K. (eds.) *IWCF 2008*. LNCS, vol. 5158, pp. 96–104. Springer, Heidelberg (2008)



28. Jingl, M.Q., Ho, W.J., Chen, L.H.: A novel method for shoeprints recognition and classification. In: International Conference on Machine Learning and Cybernetics, vol. 5, pp. 2846–2851 (2009)
29. Nibouche, O., Bouridane, A., Gueham, M., Laadjel, M.: Rotation invariant matching of partial shoeprints. In: International Machine Vision and Image Processing Conference, pp. 94–98 (2009)
30. Cervelli, F., Dardi, F., Carrato, S.: Comparison of footwear retrieval systems for synthetic and real shoe marks. In: Proc. Sixth Intl. Symp. Image and Signal Processing and Analysis, Salzburg, Austria, pp. 684–689 (2009)
31. Dardi, F., Cervelli, F., Carrato, S.: A combined approach for footwear retrieval of crime scene shoe marks. In: Proc. ICDP 2009, Third International Conference on Imaging for Crime Detection and Prevention, Paper No. P09, London, UK (2009)
32. Wang, R., Hong, W., Yang, N.: The research on footprint recognition method based on wavelet and fuzzy neural network. In: International Conference on Hybrid Intelligent Systems, pp. 428–432 (2009)
33. Otsu, N.: A threshold selection method from gray level histogram. *IEEE Transaction on Systems, Man and Cybernetics* 9, 62–66 (1979)
34. Ramakrishnan, V., Srihari, S.N.: Extraction of shoeprint patterns from impression evidence using conditional random fields. In: Proceedings of International Conference on Pattern Recognition, Tampa, FL. IEEE Computer Society Press (2008)
35. Koller, D., Friedman, N.: Probabilistic Graphical Models. MIT Press (2009)
36. Shetty, S., Srinivasan, H., Beal, M., Srihari, S.: Segmentation and labeling of documents using conditional random fields. In: Document Recognition and Retrieval XIV, vol. 6500, pp. 65000U–1 (2007)
37. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images. In: Neural Information Processing Systems, NIPS (2003)
38. Wallach, H.: Efficient training of conditional random fields. Master's thesis, University of Edinburgh (2002)
39. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8(6), 679–698 (1986)
40. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using MATLAB, 1st edn. Prentice Hall (2003)
41. Rui, Y., Huang, S., Chang, S.: Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation* 10, 39–62 (1999)
42. Srihari, S.N., Huang, C., Srinivasan, H.: On the discriminability of the handwriting of twins. *Journal of Forensic Sciences* 53(2), 430–446 (2008)
43. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
44. Hough, P.: Machine analysis of bubble chamber pictures. In: International Conference on High Energy Accelerators and Instrumentation, CERN (1959)
45. Srihari, S.N., Govindaraju, V.: Analysis of textual images using the Hough transform. *Machine Vision and Applications* 2, 141–153 (1989)
46. Goulermas, J., Liatsis, P.: Incorporating gradient estimations in a circle-finding probabilistic hough transform. *Pattern Analysis and Applications* 26, 239–250 (1999)
47. Wu, W.Y., Wang, M.J.J.: Elliptical object detection by using its geometric properties. *Pattern Recognition* 26, 1499–1509 (1993)
48. McLaughlin, R.: Randomized Hough transform: better ellipse detection. *IEEE TENCON-Digital Signal Processing Applications* 1, 409–414 (1996)

49. Haralick, R.M., Shapiro, L.G.: *Computer and Robot Vision*. Addison Wesley (1992)
50. Bunke, H., Irriger, C., Neuhaus, M.: Graph matching: Challenges and potential solutions. In: Roli, F., Vitulano, S. (eds.) *ICIAP 2005*. LNCS, vol. 3617, pp. 1–10. Springer, Heidelberg (2005)
51. Sanfeliu, A., Fu, K.S.: A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics* 13, 353–362 (1983)
52. Bunke, H., Messmer, B.T.: Efficient attributed graph matching and its application to image analysis. In: *Proceedings of the 8th International Conference on Image Analysis and Processing*, pp. 45–55 (1995)
53. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40, 99–121 (2000)
54. Hillier, F.S., Liebermann, G.J.: *Introduction to Mathematical Programming*, 2nd edn. McGraw-Hill (1995)
55. Kim, D.H., Yun, I.D., Lee, S.U.: Attributed relational graph matching algorithm based on nested assignment structure. *Pattern Recognition* 43, 914–928 (2010)
56. Pelillo, M., Siddiqi, K., Zucker, S.W.: Many-to-many matching of attributed trees using association graphs and game dynamics. In: Arcelli, C., Cordella, L.P., Sanniti di Baja, G. (eds.) *IWVF 2001*. LNCS, vol. 2059, pp. 583–593. Springer, Heidelberg (2001)
57. Smith, E., Szidarovszky, F., Karnavas, W., Bahill, A.: Sensitivity analysis, a powerful system validation technique. *The Open Cybernetics and Systemics Journal* 2, 39–56 (2008)
58. Aldenderfer, M., Blashfield, R.: *Cluster Analysis*. SAGE (1984)
59. Mikkonen, S., Suominen, V., Heinonen, P.: Use of footwear impressions in crime scene investigations assisted by computerised footwear collection system. *Forensic Science International* 82(1), 67–79 (1996)
60. Basu, S., Banerjee, A., Mooney, R.J.: Semi-supervised clustering by seeding. In: *Proc. of the Nineteenth International Conference on Machine Learning*, pp. 27–34 (2002)
61. Nixon, M., Aguado, A.: *Pattern Extraction and Image Processing*. Elsevier Science (2002)
62. Aitken, C., Taroni, F.: *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley (2004)
63. Evett, I.: Towards a uniform framework for reporting opinions in forensic science casework. *Science and Justice* 38(3), 198–202 (1998)
64. Evett, I., Lambert, J., Buckleton, J.: A Bayesian approach to interpreting footwear marks in forensic casework. *Science and Justice* 38(4), 241–247 (1998)
65. Biedermann, A., Taroni, F.: Inadequacies of posterior probabilities for the assessment of scientific evidence. *Law, Probability and Risk* 4, 89–114 (2005)
66. Tang, Y., Srihari, S.N.: Likelihood ratio estimation in forensic identification using similarity and rarity. *Pattern Recognition* 47(3), 945–958 (2014)

# A New Swarm-Based Framework for Handwritten Authorship Identification in Forensic Document Analysis

Satrya Fajri Pratama, Azah Kamilah Muda, Yun-Huoy Choo, and Noor Azilah Muda

Computational Intelligence and Technologies (CIT) Research Group,  
Center of Advanced Computing and Technologies,  
Faculty of Information and Communication Technology,  
Universiti Teknikal Malaysia Melaka  
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia  
satrya@student.utem.edu.my,  
{azah, huoy, azilah}@utem.edu.my

**Abstract.** Feature selection has become the focus of research area for a long time due to immense consumption of high-dimensional data. Originally, the purpose of feature selection is to select the minimally sized subset of features class distribution which is as close as possible to original class distribution. However in this chapter, feature selection is used to obtain the unique individual significant features which are proven very important in handwriting analysis of Writer Identification domain. Writer Identification is one of the areas in pattern recognition that have created a center of attention by many researchers to work in due to the extensive exchange of paper documents. Its principal point is in forensics and biometric application as such the writing style can be used as bio-metric features for authenticating the identity of a writer. Handwriting style is a personal to individual and it is implicitly represented by unique individual significant features that are hidden in individual's handwriting. These unique features can be used to identify the handwritten authorship accordingly. The use of feature selection as one of the important machine learning task is often disregarded in Writer Identification domain, with only a handful of studies implemented feature selection phase. The key concern in Writer Identification is in acquiring the features reflecting the author of handwriting. Thus, it is an open question whether the extracted features are optimal or near-optimal to identify the author. Therefore, feature extraction and selection of the unique individual significant features are very important in order to identify the writer, moreover to improve the classification accuracy. It relates to invarianceness of authorship where invarianceness between features for intra-class (same writer) is lower than inter-class (different writer). Many researches have been done to develop algorithms for extracting good features that can reflect the authorship with good performance. This chapter instead focuses on identifying the unique individual significant features of word shape by using feature selection method prior the identification task. In this chapter, feature selection is explored in order to find the most unique individual significant features which are the unique features of individual's writing. This chapter focuses on the integration of Swarm Optimized and Computationally Inexpensive Floating Selection (SOCIFS) feature selection technique into the proposed hybrid of Writer Identification framework

and feature selection framework, namely Cheap Computational Cost Class-Specific Swarm Sequential Selection ( $C_4S_4$ ). Experiments conducted to proof the validity and feasibility of the proposed framework using dataset from IAM Database by comparing the proposed framework to the existing Writer Identification framework and various feature selection techniques and frameworks yield satisfactory results. The results show the proposed framework produces the best result with 99.35% classification accuracy. The promising outcomes are opening the gate to future explorations in Writer Identification domain specifically and other domains generally.

**Keywords:** swarm-based framework, feature selection, handwritten authorship, significant features, forensic document analysis.

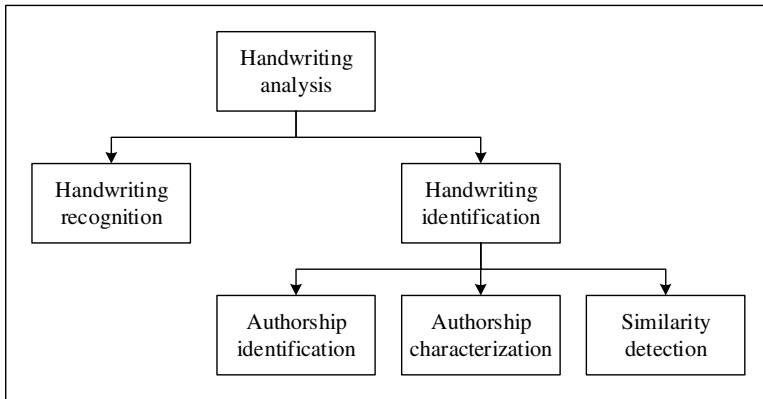
## 1 Introduction

Everyone in this world possesses their own uniqueness, whether in physical, appearance, and characteristics. These unique features are making each and every person discernible from the others. Generally, unique features used to identify an individual are biological feature, such as fingerprint, handprint, hand geometry, face, or voice. There is one feature which is not commonly used, even not a part of biological feature, which is handwriting [1]. This feature is a derivate feature of hand geometry, but also affected by other factors. The complexities of the process to produce handwriting, even the simplest alphabet letter, making this process is capable to identify someone. Even when two writers produce two handwritings that look similar, there are some features that can be used to differentiate their writings. Meaning, even someone can fake the handwriting of another person, but there are some features exist only in the original writing, this is because the original and the fake writings are having different features. Even though in the reality the handwriting will be changed due to its writer's physical and emotional condition, the unique features of one person always exist on his writing, regardless of the condition. Due to its uniqueness and consistency, the features in the handwriting are used to analyze and authenticate forensics documents [2].

The use of handwritten paper documents has never been diminished although the world has lived in digital age for quite some time. There have always been situations in which unsigned or anonymous writings on documents were potentially important. Thus, the provision of proof respecting the authorship of such documents has long been an issue [2]. The Questioned Document Examination (QDE) is an area of the Forensic Science with the main purpose to answer questions related to questioned document (authenticity, authorship and others) and has a large field of applications. There are basically two different sub-areas in the QDE: the document analysis and the handwriting analysis [2]. The first one evaluates the structural analysis of the document to find adulteration, falsification, obliteration and others, while the second investigates the originality or the association between one or more manuscripts to an author [3], for instance when validating the purchase using credit cards, where the card's owner signature on the receipt is slightly different than the signature stored by

the bank, or in the opposite situation where the forger signature is similar to the card's owner. Handwriting analysis is applied to many types of investigation like fraud, homicide, suicide and others, and it has two basic analysis subjects, manuscripts and signatures. Even with distinct features, both keep a narrow relation having the same root or origin in the writer's learning process, in other words, they carry the experiences acquired by the writer during and after his learning process through the improvement of the handwriting personal style [2].

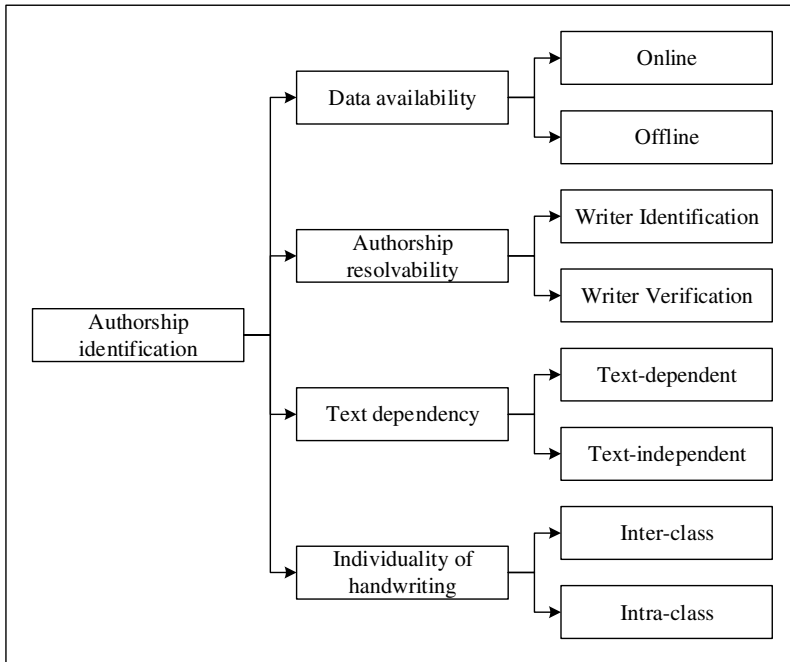
The handwriting analysis research field consists of two categories, which are handwriting recognition and handwriting identification. Fig. 1 depicts the handwriting analysis domain. Handwriting recognition deals with the contents conveyed by the handwritten word, while handwriting identification tries to differentiate handwritings to determine the author [4]. Handwriting identification can be categorized into handwritten authorship identification, handwritten authorship characterization, and similarity detection. Authorship characterization is aimed at inferring an author's background characteristics rather than identity. Similarity detection compares multiple pieces of writing without identifying the author. Handwritten authorship identification, or simply known as authorship identification, evaluates the possibility of one author produces a written document by examining other documents produced by that author [5]. Although authorship identification is categorized as QDE research area, it has evolved into its own matured domain, where the application of authorship identification is not always related to QDE. Authorship identification contributes great importance towards the criminal justice system and has been widely explored in forensic handwriting analysis [4, 6-12]. Nevertheless, there are also many issues and scenarios in authorship identification that pose as challenges which require further investigations and explorations.



**Fig. 1.** Handwriting analysis domain [5]

The performance of pattern recognition applications is heavily depended on the feature extraction and classification method employed [13, 14], which leads to the key concern issue in authorship identification: acquiring the features reflecting the author of handwriting, namely unique individual significant features [4, 11, 15-20].

The essence of authorship identification is to identify a set of features that remain relatively constant among a number of writings by a particular author, and in such a process, the classification technique is very important to the performance of authorship identification [5]. A survey conducted by [5] found a number of studies that show the discriminating power of different types of features, by which researchers attempt to identify an optimal set of features for authorship identification. There are several broad categories for authorship identification, which are platform, author resolvability, text dependency, and individuality of handwriting [4, 21, 22], and shown in Fig. 2.



**Fig. 2.** Authorship identification category [4, 21, 22]

The first category of authorship identification is the platform of the system itself. The platform of the system can be categorized into two, which are offline system and online system [4]. The terms of offline and online system are referring to the input method of the system, rather than the location of the system (as the web application or stand-alone desktop application). Offline system acquires its input from scanned documents or images, while online system acquires its input from touch-sensitive, motion-sensitive, gesture-sensitive, and pressure-sensitive acquiring devices, such as tablets, and thus contains temporal information and theoretically should provide more accurate results [4, 21]. Therefore, online and offline systems have different set of problems and information and thus require different processing methods.

The second category is author resolvability, which consists of two domains: Writer Identification (WI) and Writer Verification (WV). WI performs a one-to-many search in a large database with handwriting samples of known authorship and returns a likely list of candidates, while WV involves a one-to-one comparison with a decision whether or not the two samples are written by the same person, by determining whether the distance between two chosen samples is smaller than a predefined threshold [22]. Furthermore, there are two modes of WV, claim verification and questioned document verification. In the first mode, the system verifies the claim made by a person previously enrolled in the system, while in the second mode, verification problem verifies whether two given documents, questioned document, whose identity need to be verified and reference document, which is collected from the writer for comparison, belong to the same writer or not. The writer of the reference document may or may not be known. The difference between the two is that in this case no database of writers is available and thus, a threshold cannot be computed. In order to solve the problem, some statistical measure such as hypothesis testing, standard deviation, and mean square error is needed to compute the significance of the score [21, 22].

On the other hand, WI can be included as a particular kind of dynamic biometric in pattern recognition for forensic application. WI distinguishes writers based on the shape or individual writing style while ignoring the meaning of the word or character written, due to the differences between one author to another in terms of character association, shape, and the writing style [4, 9, 11, 23-26]. Although there are variances of writing in times, the individual writing style is persistent [4, 9, 11, 23, 27, 28]. And thus, the significant individual features are generalized as the unique features that are persistent regardless of the handwriting shape. The key concern in WI is in acquiring the features reflecting the author of handwriting [4, 11, 15-20]. Thus, it is an open question whether the extracted features are optimal or near-optimal to identify the author. [29] discussed several experiments conducted by various researchers in order to improve WI. [30] treated WI as a texture analysis problem using multichannel Gabor filtering and grey-scale co-occurrence matrix techniques, [31] and [32] addressed the problem of writer verification by casting it as a classification problem with two classes: authorship and non-authorship, [33] morphologically processed horizontal projection profiles on single words, [34] and [35] proposed edge-based directional probability distributions and connected component contours as features, [36] introduced graphemes as features for describing the individual properties of handwriting, and [37] presented a set of eleven features which can be extracted easily and used for the identification and verification of documents containing handwritten digits.

From text dependency point of view, authorship identification can be divided into two broad categories, which are text-dependent and text-independent methods. The text-dependent methods are very similar to signature verification techniques and use the comparison between individual characters or words of known semantic content, and therefore require the prior localization and segmentation of the relevant information. The text-independent methods use statistical features extracted from the entire image of a text block, and thus a minimal amount of handwriting is necessary in order

to derive stable features insensitive to the text content of the samples [4, 22]. Text-dependent methods provide high accuracy and confidence with small amount of data, which is practically not possible for text-independent systems. However, they are more prone to forgery, as the verification text is known in advance. In case of text-independent systems, forgery is not a major problem as the text-independent systems extract less frequent properties from the handwritten document that are difficult to forge [4, 21, 38].

The last category, individuality of handwriting is deemed as the most important issue in authorship identification, which is the main key to identify the author and is closely related to feature extraction task, and thus it is defined as the variance between features for intra-class must be lower than variance between features for inter-class [4]. It relies on two principles: (1) habituation, since people are primarily creatures of habits and writing is the collection of those habits, which are considered neither instinctive nor hereditary but are complex processes that are developed gradually, and (2) individuality or heterogeneity of handwriting, in which each individual had his own style of writing and no two individuals can have the same handwriting [21]. It is only possible to the extent that the variation in handwriting style between different writers exceeds the variations intrinsic to every single writer considered in isolation [22]. It can be proven using similarity error [25, 33, 37, 39] and has been explored by many researchers [4, 26, 28, 39].

In theory, the discriminating power directly relates to the number of features, nevertheless the vast machine learning algorithms practical experiences often proves this does not always apply. The learning process becomes more and more difficult during the training phase if there are too many irrelevant and redundant information, or worse, if the data is noisy and unreliable [40, 41]. Coherent with this traditional concept, the search for the unique feature for every individual in WI domain must consider the condition where the feature for one author may be similar to other authors, and thus should be omitted because of its non-uniqueness. This search objective is similar to the purpose of the feature selection, where the resulting subset is the discriminator between one classes to other classes. Hence, the feature selection phase should be incorporated after feature extraction phase in WI framework, and thus reduce the number of features used and improve the classification performance and accuracy [42]. Since features are regarded as an abstract representation of handwriting, the quality of the feature selection directly influences this representation [5]. Therefore, the purpose of feature selection in this chapter is to acquire the unique features that represent the author of the handwriting in WI domain.

Many previous works have explored the use of feature selection in WI domain [5, 29, 37, 43, 44]. And yet, these studies have not fully addressed the issue in WI domain itself, because instead of acquiring the unique individual significant features to reflect the author of handwriting, these studies focus on the acquiring the features that distinguish one author to another. While the general and common approach does produce good result, it has no significant differences with other pattern recognition problems, since the concept of Individuality of Handwriting is not apparent. Individuality of Handwriting is the most important issue in WI domain, which is the main key to identify the handwritten authorship and is closely related to feature extraction task.



Motivated by the success of the framework proposed by [4], where the global features of handwriting is extracted and thus the Individuality of Handwriting is preserved by using Invariant Discretization, this chapter is trying to further improve the quality of the global features of handwriting produced by acquiring the features that is representing the author of the handwriting using feature selection technique. These representative features must always be existed in every handwriting produced by the same author and should provide enough discriminating power to differentiate the author from other authors. These discriminative features are called unique individual significant features. Because the unique individual significant features are different from one author to other authors, general pattern recognition framework may not be suitable for acquiring these features. The framework employed for this specific task must be capable of acquiring different set of significant features for every author. There are several existing frameworks that is capable of acquiring class-specific features subset, however these frameworks should be modified prominently or they employs feature selection technique that is not suitable for acquiring unique individual significant features.

Therefore, a robust framework to cater this problem must be developed, and at the same time, employs the effectiveness of feature selection to acquire the unique individual significant features. Embarking from these motivations, this chapter is conducted in order to devise a novel feature selection technique which is capable to acquire these unique individual significant features. Furthermore, the proposed technique itself is not working on its own. The proposed technique is developed as a part of vigorous framework, specifically devised for WI domain. The proposed framework employs proposed feature selection technique as the mechanism to acquire the unique individual significant features which is unique to each author. The acquisition of unique significant features also allows the performance of the proposed framework to exceed the performance of existing WI framework [4].

## **2 Existing frameworks for Handwritten Authorship Identification in Forensic Document Analysis**

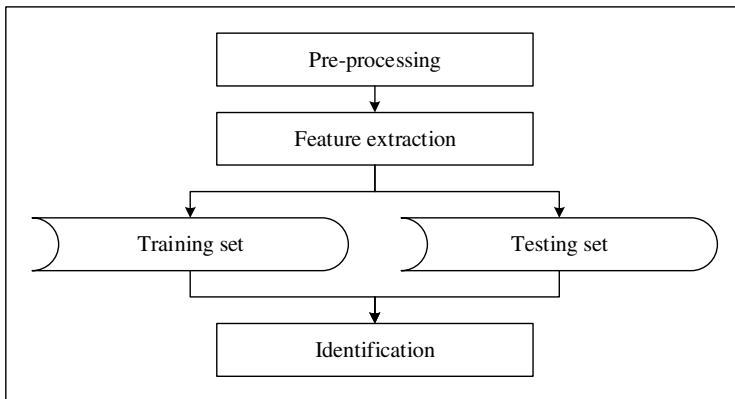
Writer Identification (WI) is an active area of research in pattern recognition due to extensive exchange of paper documents, although currently the world has already moved toward the use of digital documents. WI distinguishes writers based on the handwriting, and ignoring the meaning of the words. Previous studies have explored various methods to improve WI domain, and these studies produced the satisfying performance. However, the use of feature selection as one of important machine learning task is often disregarded in WI domain, which has been proven in the literature where only a handful of studies implemented feature selection task in the WI domain [29, 43, 44].

The key concern in WI is in acquiring the features reflecting the author of handwriting. Although WI is still attracting a vast array of researches since a long time, predominantly in forensic and biometric applications, the question of whether

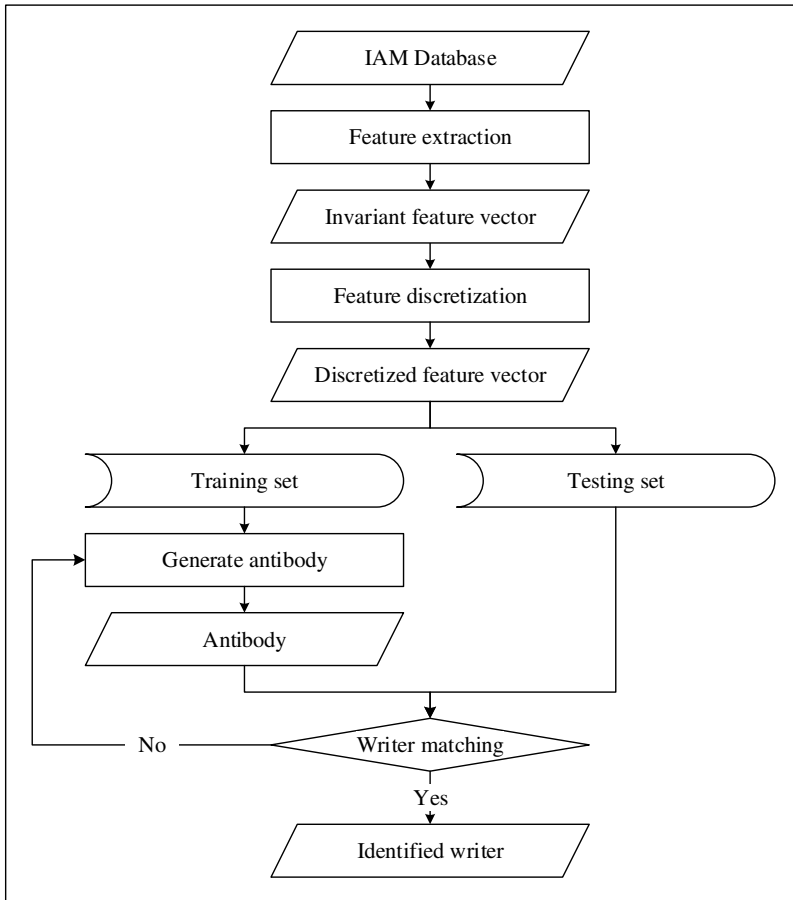
the extracted features are optimal or near-optimal to identify the author is still remain unanswered. This is because the extracted features may include many garbage features. Such features are not only useless in classification, but sometimes degrade the performance of a classifier designed on a basis of a finite number of training samples [4, 42, 45-49]. The features may not be independent of each other or even redundant. Moreover, there may be features that do not provide any useful information for the task of WI [29, 41, 42]. Therefore, feature extraction and selection of the unique individual significant features are very important in order to identify the writer, moreover to improve the classification accuracy.

Handwritten words are very effective in discriminating handwriting, and thus in the study conducted by [4], the holistic approach of global features is used where cursive word is defined as one indivisible entity and extracted by using United Moment Invariant (UMI) [50] technique. Individual features can be acquired by using feature selection technique, by selecting the subset of features. Although in theory, more features provide more discerning power, but in the reality it will degrade significantly the performance [40]. Thus, it is vital to acquire individual features and to perform feature selection for these features, because this will provide simpler identification process and improve the performance of identification in identifying the author.

WI is a part of pattern recognition domain, specifically in handwriting analysis. Thus, traditional pattern recognition framework is appropriate for solving the problem of WI, which is pre-processing, feature extraction and classification. The most recent work to enhance the traditional WI framework is the introduction of an enhanced framework specifically for WI domain proposed by [4], termed as Enhanced WI Framework (EWIF), which consists of feature extraction, feature discretization, and classification. The framework design for traditional pattern recognition framework and EWIF are shown in Fig. 3 and Fig. 4 respectively.



**Fig. 3.** Traditional pattern recognition framework



**Fig. 4.** Enhanced WI Framework [4]

Feature extraction is a process of converting input object into feature vectors. The extracted features are in real value and unique for each word. By using UMI, a digital image is converted to a set of moments which represents the global characteristics of an image shape. Global Moment Function can be used to generate a set of moments that uniquely represent the global characteristic of an image. Moments are scalar quantities used to characterize a function and to capture its significant features. Moment Invariants are very useful tools for pattern recognition [50]. The first introduction of Moment Invariants to pattern recognition and image processing was the employment of algebraic invariants theory by [51], which derived his renowned seven invariants to the rotation of 2D objects. And thus ever since, it has been chosen as one of the most important and frequently used shape descriptors options. Even though they suffer from certain intrinsic limitations (the worst of which is their globalness,

which prevents direct utilization for occluded object recognition), they frequently serve as “first-choice descriptors” and as a reference method for evaluating the performance of other shape descriptors [52]. Geometric Moment Invariants (GMI) [51] presents a set of moments based on combinations of algebraic invariants. This is complied with the definition of invariants given by [53]: an image or a shape feature is invariant if that image or shape undergoes one or a combination of linear transformations. The moments are normalized using (1).

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{(p+q+2)/2}} \tag{1}$$

where  $\mu_{pq}$  is the first, second, and third order of moment which represent the center of the image, measure the variance of the image intensity distribution, and denotes the projection of the image respectively,  $\mu_{00}$  is the zero-th order moment which represents the total intensity of the image, and  $p + q = 2, 3, 4, \dots$ . These moments are invariant under the image scale, translation and rotation, and thus there are seven tuples of moment invariant proposed, which are shown in (2).

$$\begin{aligned} \phi_1 &= \eta_{20} + \eta_{02} \\ \phi_2 &= \phi_1^2 + 4\eta_{11}^2 \\ \phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2 \\ \phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\ \phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ \phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\ \phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ &+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \tag{2}$$

However, [54] found that GMI lose its scale invariance in discrete condition. Several improvements to maintain scale invariance are made by [55-57]. All these improvements are not valuable based on both regions and boundaries simultaneously or the formulas are not coincident with Hu’s moments. Therefore, [50] proposed new Moment Invariants called United Moment Invariants (UMI), which is capable of keeping invariant to region and closed and unclosed boundary, both in discrete and continuous condition. The equation of UMI is as shown in (3).

$$\begin{aligned}
\theta_1 &= \frac{\sqrt{\phi_2}}{\phi_1} \\
\theta_2 &= \frac{\phi_6}{\phi_1\phi_4} \\
\theta_3 &= \frac{\sqrt{\phi_4}}{\phi_4} \\
\theta_4 &= \frac{\phi_5}{\phi_3\phi_4} \\
\theta_5 &= \frac{\phi_1\phi_6}{\phi_2\phi_3} \\
\theta_6 &= \left(\phi_1 + \sqrt{\phi_2}\right)\frac{\phi_3}{\phi_6} \\
\theta_7 &= \frac{\phi_1\phi_5}{\phi_3\phi_6} \\
\theta_8 &= \frac{\phi_3 + \phi_4}{\sqrt{\phi_5}}
\end{aligned} \tag{3}$$

where  $\phi_i$  are GMI. The features extracted by UMI are the pattern to represent the image shape. It is also worth mentioning that [50] also found the scale invariance of GMI is untenable in discrete condition and the disunion of invariants formula based on region and boundary. The information of different types of geometrical features of the image is also provided by UMI [58]. The feature extraction phase in this chapter is achieved by using global representation of UMI [50] to acquire the global features of handwriting image, due to the requirement of cursive word is needed to extract as one single indivisible entity.

According to [4], the advantages of global approach are including its capabilities to show the individuality of handwriting [23], is shown to be very effective in reducing the complexity of the word [59], moreover to increase the accuracy of classification, and it is invariant with respect to all different writing styles; hence it holds immense promise for realizing near-human performance [60] and very robust in detecting similar object when it is used in similarity search. Table 1 is the example of feature invariant of words using UMI with eight features vector for each image, with f1 represents the first feature, f2 for second feature, and henceforth.

Many real-world classification tasks exist that involve continuous features where such algorithms could not be applied unless the continuous features are first discretized. Discretization is a process of dividing a range of continuous features into disjoint intervals, which labels can then be used to replace the actual data values [61]. Discretization engages searching for cut-off points that determine intervals and thus unifying the values over each interval. All values that lie within an interval are

**Table 1.** UMI representation for handwritten word image

Word	f1	f2	f3	f4	f5	f6	f7	f8
<i>alone</i>	1.84	1.79	0.91	1.31	0.84	1.00	0.73	1.79
<i>bowed</i>	1.53	1.08	1.12	1.96	0.72	1.49	1.82	1.46
<i>swish</i>	1.61	1.53	0.53	0.38	0.80	1.26	0.25	3.29
<i>scheme</i>	1.99	8.24	0.65	0.76	3.77	0.20	0.09	2.40
<i>the</i>	3.08	2.06	0.52	0.64	0.52	0.82	0.31	2.75

mapped to the same value, in effect converting numerical attributes that can be treated as being symbolic [62]. Discretization largely contributes to rough set theory [63] and provides more comprehensible knowledge representation which is reduced and simplified [64], and thus more accurate and faster. Continuous variable discretization has recently received significant attention in the machine learning domain [65]. The goal of discretization is to find a set of cut points to partition the range into a small number of intervals that have good class coherence, which is usually measured by an evaluation function. In addition to the maximization of interdependence between class labels and attribute values, an ideal discretization method should have a secondary goal to minimize the number of intervals without significant loss of class-attribute mutual dependence [66].

Discretization is usually performed prior to the learning process and it can be broken into two tasks. The first task is to find the number of discrete intervals. Only a few discretization algorithms perform this; often, the user must specify the number of intervals or provide a heuristic rule. The second task is to find the width, or the boundaries, of the intervals given the range of values of a continuous attribute [66]. Discretization is in EWIF employed because the features extracted from feature extraction phase are in continuous forms [61]. [4] argues that discretization is important in order to obtain the detachment of writers' individuality and produce better data representation, thus [4] proposes a new discretization technique namely Invariant Discretization, which provides standard representation of individual features, which allows small variance between features for intra-class (same author) and large variance for inter-class (different authors). Although feature discretization provides better representation of individual features, this mechanism only partially reflect the key concern in WI domain.

In the classification phase, the correct identification accuracy, or termed classification accuracy is calculated to measure the quality of feature subset produced from feature selection phase. EWIF applies Modified Immune Classifier (MIC) [4]. In order to detect the same features of a writer, the detector must complement the antigen. The detector in this case is the features extracted from training dataset, while the antigen is the features extracted from testing dataset or questioned handwriting [4]. The classifier consists of two modules: censoring module and monitoring module. The detector is generated with the complementary of the self-cell in the censoring module.

On the other hand, monitoring module is the matching process, and the term “bind” is adopted in order to describe the matching process, which is due to the complementary of the self-cell which is defined as the detector in the censoring module [4]. MIC uses several binary matching techniques, which are Hamming distance,  $r$ -Chunk,  $r$ -Contiguous, and Multiple  $r$ -Contiguous. The binary strings are used to represent the detectors and antigens, which forms the binary matching rule.

The inclusion of feature selection calls for the further improvement to the EWIF. This is because the main drawback of EWIF is that the mechanism to acquire the unique significant features is not present and is not defined as the part of the framework; instead the whole features are used for the identification phase. The acquisition of the unique significant features is apparently one of the important issues on WI domain because it provides more effective way to identify the handwritten authorship [4], and this issue is not addressed in the EWIF.

### 3 Swarm-Based Feature Selection Technique

Feature selection has become an active research area for decades, and has been proven in both theory and practice [40]. The main objective of feature selection is to select the minimally sized subset of features as long as the classification accuracy does not significantly decreased and the result of the selected features class distribution is as close as possible to original class distribution [42].

The feature set produced from feature extraction phase in traditional framework or discretization phase in Enhanced Writer Identification Framework (EWIF), may consist of relevant and irrelevant features. There will be more complexities produced in terms of accuracy and performance, if these features are used directly in classification phase. Although in theory, more features provide more discerning power, but in the reality it will degrade significantly the performance [40]. Hence, the feature selection phase should be incorporated after discretization phase, and thus reduce the number of features used and improve the classification performance and accuracy [42]. Feature selection phase should be able to filter those features and select the most unique individual significant features in the process. Therefore, selection of the unique individual significant features is very important in order to identify the writer.

Wrapper feature selection method possesses unique advantages and disadvantages. A wrapper algorithm explores the space of features subsets to optimize the induction algorithm that uses the subset for classification. The rationale for wrapper methods is that the induction method that will ultimately use the feature subset should provide a better estimate of accuracy than a separate measure that has an entirely different inductive bias [41, 67]. These methods based on penalization face a combinatorial challenge when the set of variables has no specific order and when the search must be done over its subsets since many problems related to feature extraction have been shown to be NP-hard [68]. Advantages of wrapper method are the ability to include the interaction between feature subset search and model selection, and take into account feature dependencies. On the other hand, the disadvantages are that it has higher risk of over-fitting than filter methods and are very computationally intensive [69].

Therefore, this section describes the method to optimize selected feature selection technique, particularly in diminishing computational cost.

Several techniques have been introduced throughout the last decade to reduce the complexity of wrapper method, for instance is by infusing it with recent stochastic optimization [44, 70-75], controlling the number of cross-validation [76], and hybridizing with filter methods [77, 78]. However, very few studies conducted in utilizing concurrent programming techniques [79-81]. Studies shown that implementing concurrent programming, specifically multithreading, sanctions much lesser processing time [79-83]. Therefore, the first optimization applied towards wrapper method is multithreading. This decision is motivated by the fact that wrapper technique is computationally expensive; therefore it constrained the possibility of hybridization since it will consume more resources and requires higher computational cost, and hence direct hybridization with stochastic optimization may not be the wisest option.

Considering the advantages of switching from sequential programming towards concurrent programming, or in this case is multithreading, and the lack of focus for multithreading in feature selection techniques, leads to the decision to adapt multithreading in Sequential Forward Floating Selection (SFFS) [84]. SFFS is an extension of Sequential Forward Selection [85], which suffers from the nesting effect, meaning that once a feature is included in some step of the iterative process, it cannot be excluded in a later step. SFFS performs a simple hill-climbing search. The best feature subset  $S$  is initialized as the empty set and perform the forward selection, where in each step a new subset is generated first by adding a feature  $x^+$ , but after that features  $x^-$  is searched for to be eliminated from  $S$  until the classification accuracy  $J(S \setminus x^-)$  decreases, which is called as backward selection. The iterations continue until no new feature can be added because the classification accuracy  $J(S \cup x^+)$  does not increase.

Multithreaded SFFS is capable to reduce the computational cost of original SFFS, not only because of the introduction of multithreading, but also because of the introduction of a novel mechanism called merit pooling. Merit pooling refers to the process of pre-calculating and storing the merit of each feature before the selection process take hand. This mechanism reduce a great deal of processing time, because instead of recalculating the merit of the feature subset every time a feature is added or removed, the proposed technique will simply sum up the merit values for each individual feature in the subset which has been stored in the merit pool previously. The resulting merit value will also be stored in the merit pool, so that future subset that has same feature member will simply use this value, without having to re-looking up the merit of individual member in the merit pool. In the original implementation of SFFS, each time a feature is added or removed from the feature subset, the merit of the subset will be calculated by repeatedly calling the induction algorithm. The process of calculating the merit is oftentimes the primary source of high computational cost of wrapper methods [41].

However, it is found that multithreaded SFFS performs not as well as original SFFS, although it opens the possibility of hybridization with swarm intelligence. In this chapter, Particle Swarm Optimization (PSO) [86, 87] is selected as the best way to optimize multithreaded SFFS. PSO is a population-based optimization method,



which can be used to solve a wide array of different optimization problems. PSO is a stochastic algorithm that does not need gradient information derived from the error function. This allows the PSO to be used on functions where the gradient is either unavailable or computationally expensive to obtain. The origin of the PSO was based on the sociological behavior of bird flocking [87]. PSO initially identifies some particle as the best particle in a neighborhood of particles based on its fitness. All the particles are then accelerated in the direction of this particle, but also in the direction of their own best solutions that they have discovered previously. All particles also have the opportunity to discover better particles, in which case the other particles will change direction and head towards the new “best” particle. By approaching the current best solution, the neighboring solutions will be discovered by some of the particles. It is important to realize that the velocity term models the rate of change in the position of the particle.

The success of PSO implementation on the Writer Identification (WI) domain has also been demonstrated by [88]. Other consideration taken for selecting PSO is also due to its simple yet effective implementation. Because of this characteristic, PSO is not increasing the computational complexity of multithreaded SFFS more than necessary. The hybridization with PSO is primarily to prevent the multithreaded SFFS selects the local optima, and forces it to reevaluate the candidates with the same merit in every iteration to find the global optima. The hybrid between two techniques is dubbed Swarm Optimized and Computationally Inexpensive Floating Selection (SOCIFS). The main idea of SOCIFS is that fitness function of PSO is modified, by implementing the classification accuracy of unique individual significant features acquired by using multithreaded SFFS. This is to allow the most optimal interaction between PSO and multithreaded SFFS, and thus allow for wider search space exploration. Furthermore, there are multiple instances of multithreaded SFFS executed concurrently; each of it is executed in PSO particle. Fitness function  $f(X(t))$  in SOCIFS is defined in (4) and (5), derived from [87].

$$X = \begin{cases} S \cup x^+, & \text{if forward selection} \\ S \setminus x^-, & \text{if backward selection} \end{cases} \quad (4)$$

$$f(X(t)) = \alpha \times \gamma_{X(t)} + \beta \times \frac{|N| - |X(t)|}{|N|} \quad (5)$$

where  $\gamma_X(t)$  is the merit of particle  $i$  current subset  $X$  in iteration  $t$ , where the value is obtained by multithreaded SFFS.  $|N|$  is the number of features, while  $|X(t)|$  is the size of selected feature subset.  $\alpha$  and  $\beta$  are the parameters used to determine the importance of classification accuracy and the subset size, where  $\alpha \in [0, 1]$  and  $\beta = 1 - \alpha$ . Each particle will examine different feature subset and thus produce unique results, this is because the examined feature subset and its results are recorded, to prevent different particles examine the same subset multiple times. The algorithm of SOCIFS is illustrated in Fig. 5.

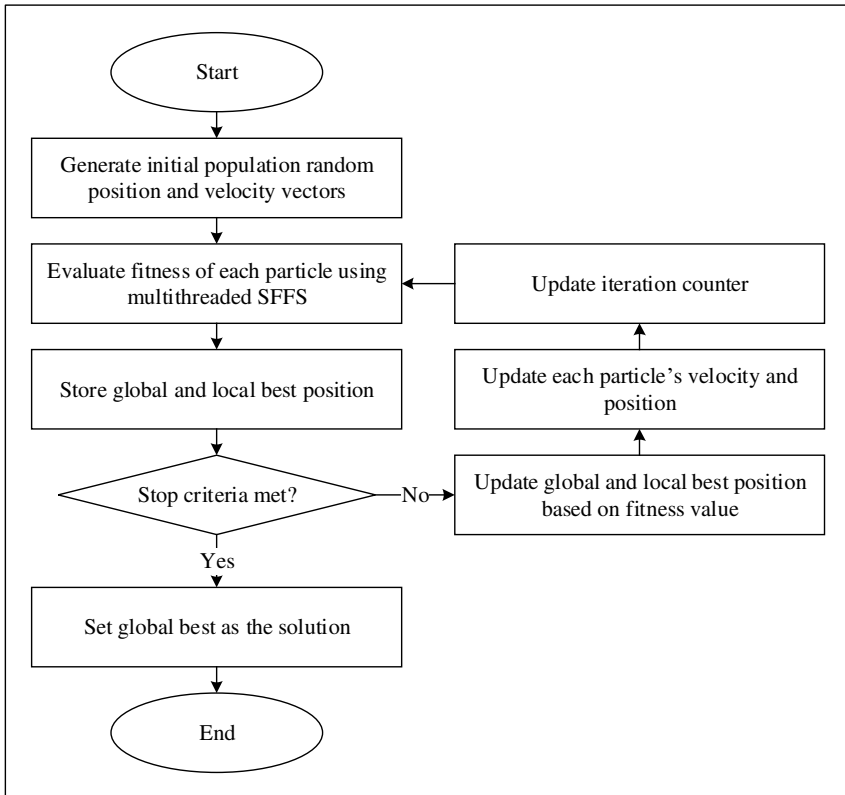
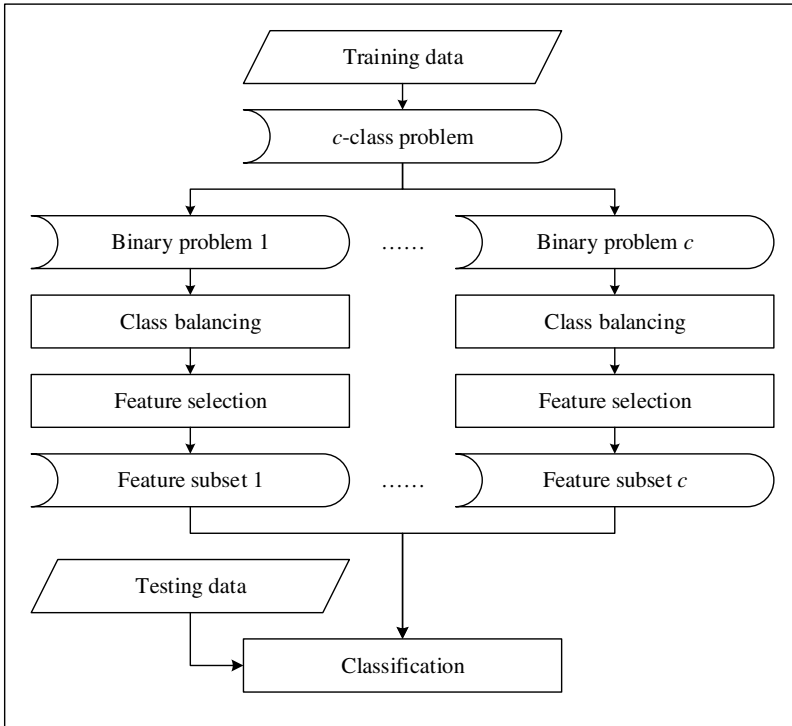


Fig. 5. Swarm Optimized and Computationally Inexpensive Floating Selection (SOCIFS)

#### 4 Swarm-Based Framework for Handwritten Authorship Identification

The main issue in Writer Identification (WI) is to acquire the individual features from various handwritings [4]. Among these features are exists the significant individual features which directly unique to those individual. Based on this description, it is concluded that each individual possess different unique significant feature. Therefore, class-specific feature selection must be incorporated in order to capture these unique individual significant features. Even though traditional feature selection techniques can be used for acquiring these unique individual significant features [89-92], it may not be appropriate and feasible. And thus, the traditional handwriting identification framework, which consists of pre-processing, feature extraction and classification [93] is not adequate for this issue. Enhanced WI Framework (EWIF) shown in Fig. 4 [4], consists of feature extraction, feature discretization, and identification has been adopted by [90, 92] and produced good result, and therefore it can be concluded that this framework is can be further improved.

This section describes proposed swarm-based framework to cater with this class-specific feature selection issue, namely Cheap Computational Cost Class-Specific Swarm Sequential Selection ( $C_4S_4$ ). Furthermore, the proposed framework is similar with General Framework for Class-Specific (GFCS) feature selection framework [94], which is shown in Fig. 6. Therefore, it can be assumed the proposed framework is a hybrid of GFCS and EWIF. And thus, several modifications should be implemented in GFCS, considering that GFCS is proposed to handle wide-range of application and domain. The proposed framework differs from GFCS and EWIF in several aspects. The differences between these frameworks are summarized in Table 2.



**Fig. 6.** General Framework for Class-Specific Feature Selection (GFCS) [94]

**Table 2.** Summary of EWIF, GFCS, and  $C_4S_4$  differences

Criteria	EWIF	GFCS	$C_4S_4$
Feature extraction	Yes	-	Yes
Feature discretization	Yes	-	Yes
Class binarization	-	Yes	Yes
Class balancing	-	Yes	Yes
Feature selection	-	Yes	Yes
Antibody pool	Yes	-	Yes

GFSC is selected in this as the basis for the proposed feature selection framework because it is designed to select the class-specific feature subset, which is similar to the concept of acquiring the unique significant features in WI domain. The first difference of  $C_4S_4$  to GFCS is that the  $C_4S_4$  includes feature extraction and feature discretization stage, originating from EWIF, and thus produced training and testing set. After that, the framework works similarly with GFCS, which is to use the one-against-all class binarization in order to transform a  $c$ -class problem into  $c$  binary problems. For each class  $w_i$ ,  $i = 1, \dots, c$ ; a binary problem  $\langle w_i \Omega_i \rangle$  where  $\Omega_i = \bigcup_{j=1, j \neq i}^c w_j$ , is created for

the training data. For each binary problem the instances of the class  $w_i$  are used as positive examples, and the instances of all other classes are used as negative examples. The generated binary problems could be imbalanced; therefore the next stage is necessary to balance the classes by applying an oversampling by repeating training instances method.  $\beta_i = |w_i| - |\Omega_i|$  is then computed in the next stage, where  $|w_i|$  is the number of instances in class  $w_i$ , and  $|\Omega_i|$  is the number of instances in the remaining classes. If  $\beta_i > 0$ , the classes will be balanced by repeating instances in the class  $w_i$  until the number of instances in  $w_i$  and  $\Omega_i$  are the same. For each binary problem, features are selected in the third stage by using Swarm Optimized and Computationally Inexpensive Floating Selection (SOCIFS), and the selected features are assigned to the class from which the binary problem was constructed. In this way,  $c$  possible different feature subsets are obtained, one for each class of the original  $c$ -class supervised classification problem, or unique individual significant features in this domain. These  $c$ -feature subsets are in turn is transformed into  $c$ -antibodies and stored in antibody pool that consists of all antibodies, which in turn is used in identification stage.

On the other hand, the first difference between  $C_4S_4$  and EWIF is that the feature discretization is conducted before splitting dataset into training and testing dataset in EWIF, whereas the feature discretization is conducted after the dataset has been split into training and testing dataset in  $C_4S_4$ . This process is closely representing the real-life applications, where the testing dataset is not available to the system beforehand and thus should not be included in the training process. However, this process aroused another problem, since the training dataset is discretized while the testing dataset is not. The same discretization method cannot be directly applied to the testing dataset, because it will produce different set of data due to different cut-off points and intervals is employed. This problem is solved in  $C_4S_4$  by storing the discretization rules for each class, which are the cut-off points and intervals. These discretization rules are employed during the classification phase, where the testing data will be discretized using each class discretization rule before the matching process is performed. In another word, an instance of the testing data will be casted into  $c$  number of instance with different values due to different rule before the identification phase will take place. The classification results for these  $c$ -instances of testing data will be ranked. The class that corresponds to the discretization rule with the highest ranking will be identified as the final class. These processes are the framework of  $C_4S_4$ , which is illustrated in Fig. 7.

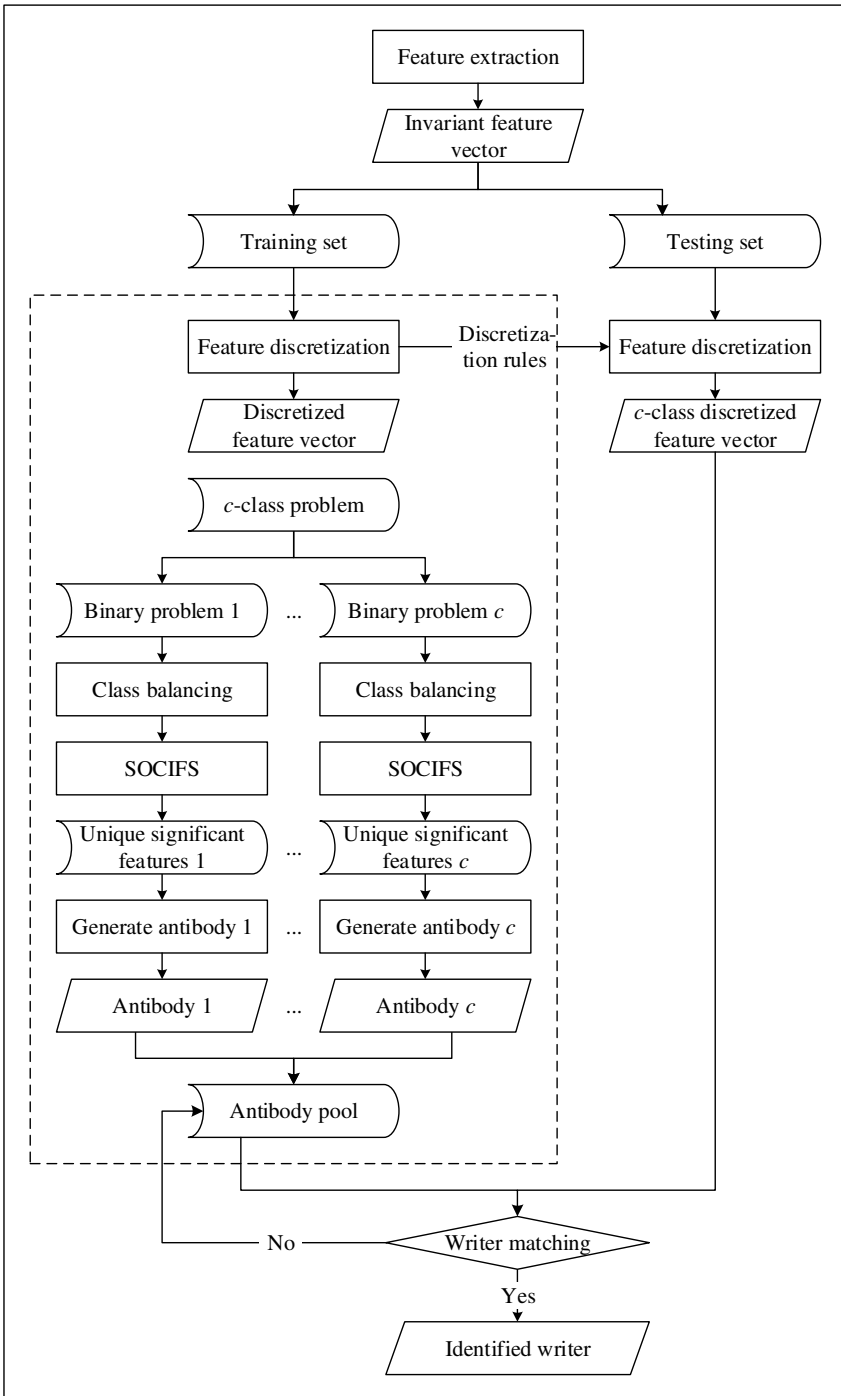
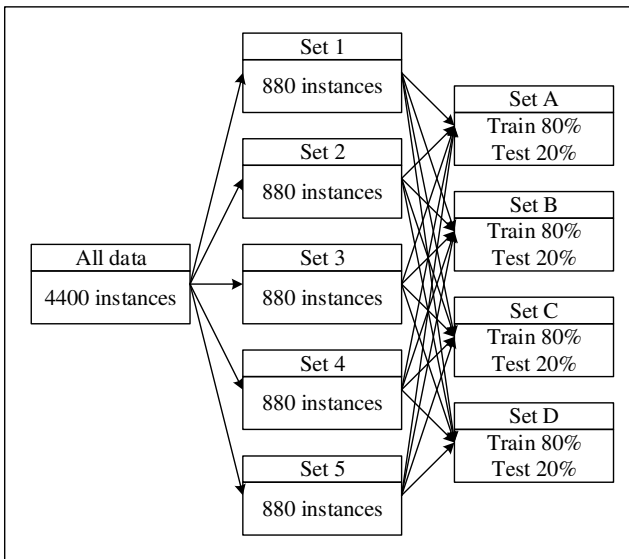


Fig. 7. Cheap Computational Cost Class-Specific Swarm Sequential Selection (C<sub>4</sub>S<sub>4</sub>)

## 5 Results and Discussions

The quality of proposed framework must be justified via performance measurements. Dataset used for the performance measurements comes from IAM Handwriting Database [95], which is developed by Research Group on Computer Vision and Artificial Intelligence at Instituts für Informatik und angewandte Mathematik (IAM) in Universität Bern, Switzerland. This database contains forms of handwritten English text. It can be used to train and test handwriting recognition techniques, and to perform writer identification and verification experiments.

Sixty (60) classes are used for research. From these 60 classes, 4400 instances are collected, and are randomly divided into four different datasets to form training and testing dataset in the classification task. The ratio between the number of training and testing dataset is 4:1, which is actually the simple way of describing 5-fold cross-validation. These four datasets are as depicted in Fig. 8.

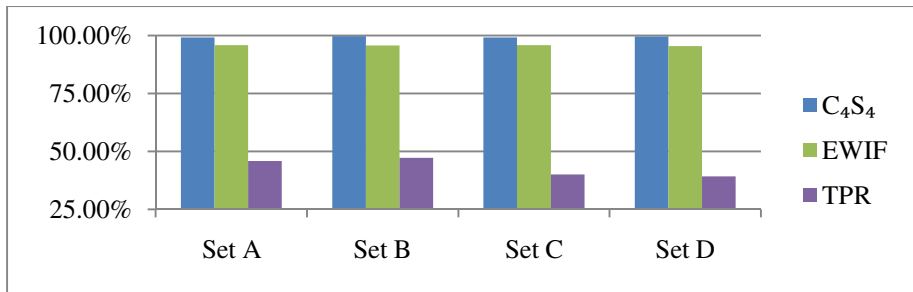


**Fig. 8.** Data collection procedure

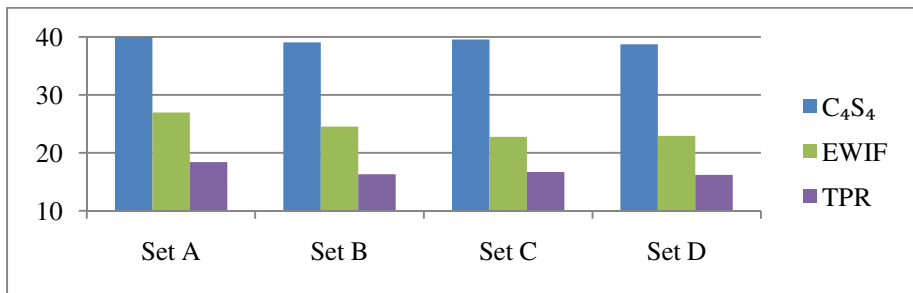
The three commonly used performance measurements for evaluating the performance of feature selection technique are number of selected features, classification accuracy, and processing time. However, considering that the Cheap Computational Cost Class-Specific Swarm Sequential Selection ( $C_4S_4$ ) will produce different size of feature subset for different class, number of selected features performance measurement will be omitted in this analysis. This analysis will compare the performance of proposed framework to Enhanced Writer Identification Framework (EWIF) and traditional pattern recognition (TPR) framework. Table 3 presents the classification accuracy and processing time results for  $C_4S_4$ , EWIF, and TPR in four datasets. The results are also depicted in bar chart format in Fig. 9 and Fig. 10.

**Table 3.** C4S4, EWIF, and TPR results on classification accuracy and processing time

Criteria	Framework	Set A	Set B	Set C	Set D
Classification accuracy	C4S4	99.10%	99.65%	99.09%	99.55%
	EWIF	95.82%	95.65%	95.78%	95.35%
	TPR	45.88%	47.24%	40.14%	39.23%
Processing time	C4S4	39.97 sec.	39.05 sec.	39.55 sec.	38.72 sec.
	EWIF	26.97 sec.	24.56 sec.	22.75 sec.	22.91 sec.
	TPR	18.41 sec.	16.29 sec.	16.73 sec.	16.18 sec.



**Fig. 9.** C4S4, EWIF, and TPR results for classification accuracy



**Fig. 10.** C4S4, EWIF, and TPR results for processing time

The classification accuracy of proposed framework and feature selection techniques are the primary consideration of this chapter. Based on the results shown in Table 3 and presented graphically in Fig. 9 and Fig. 10, the proposed framework produces the best average of classification accuracy, 99.35%; moreover, the result is significantly exceeding the result of EWIF (95.65%) and TPR (43.12%). The results produced by C4S4 shows that the incorporation of feature selection to EWIF is capable to improve its performance. The second measurement of this chapter is processing time of proposed framework and feature selection techniques. Based on the result, it is shown that there is no trade-off between classification accuracy of C4S4 and its processing time. The average processing time of C4S4 is only approximately 15 seconds longer than EWIF (39.32 to 24.30 seconds) and approximately 23 seconds longer than TPR (39.32 to 16.90 seconds).

## 6 Conclusions

The purpose of this section is to discuss the summary of this chapter. This chapter is inspired by the fact that every person has unique and significant features that can distinguish oneself to other person, which is always consistent in every handwriting, regardless of words written. These unique individual significant features however, are hidden in the shape and of writing, and thus, key concern in Writer Identification (WI) is in acquiring the features reflecting the author of handwriting using various writing styles.

In this chapter, the word shape is first obtained via feature extraction phase using holistic approach of global representation technique in Moment Function. These extracted features are then selected in the feature selection phase using proposed technique. These selected features are the unique individual significant features which are unique to each person, and used in the classification phase in order to identify the handwritten authorship.

The focus of this chapter is to develop a swarm-based framework which is suited in WI domain, specifically in obtaining the significantly unique features of an individual. The development of the proposed technique and framework has been thoroughly discussed. The proposed framework is unique due to the fact that rather than trying to acquire the features which can differentiate one person to another, the proposed framework instead determine which features are unique to one author. The prior method is commonly used in other domains, where it is important to discriminate one class to another class. However, this is not the case in WI domain. If the prior method is used, the features capable to differentiate one author to another author may not exist, because it is possible for one author possess similar features to another author, although this possibility is rather insignificant. Therefore, the latter method is more suitable, because as mentioned earlier, every individual possess unique and individualistic significant features.

As a conclusion, this chapter has successfully proposed a novel swarm-based framework namely Cheap Computational Cost Class-Specific Swarm Sequential Selection ( $C_4S_4$ ) which serves as the major contribution of this chapter. While the proposed technique is still not perfect, it still performs better than existing handwritten authorship identification frameworks. The results validate the quality of the proposed technique and framework and open the opportunity for further exploration in WI domain specifically, and other domains generally.

## References

1. Amend, K., Ruiz, M.S.: *Handwriting Analysis*. The Career Press, New Jersey (1980)
2. Huber, R.A., Headrick, A.M.: *Handwriting Identification: Facts and Fundamentals*. CRC Press, New York (1999)
3. Baranoski, F.L., Oliveira, L.S., Justino, E.J.R.: *Writer Identification Based on Forensic Science Approach*. In: XXXIII Latin American Conference on Informatics, San Jose, Costa Rica, pp. 25–32 (2007)



4. Muda, A.K.: Authorship Invarianceness for Writer Identification Using Invariant Discretization and Modified Immune Classifier. Universiti Teknologi Malaysia (2009)
5. Li, J., Zheng, R., Chen, H.: From Fingerprint to Writeprint. *Communications of the ACM* 49(4), 76–82 (2006)
6. Al-Ma'adeed, S., Mohammed, E., Kassis, D.A., Al-Muslih, F.: Writer Identification Using Edge-based Directional Probability Distribution Features for Arabic Words. In: *Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications 2008*, pp. 582–590. IEEE Computer Society (2008), 1544403
7. Niels, R., Vuurpijl, L., Schomaker, L.: Automatic Allograph Matching in Forensic Writer Identification. *International Journal of Pattern Recognition and Artificial Intelligence* 21(1), 61–81 (2007), doi:10.1142/S0218001407005302
8. Pervouchine, V., Leedham, G.: Extraction and Analysis of Forensic Document Examiner Features Used for Writer Identification. *Pattern Recognition* 40(3), 1004–1013 (2007), doi:10.1016/j.patcog.2006.08.008
9. Srihari, S.N., Huang, C., Srinivasan, H., Shah, V.: Biometric and Forensic Aspects of Digital Document Processing. In: Chaudhuri, B.B. (ed.) *Digital Document Processing. Advances in Pattern Recognition*, pp. 379–405. Springer, Heidelberg (2007)
10. Tapiador, M., Sigüenza, J.: Writer Identification Method Based on Forensic Knowledge. In: Zhang, D., Jain, A.K. (eds.) *ICBA 2004. LNCS, vol. 3072*, pp. 555–561. Springer, Heidelberg (2004)
11. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of Handwriting. *Journal of Forensic Sciences*, 856–872 (2002)
12. Franke, K., Köppen, M.: A Computer-based System to Support Forensic Studies on Handwritten Documents. *International Journal on Document Analysis and Recognition* 3(4), 218–231 (2001), doi:10.1007/PL00013565
13. Liu, C.-L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten Digit Recognition: Benchmarking of the State-of-the-art Techniques. *Pattern Recognition* 36(10), 2271–2285 (2003)
14. Liu, C.-L., Nakashima, K., Sako, H., Fujisawa, H.: Handwritten Digit Recognition: Investigation of Normalization and Feature Extraction Techniques. *Pattern Recognition* 37(2), 265–279 (2004)
15. Xu, D.-Y., Shang, Z.-W., Tang, Y.-Y., Fang, B.: Handwriting-based Writer Identification with Complex Wavelet Transform. In: *International Conference on Wavelet Analysis and Pattern Recognition*, pp. 597–601 (2008)
16. Bensefia, A., Paquet, T., Heutte, L.: A Writer Identification and Verification System. *Pattern Recognition Letters* 26(13), 2080–2092 (2005)
17. He, Z.-Y., Tang, Y.-Y.: Chinese Handwriting-based Writer Identification by Texture Analysis. In: *Proceedings of International Conference on Machine Learning and Cybernetics*, pp. 3488–3491 (2004)
18. Yu, K., Wang, Y., Tan, T.: Writer Identification Using Dynamic Features. In: Zhang, D., Jain, A.K. (eds.) *ICBA 2004. LNCS, vol. 3072*, pp. 512–518. Springer, Heidelberg (2004)
19. Schlaphbach, A., Bunke, H.: Off-line Handwriting Identification Using HMM Based Recognizers. In: *17th International Conference on Pattern Recognition, Cambridge*, pp. 654–658 (2004)
20. Shen, C., Ruan, X.-G., Mao, T.-L.: Writer Identification Using Gabor Wavelet. In: *Proceedings of the 4th World Congress on Intelligent Control and Automation*, pp. 2061–2064 (2002)
21. Gupta, S.: *Automatic Person Identification and Verification using Online Handwriting*. International Institute of Information Technology (2008)

22. Bulacu, M.L.: *Statistical Pattern Recognition for Automatic Writer Identification and Verification*. University of Groningen (2007)
23. Zhang, B., Srihari, S.N.: Analysis of Handwriting Individuality Using Word Features. In: *Proceedings of the Seventh International Conference of Document Analysis and Recognition*, pp. 1142–1146 (2003)
24. Marti, U.V., Messerli, R., Bunke, H.: Writer Identification Using Text Line Based Features. In: *Proceedings of Sixth International Conference on Document Analysis and Recognition*, pp. 101–105 (2001)
25. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of Handwriting: A Validation Study. In: *Sixth IAPR International Conference on Document Analysis and Recognition*, pp. 106–109 (2001)
26. Yong, Z., Tieniu, T., Yunhong, W.: Biometric Personal Identification Based on Handwriting. In: *Proceedings of 15th International Conference on Pattern Recognition*, pp. 797–800 (2000)
27. Zheng, Z., Srihari, R., Srihari, S.N.: A Feature Selection Framework for Text Filtering. In: *Proceedings of the Third IEEE International Conference on Data Mining (ICDM 2003)*, pp. 705–708. IEEE (2003)
28. Srihari, S.H., Cha, S.-H., Lee, S.: Establishing Handwriting Individuality Using Pattern Recognition Techniques. In: *Sixth International Conference on Document Analysis and Recognition*, pp. 1195–1204 (2001)
29. Schlapbach, A., Kilchherr, V., Bunke, H.: Improving Writer Identification by Means of Feature Selection and Extraction. In: *Eight International Conference on Document Analysis and Recognition*, pp. 131–135. IEEE (2005)
30. Said, H.E.S., Tan, T., Baker, K.: Personal Identification Based on Handwriting. *Pattern Recognition* 33, 149–160 (2000)
31. Cha, S.-H., Srihari, S.: Multiple Feature Integration for Writer Verification. In: *Proceedings of 7th International Workshop on Frontiers in Handwriting Recognition*, pp. 333–342 (2000)
32. Zhang, B., Srihari, S.N., Lee, S.: Individuality of Handwritten Characters. In: *Proceedings of 7th International Conference on Document Analysis and Recognition*, pp. 1086–1090 (2003)
33. Zois, E.N., Anastassopoulos, V.: Morphological Waveform Coding for Writer Identification. *Pattern Recognition* 33, 385–398 (2000)
34. Bulacu, M., Schomaker, L., Vuurpijl, L.: Writer Identification Using Edge-based Directional Features. In: *Proceedings of 7th International Conference on Document Analysis and Recognition*, pp. 937–941 (2003)
35. Schomaker, L., Bulacu, M.: Automatic Writer Identification using Connected-component Contours and Edge-based Features of Uppercase Western Script. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 787–798 (2004)
36. Bensefia, A., Paquet, T., Heutte, L.: Handwriting Analysis for Writer Verification. In: *Proceedings of 9th International Workshop on Frontiers in Handwriting Recognition*, pp. 196–201 (2004)
37. Leedham, G., Chachra, S.: Writer Identification Using Innovative Binarised Features of Handwritten Numerals. In: *Proceedings of 7th International Conference on Document Analysis and Recognition*, pp. 413–417 (2003)
38. Siddiqi, I.: *Classification of Handwritten Documents: Writer Recognition*. Université Paris Descartes (2009)
39. He, Z., Youb, X., Tang, Y.-Y.: Writer Identification Using Global Wavelet-based Features. *Neurocomputing* 71(10), 1831–1841 (2008)

40. Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, 1205–1224 (2004)
41. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. University of Waikato (1999)
42. Dash, M., Liu, H.: Feature Selection for Classification. *Journal of Intelligent Data Analysis*, 131–156 (1997)
43. Zhang, P., Bui, T.D., Suen, C.Y.: Feature Dimensionality Reduction for the Verification of Handwritten Numerals. *Pattern Analysis Application*, 296–307 (2004)
44. Kim, G., Kim, S.: Feature Selection Using Genetic Algorithms for Handwritten Character Recognition. In: *Seventh International Workshop on Frontiers in Handwriting Recognition*, Amsterdam, pp. 103–112. International Unipen Foundation (2000)
45. Sewell, M.: Feature Selection (2007), <http://machine-learning.martinsewell.com/feature-selection/feature-selection.pdf> (accessed October 25, 2009)
46. Ahmad, A., Dey, L.: A Feature Selection Technique for Classificatory Analysis. *Pattern Recognition Letters*, 43–56 (2004)
47. Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3, 1157–1182 (2003)
48. Portinale, L., Saitta, L.: Feature Selection: State of the Art. In: *Feature Selection*, pp. 1–22. Università del Piemonte Orientale, Alessandria (2002)
49. Kudo, M., Sklansky, J.: Comparison of Algorithms that Select Features for Pattern Classifiers. *Journal of Pattern Recognition* 33, 25–41 (2000)
50. Yinan, S., Weijun, L., Yuechao, W.: United Moment Invariants for Shape Discrimination. In: *International Conference on Robotics, Intelligent Systems and Signal Processing*, Changsha, pp. 88–93. IEEE (2003)
51. Hu, M.K.: Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, 179–187 (1962)
52. Flusser, J., Suk, T., Zitová, B.: *Moments and Moment Invariants in Pattern Recognition*, vol. 1. John Wiley and Sons, Ltd., West Sussex (2009)
53. Belkasim, S.O., Shridhar, M., Ahmadi, M.: Pattern Recognition with Moment Invariants: A Comparative Study and New Results. *Pattern Recognition* 24(12), 1117–1138 (1991)
54. Ding, M., Chang, J., Peng, J.: Research on Moment Invariants Algorithm. *Journal of Data Acquisition and Processing* 7(2), 1–9 (1992)
55. Lv, H., Zhou, J.: Research on Discrete Moment Invariance Algorithm. *Journal of Data Acquisition and Processing* 1(2), 151–155 (1993)
56. Wang, B., Sun, J., Cai, A.: Relative Moments and Their Applications to Geometric Shape Recognition. *Journal of Image and Graphics* 6(3), 296–300 (2002)
57. Liu, J., Zhang, T.: Construction and Expansion of Target's Moment Invariants. *The Transaction of Photo Electricity Technique* 2002, 123–130 (2002)
58. Mukundan, R., Ramakrishnan, K.R.: *Moment Functions in Image Analysis Theory and Application*. World Scientific Publishing Co. Pte. Ltd., Singapore (1998)
59. Vinciarelli, A.: A Survey on Off-line Cursive Word Recognition. *Pattern Recognition*, 1433–1446 (2002)
60. Madvanath, S., Govindaraju, V.: The Role of Holistic Paradigms in Handwritten Word Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 149–164 (2001)
61. Agre, G., Peev, S.: On Supervised and Unsupervised Discretization. *Cybernetics and Information Technologies* 2(2), 43–57 (2002)

62. Nguyen, H.S.: Discretization Problems for Rough Set Methods. In: Polkowski, L., Skowron, A. (eds.) *RSCTC 1998*. LNCS (LNAI), vol. 1424, pp. 545–552. Springer, Heidelberg (1998)
63. Xin, G., Xiao, Y., You, H.: Discretization of Continuous Interval-valued Attributes in Rough Set Theory and Application. In: *International Conference on Machine Learning and Cybernetics*, pp. 3682–3686 (2007)
64. Liu, H., Hussain, F., Tan, C.-L., Dash, M.: Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery* 6, 292–423 (2002)
65. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. In: *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 194–202 (1995)
66. Kotsiantis, S., Kanellopoulos, D.: Discretization Techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering* 32(1), 47–58 (2006)
67. Kohavi, R., John, G.H.: Wrappers for Feature Subset Selection. *Artificial Intelligence* 97(1-2), 1–43 (1997)
68. Gadat, S., Younes, L.: A Stochastic Algorithm for Feature Selection in Pattern Recognition. *Journal of Machine Learning Research*, 509–547 (2007)
69. Saeys, Y., Inza, I., Larranaga, P.: A Review of Feature Selection Techniques in Bioinformatics. *Journal of Bioinformatics*, 2507–2517 (2007)
70. Liu, Y., Wang, G., Chen, H., Dong, H., Zhu, X., Wang, S.: An Improved Particle Swarm Optimization for Feature Selection. *Journal of Bionic Engineering* 8(2), 191–200 (2011)
71. Unler, A., Murat, A.: A Discrete Particle Swarm Optimization Method for Feature Selection in Binary Classification Problems. *European Journal of Operational Research* 206, 528–539 (2010)
72. Deriche, M.: Feature Selection Using Ant Colony Optimization. In: *6th International Multi-Conference on Systems, Signals and Devices*, pp. 1–4 (2009)
73. Zhuo, L., Zheng, J., Wang, F., Li, X., Ai, B., Qian, J.: A Genetic Algorithm Based Wrapper Feature Selection Method for Classification of Hyperspectral Images Using Support Vector Machine. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 37, 397–402 (2008)
74. Subbotin, S., Oleynik, A.: Modifications of Ant Colony Optimization Method for Feature Selection. In: *9th International Conference - The Experience of Designing and Applications of CAD Systems in Microelectronics*, pp. 493–494 (2007)
75. Wang, X., Yang, J., Teng, X., Xia, W., Jensen, R.: Feature Selection based on Rough Sets and Particle Swarm Optimization. *Pattern Recognition Letters* 28(4), 459–471 (2007)
76. Ververidis, D., Kotropoulos, C.: Fast and Accurate Sequential Floating Forward Feature Selection with the Bayes Classifier Applied to Speech Emotion Recognition. *Signal Processing* 88(12), 2956–2970 (2008)
77. Xie, J., Xie, W., Wang, C., Gao, X.: A Novel Hybrid Feature Selection Method Based on IFSFFS and SVM for the Diagnosis of Erythematous-Squamous Diseases. In: *Workshop on Applications of Pattern Analysis*, pp. 142–151 (2010)
78. Chuang, L.-Y., Chang, H.-W., Tu, C.-J., Yang, C.-H.: Improved Binary PSO for Feature Selection using Gene Expression Data. *Computational Biology and Chemistry* 32(1), 29–38 (2008)
79. Deisy, C., Subbulakshmi, B., Baskar, S., Ramaraj, N.: Efficient Dimensionality Reduction Approaches for Feature Selection. In: *International Conference on Computational Intelligence and Multimedia Applications*, pp. 121–127 (2007)

80. Pizzi, N.J., Pedtycz, W.: Classification of Magnetic Resonance Spectra using Parallel Randomized Feature Selection. In: Proceedings of 2004 IEEE International Joint Conference on Neural Networks, pp. 2455–2459 (2004)
81. Melab, N., Cahon, S., Talbi, E.-G.: Parallel GA-based Wrapper Feature Selection for Spectroscopic Data Mining. In: Proceedings of International Parallel and Distributed Processing Symposium, pp. 201–208 (2002)
82. Qian, X.-J., Xu, J.-B.: Optimization and Implementation of Sorting Algorithm Based on Multi-core and Multi-thread. In: IEEE 3rd International Conference on Communication Software and Networks, pp. 29–32 (2011)
83. Cruz, C., Pelta, D.A., Royo, A.S., Verdegay, J.L.: Soft Computing and Cooperative Strategies for Optimization. In: IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications 2005, pp. 75–78 (2005)
84. Pudil, P., Novovicova, J., Kittler, J.: Floating Search Methods in Feature Selection. *Pattern Recognition Letters* 15, 1119–1125 (1994)
85. Whitney, A.W.: A Direct Method of Nonparametric Measurement Selection. *IEEE Transaction in Computational*, 1100–1103 (1971)
86. Eberhart, R.C., Kennedy, J.: A New Optimizer using Particle Swarm Theory. In: Proceedings of 6th International Symposium on Micro Machine and Human Science, pp. 39–43 (1995)
87. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of IEEE International Conference on Neural Networks 1995, pp. 1942–1948 (1995)
88. AbdI, K.M., Mohd Hashim, S.Z.: Swarm-Based Feature Selection for Handwriting Identification. *Journal of Computer Science* 6(1), 80–86 (2010)
89. Pratama, S.F., Muda, A.K., Choo, Y.-H., Muda, N.A.: A Comparative Study of Feature Selection Methods for Authorship Invarianceness in Writer Identification. *International Journal of Computer Information Systems and Industrial Management Applications* 4, 467–476 (2012)
90. Pratama, S.F., Muda, A.K., Choo, Y.-H., Muda, N.A.: PSO and Computationally Inexpensive Sequential Forward Floating Selection in Acquiring Significant Features for Handwritten Authorship. In: 11th International Conference on Hybrid Intelligent Systems, Melaka, Malaysia, pp. 358–363 (2011)
91. Pratama, S.F., Muda, A.K., Choo, Y.-H., Muda, N.A.: Computationally Inexpensive Sequential Forward Floating Selection for Acquiring Significant Features for Authorship Invarianceness in Writer Identification. *International Journal of New Computer Architectures and Their Applications* 1(3), 581–598 (2011)
92. Pratama, S.F., Muda, A.K., Choo, Y.-H., Muda, N.A.: SOCIFS Feature Selection Framework for Handwritten Authorship. *International Journal of Hybrid Intelligent Systems* 10(2), 83–91 (2013), doi:10.3233/HIS-130167
93. Muda, A.K., Shamsuddin, S.M., Darus, M.: Invariants Discretization for Individuality Representation in Handwritten Authorship. In: Srihari, S.N., Franke, K. (eds.) *IWCF 2008*. LNCS, vol. 5158, pp. 218–228. Springer, Heidelberg (2008)
94. Pineda-Bautista, B.B., Carrasco-Ochoa, J.A., Martinez-Trinidad, J.F.: General Framework for Class-Specific Feature Selection. *Expert Systems with Applications* 38, 10018–10024 (2011)
95. Marti, U., Bunke, H.: The IAM-database: an English Sentence Database for Off-line Handwriting Recognition. *International Journal on Document Analysis and Recognition* 5, 39–46 (2002)

# Data Mining Methods Applied to a Digital Forensics Task for Supervised Machine Learning

Antonio J. Tallón-Ballesteros and José C. Riquelme

Department of Languages and Computer Systems, University of Seville  
Reina Mercedes Avenue, Seville, 41012 Spain  
atallon@us.es

**Abstract.** Digital forensics research includes several stages. Once we have collected the data the last goal is to obtain a model in order to predict the output with unseen data. We focus on supervised machine learning techniques. This chapter performs an experimental study on a forensics data task for multi-class classification including several types of methods such as decision trees, bayes classifiers, based on rules, artificial neural networks and based on nearest neighbors. The classifiers have been evaluated with two performance measures: accuracy and Cohen's kappa. The followed experimental design has been a 4-fold cross validation with thirty repetitions for non-deterministic algorithms in order to obtain reliable results, averaging the results from 120 runs. A statistical analysis has been conducted in order to compare each pair of algorithms by means of t-tests using both the accuracy and Cohen's kappa metrics.

**Keywords:** Digital forensics, Glass evidence, Data mining, Supervised machine learning, Classification model.

## 1 Introduction

Forensic science can be defined as the application of the science to matters of the law. A fundamental principle of forensic science is that a criminal act, or more generally a human-initiated event, produces a record of itself. The record, however imperfect, is the results of human actor(s) and the events they set in motion producing interactions that result in changes in the environment. Object get moved or broken, marks are made, and materials are changed or transferred [1]. Forensic analysis is usually performed through experiments in lab which is expensive both in cost and time. Nowadays, data availability is increasing and the computational intelligence [2] techniques are very important in order to do automatically an accurate and fast analysis. Popescu and Farid [3] did a research about the use of statistical tools for altered photographs in the digital forensics context. Although digital forensics has been around for several decades, it is still a young science, and the body of peer-reviewed, academic literature that is essential for every science is currently relatively small, but it is growing [4]. Several kinds of evidences may be present in a forensic activity, such as fibres, paint, glass, soil, fingerprints. Depending on the types chemical tests, microscopic

techniques, molecular spectroscopy, elemental analysis, mass spectrometry, separation techniques or thermal analysis could be conducted [5].

An important area in forensic science called forensic interpretation of glass evidence is devoted to the study of several kind of glass properties (shape, structure, colour, size, thickness,...) after their breakage [6]. Forensic glass analysis tries to discriminate between several types of glasses. Moreover sometimes, a subsequent work once the glasses have been passed by an annealing process is performed. It has been applied in the case of a bi-class problem with toughened and laminated glasses [7]. Winstanley and Rydeard [8] were pioneered in talking about some annealing concepts about small glass fragments. Terry et al. [9] performed a quantitative analysis of glasses used in Australia depending on the source country.

The classification of glass fragments has been addressed with three data mining approaches in [10]. Ahmad et al. [11] worked with several samples of glass from cars or shops. The purpose of the classifier in the former case was to separate the rear, wind and side glass, and in the latter one was to distinguish heat absorbing, clear, reflective and figured floats. In the context of glass microtraces, Zadora et al. [12] proposed a quantitative elemental analysis using a scanning electron microscope with an energy dispersive X-ray spectrometer (SEM-EDX) in order to achieve a classification scheme for samples collected in Poland. Float glass samples of relevant cars in New Zealand using laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS) have been analyzed in [13]. UzKent et al. [14] have developed a system for classifying the sound produced by the glass breaking.

## 2 The Problem and the Data Set

Glass may be scattered in several locations and can be produced in a wide variety of forms and compositions, and these affect the properties of this material [15]. It can occur as evidence when it is broken during the commission of a crime. Broken glass fragments ranging in size from large pieces to tiny shards may be transferred to and retained by nearby persons or objects. The mere presence of fragments of glass on the clothing of an alleged burglar in a case involving entry through a broken window may be significant evidence if fragments are found. The significance of such evidence will be enhanced if the fragments are determined to be indistinguishable in all measured properties from the broken window. On the other hand, if the recovered fragments differ in their measured properties from the glass from the broken window, then that window can be eliminated as a possible source of the glass on the subject's clothing [16].

Our digital forensics problem is to forecast the type of class on basis of the chemical analysis. The study of classification of types of glass was motivated by criminological investigation. Their data set is named Glass Identification [17], whose data come from USA Forensic Science Service. It was created by B. German and was donated by V. Spiehler to UCI (University of California, Irvine) repository [18]. Instances belong to one of the types of glass, defined in terms of their oxide content (i.e. Na, Fe, K, etc). Now, we proceed to describe the semantics of the features and the class label.

- Attribute 1. Id number: 1 to 214.
- Attribute 2. RI: refractive index.
- Attribute 3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10).
- Attribute 4. Mg: Magnesium.
- Attribute 5. Al: Aluminum.
- Attribute 6. Si: Silicon.
- Attribute 7. K: Potassium.
- Attribute 8. Ca: Calcium.
- Attribute 9. Ba: Barium.
- Attribute 10. Fe: Iron.
- Class label. Type of glass. There are seven possible values:
  - Building\_windows\_float\_processed (value 1).
  - Building\_windows\_non\_float\_processed (value 2).
  - Vehicle\_windows\_float\_processed (value 3).
  - Vehicle\_windows\_non\_float\_processed (value 4). However, there are no instances containing this glass type.
  - Containers (value 5).
  - Tableware (value 6).
  - Headlamps (value 7).

We have deleted the information related with the identifier of the instances and we have considered six possible output values. Table 1 summarizes the main properties of the data set taking into account the previous remarks and Table 2 depicts the values of the statistics related with glass identification features.

This problem has been used in several works. V. Spiehler experienced with a binary classification problem for the determination whether the glass was a type of "float" glass or not. She conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm and discriminant analysis. Also, Buscema [19] has tested the Glass Identification problem in a binary form with four classifiers. Previously, Parvin et al. [20] introduced an ensemble approach and tested it with the 6-class glass problem. The multi-class version of this digital forensics task is very complex since it is difficult to classify, as literature have reported, with a high accuracy and thus this is the motivation of this chapter.

**Table 1.** Summary of the digital forensics problem

Patterns	Attributes	Numeric Attributes	Domain and Type	Nominal Attributes	Classes
214	9	9	Real (continuous)	0	6



**Table 2.** Problem statistics

Attribute	Mean	SD	Min	Max	Correlation with class
2. RI	1.5184	0.0030	1.5112	1.5339	-0.1642
3. Na	13.4079	0.8166	10.73	17.38	0.5030
4. Mg	2.6845	1.4424	0	4.49	-0.7447
5. Al	1.4449	0.4993	0.29	3.5	0.5988
6. Si	72.6509	0.7745	69.81	75.41	0.1515
7. K	0.4971	0.6522	0	6.21	-0.0100
8. Ca	8.9570	1.4232	5.43	16.19	0.0007
9. Ba	0.1750	0.4972	0	3.15	0.5751
10. Fe	0.0570	0.0974	0	0.51	-0.1879

*SD standard deviation.*

### 3 The Algorithms

Classifiers can be divided in several types [21-22]:

- **Decision trees.** A possible definition of a decision tree is a simple structure based on a tree that can be used as a classifier. Each non-leaf or internal node is associated with a decision and the leaf nodes are generally associated with an outcome or class label. Each internal node tests one or more attribute values leading two or more links or branches. Each link in turn is associated with a possible value of the decision. These links are mutually distinct and collectively exhaustive. This means that it is possible to follow only one of the links and all possibilities will be taken care of—there is a link for each possibility. The interested reader is referred to Murthy's paper [23] for a deep review. We have used two representative methods like C4.5 [24] and CART [25] that stands for Classification and Regression Tree.
- **Bayes classifiers.** In pattern recognition, Bayes classifier [26] is popular because it is an optimal classifier. It is possible to show that the resultant classification minimizes the average probability of error. Bayes classifier is based on the assumption that information about classes in the form of prior probabilities and distributions of patterns in the class are known. It employs the posterior probabilities to assign the class label to a test pattern; a pattern is assigned the label of the class that has the maximum posterior probability. The classifier employs Bayes theorem to convert the prior probability into posterior probability based on the pattern to be classified, using the likelihood values. We have used a Bayesian network (BayesNet) which is a probabilistic graphical model that represents a set of variables and their probabilistic dependencies. Formally, Bayesian networks are directed acyclic graphs whose nodes represent variables, and whose arcs encode conditional dependencies between the variables. There are efficient algorithms that perform inference and learning in Bayesian networks [22].
- **Rule-Based classifiers.** Also named rule induction classifiers. The learned model is represented as a set of IF-THEN rules. Rules are a good way of representing

information or bits of knowledge [27]. In problems where classes can be characterized by general relationships, rather than just by examples (instances). It becomes attractive to build classifiers based on rules. Humans generally like explanations for most decisions. Rules, one at a time, can be directly learned from the data that is called rule induction. Each rule is a combination of conditions [22]. As an example of this classifier type we have used RIPPER (Repeated Incremental Pruning to Produce Error Reduction) [28].

- Artificial neural networks. The field of neural networks has arisen from diverse sources, ranging from the fascination of mankind with understanding and emulating the human brain, to broader issues of copying human abilities such as speech and the use of language, to the practical commercial, scientific, and engineering disciplines of pattern recognition, modeling, and prediction [29]. There are several approaches: feed-forward and recurrent neural networks [30]. We have used the feed-forward type including two well-known approaches like the Multi-Layer Perceptron (MLP) neural network [31] with a back-propagation algorithm and the Radial Basis Function (RBF) neural network [32].
- Classifiers based on nearest neighbours. One of the simplest decision procedures that can be used for classification is the nearest neighbour (NN) rule [33-34]. It classifies a sample based on the category of its nearest neighbour. When large samples are involved, it can be shown that this rule has a probability of error which is less than twice the optimum error—hence there is less than twice the probability of error compared to any other decision rule. The nearest neighbour based classifiers use some or all the patterns available in the training set to classify a test pattern. These classifiers essentially involve finding the similarity between the test pattern and every pattern in the training set. The nearest neighbour (1-NN) algorithm assigns to a test pattern the class label of its closest neighbor. We have used three 1-NN variants that differ in the distance function that compute the dissimilarity measure or distance. Euclidean, Manhattan and Chebyshev (also written as Tchebyshev) distance [35] measures have been tested in the current chapter. The resulting classifiers are called Classical 1-NN (sometimes referred as Euclidean 1-NN), Manhattan 1-NN and Chebyshev 1-NN. The first one is very common in machine learning community. Recently, Boularias and Chaib-draa [36] have compared Euclidean and Manhattan distances in the k-NN algorithm for apprenticeship learning. In other research, related with granular data modeling, Chebyshev and Euclidean distances have been used [37].

## 4 Experimentation

### 4.1 Validation Technique

The experimental design used in this chapter has been a stratified four-fold cross validation [38]. The primary idea of the four-fold cross validation procedure is to divide the full data set in four partitions of the same size; each one is used as a test set and the remaining are used as a train set. The stratification subjects to that the partitions

maintain the class distribution of the samples approximately equal as in the original data set [39]. Stochastic algorithms have been run thirty times and since we have four folds, the results are averaged by one hundred and twenty runs in order to obtain reliable results.

## 4.2 Performance Measures

There are several measures for assessing the models obtained by the classifiers [40]. We have gathered the following performance measures:

- Accuracy. Generally speaking, the accuracy of a classifier is the probability of correctly classifying a randomly selected instance [39]. It is also known as the number of successful hits [41]. Mathematically, the accuracy is given by:

$$Accuracy = \frac{\sum_{n=1}^N I(C(\mathbf{x}_n) = y_n)}{N} * 100 \quad (1)$$

where  $I(g)$  is a function that returns 1 if  $g$  is true and 0 otherwise,  $C(\mathbf{x}_n)$  the class label assigned to the  $\mathbf{x}_n$  pattern by the classifier and  $N$  the total number of patterns.

- Cohen's kappa. It is an interesting alternative measure to the accuracy, since it compensates for random hits [42]. It was first introduced as a measure of agreement between observers of psychological behavior. The original intent of Cohen's kappa was to measure the degree of agreement, or disagreement, between two people observing the same phenomenon. The range of Kappa values extends from positive to negative one, with positive one indicating strong agreement, negative one indicating strong disagreement, and zero indicating chance-level agreement. In order to illustrate, the calculation of Cohen's kappa from the confusion matrix we will take as a starting point a 3-class problem which confusion matrix including marginal values is shown in Table 3.

**Table 3.** Confusion matrix for a 3-class problem

		Predicted class			Total
		<i>C1</i>	<i>C2</i>	<i>C3</i>	
Correct class	<i>C1</i>	<i>a</i>	<i>b</i>	<i>c</i>	$a+b+c=C1_{corr}$
	<i>C2</i>	<i>d</i>	<i>e</i>	<i>f</i>	$d+e+f=C2_{corr}$
	<i>C3</i>	<i>g</i>	<i>h</i>	<i>i</i>	$g+h+i=C3_{corr}$
Total		$a+d+g=C1_{pred}$	$b+e+h=C2_{pred}$	$c+f+i=C3_{pred}$	<i>N</i>

Being  $N$  the total number of patterns,  $C1$ ,  $C2$  and  $C3$  the label related with class 1, 2 or 3, respectively. Their Cohen's kappa is given by:

$$Cohen's\ kappa = \frac{N * (a + e + i) - (C1_{corr} * C1_{pred} + C2_{corr} * C2_{pred} + C3_{corr} * C3_{pred})}{N^2 - (C1_{corr} * C1_{pred} + C2_{corr} * C2_{pred} + C3_{corr} * C3_{pred})} \quad (2)$$

It can be generalized to the  $m$  classes

$$Cohen's\ kappa = \frac{N \sum_{i=1}^m CM_{ii} - \sum_{i=1}^m C_{i_{corr}} C_{i_{pred}}}{N^2 - \sum_{i=1}^m C_{i_{corr}} C_{i_{pred}}} \quad (3)$$

where  $CM_{ii}$  represent the diagonal elements of the confusion matrix.

Next, we will compute both performance measures for a numeric example taken from [40] that is depicted in Table 4. The accuracy and Cohen's kappa of the confusion matrix example for the 3-class problem is as follows.

$$Accuracy = \frac{15+15+45}{100} * 100 = 75\%$$

$$Cohen's\ kappa = \frac{N \sum_{i=1}^m CM_{ii} - \sum_{i=1}^m C_{i_{corr}} C_{i_{pred}}}{N^2 - \sum_{i=1}^m C_{i_{corr}} C_{i_{pred}}} = \frac{100 \sum (15+15+45) - \sum (20 * 24 + 30 * 20 + 50 * 56)}{100^2 - \sum (20 * 24 + 30 * 20 + 56 * 40)} = 0.5915$$

**Table 4.** Confusion matrix example for a 3-class problem

		Predicted class			Total
		C1	C2	C3	
Correct class	C1	15	2	3	20
	C2	7	15	8	30
	C3	2	3	45	50
Total		24	20	56	100

Source: [40].

The Cohen's kappa value is greater than 0 (random classification) and more close to 1 (perfect classification), that indicates some classification errors. The performance is good, but it can be improved for instance by correctly classifying more samples of the class number 2.

### 4.3 Algorithm Implementation and Parameters

For the experimentation we have used the implementations of the algorithms described in Section 3 that are included in framework WEKA (Waikato Environment for Knowledge Analysis) version 3.7.4 [43], with the exceptions of CART and RBF taken from the version 3.5.7. We have tested the methods related with different supervised machine learning approaches such as decision trees, bayes classifiers, rule-based classifiers, artificial neural networks and classifiers based on nearest neighbours. More specifically, we have carried out experiments with the following nine algorithms: C4.5 (J48), CART (SimpleCart), BayesNet, RIPPER (JRip), MLP with a back-propagation method, RBF, Euclidean 1-NN, Manhattan 1-NN and Chebyshev 1-NN. Regarding the parameters, in the first experiment the algorithms have been run with the default values which are according to the recommendations of their own authors. In addition, these values have been used by us in some previous studies and showed a robust behavior [44]. In the second experiment we have reported the results with fined-tuned parameter values that are described in the next section.

### 4.4 Statistical Tests

A statistical analysis has been performed in order to find out significant differences between the results obtained by the stochastic algorithms that we have dealt with. For the non-stochastic algorithms it is not possible to carry out the analysis because we have only one result per fold and the number of freedom degrees would be low for it. Since we have one problem and several stochastic algorithms we have performed a paired t-test for comparing the algorithms two by two [45]. More specifically, we have done a two-tailed t-test at a significant level of 0.05. Let  $\mu_1$  be the mean performance of the first algorithm and  $\mu_2$  be the mean performance of the second algorithm, and  $\mu_d = \mu_1 - \mu_2$ , the hypotheses are the following:

- $H_0 : \mu_d = 0$ . There is no difference in the mean performances of the two algorithms.
- $H_1 : \mu_d > 0$ . The first algorithm seems to work better.

The t statistic is computed. For the t value we will obtain the tail area (p-value) from the t-distribution table with a number of number of freedom degrees equal to the sample size minus one of the repetitions performed by each algorithm (in our case  $120-1=119$ ). If the p-value is lower than 0.05 we reject the null hypothesis concluding that there significant differences and the first algorithm is significantly better to the second one.

Statistical analysis have been conducted for the both performances measures reported in this chapter in order to extract general conclusions about which are the most stochastic appropriate algorithms for the digital forensics problem.

## 5 Results

This section is structured in two subsections. The first one is devoted to report the results with the default parameter values that were proposed by their authors due they are robust in general terms; in addition we have included a statistical comparison in order to obtain an overview if there are significant differences between stochastic algorithms. The second one show the results with fined-tuned parameter values of the algorithms by means of a grid search using the training set of each fold; since trials with a different range of the parameters could conduct to other ordering of the algorithms we have not performed any kind of statistical test. In both subsections we have divided the results in two parts: one for non-stochastic algorithms and another for the stochastic ones. The accuracy and Cohen's kappa measures have been reported for each algorithm regarding to training and test phases.

### 5.1 Results with Default Parameter Values

Table 5 shows the results obtained with the default parameters. Taking into account the non-stochastic algorithms, the best one is Manhattan 1-NN with a test accuracy over 70% and a test Cohen's kappa very close to 0.6. The second best is Euclidean 1-NN with differences of approximately 0.5 for accuracy and 0.01 for Cohen's kappa. The next best algorithms are BayesNet, C4.5 and Chebyshev 1-NN. The best stochastic algorithms ordered by decreasing performance for both measures are CART and RIPPER. The followers are MLP and/or RBF algorithms, depending on the evaluation measure. Statistical test will let to refine these remarks by means of a two-tailed t-test for each pair of algorithms. We do not established a direct comparison between non-stochastic and stochastic algorithms due to the different number of iteration for each kind of method.

The best results published recently in the paper authored by Silva and Hruscka [46] using the same data set are similar (70% with k-NN instead of 1-NN) although there are important differences in the kind of cross validation (ten-fold versus four-fold) and that work does not contain any statistical analysis for the aforementioned problem. They have reported the mean accuracy without including neither the SD nor the Cohen's kappa measure, thus it is not possible to comment some issues about the homogeneity of the solutions or to get an overview about the global classifier performance for the different labels of the instances. Two years ago, Wang et al. [47] presented a study about the performance of extreme learning machine (ELM) and introduced a new proposal called effective ELM (EELM). Their experimental designed was performed by a hold-out getting an accuracy test (from 0 to 1) with ELM and EELM over 0.42 averaged by fifty trials.

Now, we present the statistical analysis results for the stochastic algorithms. We have done two independent kinds of tests: one for accuracy and another for Cohen's kappa that are reported in Tables 6 and 7.

**Table 5.** Training and test results with the accuracy and Cohen’s kappa measures for 6-class glass identification problem

Algorithm type	Classifier approach	Method	Accuracy (%)		Cohen’s kappa	
			Training	Test	Training	Test
Non-stochastic	Decision Tree	C4.5	90.50±1.59	68.00±8.33	0.8700±0.0214	0.5663±0.1005
	Bayes	BayesNet	*	69.59±7.50	*	0.5830±0.0974
	Nearest neighbour	Euclidean 1-NN	100.00±0.00	<i>69.64±7.84</i>	1.0000±0.0000	<i>0.5867±0.1062</i>
	Nearest neighbour	Manhattan 1-NN	100.00±0.00	<b>70.13±6.85</b>	1.0000±0.0000	<b>0.5949±0.0973</b>
	Nearest neighbour	Chebyshev 1-NN	100.00±0.00	65.04±6.18	1.0000±0.0000	0.5222±0.0849
Stochastic	Decision Tree	CART	80.99±4.39	<b>67.87±2.39</b>	0.7358±0.0621	<b>0.5541±0.0343</b>
	Rules	RIPPER	80.63±4.65	<i>66.26±6.03</i>	0.7327±0.0648	<i>0.5290±0.0837</i>
	ANN	MLP	82.44±2.57	65.47±5.73	0.7557±0.0365	0.5180±0.0805
	ANN	RBF	79.15±2.78	65.27±8.33	0.7170±0.0383	0.5259±0.1118

**Best** and *second best* test results depending on the algorithm type have been highlighted in **boldface** and *italics*, respectively.

\* Training results not provided by the classifier implementation.

**Table 6.** Statistical analysis with a two-tailed t-test for accuracy measure in the 6-class glass identification problem

Two-tailed t-test for accuracy				
Algorithm 1	Algorithm 2	p-value	t(119) statistic	Statistical test conclusion
CART	RIPPER	0.0099 *	2.6226	<b>CART &gt; RIPPER</b>
CART	MLP	8.868*10 <sup>-5</sup> *	4.0586	<b>CART &gt; MLP</b>
CART	RBF	0.0023 *	3.1205	<b>CART &gt; RBF</b>
RIPPER	MLP	0.2780	1.0899	RIPPER ≥ MLP
RIPPER	RBF	0.2200	1.2331	RIPPER ≥ RBF
MLP	RBF	0.7536	0.3146	MLP ≥ RBF

Overall accuracy ranking:  $\mu_{\text{Accuracy(CART)}} > \mu_{\text{Accuracy(RIPPER)}} \geq \mu_{\text{Accuracy(MLP)}} \geq \mu_{\text{Accuracy(RBF)}} .$

\* : Significant difference at  $\alpha = 0.05$  .

According to the statistical test results for accuracy, we can assert that CART is the algorithm with a performance significantly better than the remaining algorithms. The second best algorithm is RIPPER but the differences with their competitors are not enough to be significant. Thus the best classifier belongs to decision tree approach and the next best to rules. Comparing the two models of neural networks, there are no significant differences although MLP is slightly better than RBF.

**Table 7.** Statistical analysis with a two-tailed t-test for Cohen's kappa measure in the 6-class glass identification problem

Two-tailed t-test for Cohen's kappa				
Algorithm 1	Algorithm 2	p-value	t(119) statistic	Statistical test conclusion
CART	RIPPER	0.0044 *	2.9016	<b>CART &gt; RIPPER</b>
CART	MLP	$3.4 * 10^{-5}$ *	4.3051	<b>CART &gt; MLP</b>
CART	RBF	0.0145 *	2.4813	<b>CART &gt; RBF</b>
RIPPER	MLP	0.2827	1.0791	RIPPER $\geq$ MLP
RIPPER	RBF	0.7820	0.2774	RIPPER $\geq$ RBF
MLP	RBF	0.3678	0.9041	RBF $\geq$ MLP

Overall Cohen's kappa ranking:

$$\mu_{\text{Cohen's kappa(CART)}} > \mu_{\text{Cohen's kappa(RIPPER)}} \geq \mu_{\text{Cohen's kappa(RBF)}} \geq \mu_{\text{Cohen's kappa(MLP)}} .$$

\* : Significant difference at  $\alpha = 0.05$ .

For Cohen's kappa, statistical test indicates that CART is significantly the best algorithm. The second best one is RIPPER that is quantitatively better than MLP and RBF. The last one neural network model is slightly better than MLP without significant differences.

## 5.2 Results with Fine-Tuned Parameter Values

First of all, we introduce the parameter values that we have defined for the fine setting by means of a grid search with the training set of each fold. For the 1-NN algorithm it is not possible to use specific parameters with the exception of the distance function that we have considered in the previous subsection. Table 8 presents the possible values or range of the parameters that we have selected for the fine tuning; the algorithms are sorted depending on the type, that is, first the non-stochastic ones and then the stochastic ones.

Table 9 reports the results of those algorithms obtained with the aforementioned fine-tuned parameters grouped by algorithm type and classifier approach. Also, we have included the results default with the default parameters, due to the reasons exposed at the beginning of this subsection, for the three variants of 1-NN in order to get a general view of the performance. In reference to the non-stochastic methods, the two best algorithms are C4.5 and Manhattan 1-NN, depending on the performance measure. From the stochastic aspect, the best classifier for both measures is CART, followed by RBF.

The fine setting of the parameters has shifted the performance ordering of the non-stochastic algorithms and has let to improve the results; the best classifier has now surpassed the top of 73.5% of accuracy and has reached a Cohen's kappa close to 0.595. In the context of stochastic methods, this tuning has increased the performance of the algorithms and has moved the name of the second best classifier; the best results for both measures are over 68% and 0.55 for accuracy and Cohen's kappa, respectively.



**Table 8.** Fine-tuned parameter values of the algorithms by means of a grid search on the training set of each fold

Algorithm	Parameter	Possible values or range	Default value	Best value
C4.5	Confidence factor (C)	{0.150, 0.175, 0.200, 0.225, 0.250}	0.25	0.175
	Minimum number of instances per leaf (M)	2-10	2	2
BayesNet	Alpha value (A) for Simple Estimator	{0.25, 0.50, 0.75}	0.50	0.75
CART	Minimal number of observations of the terminal nodes (M)	2-5	2	4
	The number of fold in the internal cross-validation	2-10	5	5
RIPPER	Folds: the amount of data used for pruning (F)	1-5	3	4
	The minimum total weight of the instances in a rule (N)	1-3	2	1
	The number of optimization runs (O)	1-3	2	3
MLP	TrainingTime: The number of epochs to train through (N)	{250, 500, 750, 1000}	500	500
	HiddenLayers: hidden layers of the neural network (H)	4-16	a = (attribs. + classes) / 2	15
RBF	NumClusters: The number of clusters for K-Means to generate (B)	1-6	2	4

**Table 9.** Training and test fine-tuned results with the accuracy and Cohen’s kappa measures for 6-class glass identification problem

Algorithm type	Classifier approach	Method	Accuracy (%)		Cohen’s kappa	
			Training	Test	Training	Test
Non-stochastic	Decision Tree	C4.5	88.71±2.31	<b>73.67±2.62</b>	0.8511±0.0334	<i>0.5929±0.1006</i>
	Bayes	BayesNet	*	70.07±6.84	*	0.5886±0.0897
	Nearest neighbour	Euclidean 1-NN	100.00±0.00	69.64±7.84	1.0000±0.0000	0.5867±0.1062
	Nearest neighbour	Manhattan 1-NN	100.00±0.00	<i>70.13±6.85</i>	1.0000±0.0000	<b>0.5949±0.0973</b>
	Nearest neighbour	Chebyshev 1-NN	100.00±0.00	65.04±6.18	1.0000±0.0000	0.5222±0.0849
Stochastic	Decision Tree	CART	77.78±2.98	<b>68.22±1.93</b>	0.6907±0.0388	<b>0.5576±0.0292</b>
	Rules	RIPPER	82.39±4.65	66.85±4.84	0.7584±0.0640	0.5397±0.0674
	ANN	MLP	86.95±2.45	66.47±3.37	0.8200±0.0339	0.5323±0.0456
	ANN	RBF	88.28±2.35	<i>66.86±4.67</i>	0.8407±0.0320	<i>0.5475±0.0633</i>

**Best** and *second best* test results with fine-tuned parameters depending on the algorithm type have been highlighted in **boldface** and *italics*, respectively.

\* Training results not provided by the classifier implementation.

## 6 Conclusions

In this chapter we have reviewed the state-of-the-art related with a digital forensics task called Glass Identification in the context of multi-class supervised learning. This problem have been tackled from some decades to the present, however the previous studies are focused on a particular issue. We have presented an empirical overview of the performance with a good number of classifiers from different machine learning approaches with two metrics like accuracy and Cohen's kappa for training and test stages, using the default parameter values in the first experiment and the fine-tuned values in the second one. We have included a statistical analysis in the first experiment that has revealed some valuable conclusions.

In the first experiment, related with the deterministic algorithms, Manhattan 1-NN obtains the best performance for accuracy and Cohen's kappa metrics. Their performance is slightly better than the Euclidean 1-NN. Our real-world problem is another sample in that nearest neighbours classifiers can be applied successfully. Thus, it has been proven that Manhattan 1-NN is better than Euclidean 1-NN, BayesNet, C4.5 and Chebyshev 1-NN. Moreover, we have reported the results of non-deterministic algorithms; however it is not possible to compare them with deterministic algorithms because the former methods have been smoothed by an average of one hundred and twenty runs versus four of the latter methods. The best non-deterministic algorithm is CART with statistically significant differences with the remaining non-deterministic methods. The second best classifier is RIPPER, however there are no significant differences with the classifiers with a lower performance. Best approaches for non-deterministic methods are, in this order, decision trees, rules and artificial neural networks. In the second experiment, the best deterministic algorithm is C4.5 or 1-NN Manhattan according to the performance evaluation measure. The best non-deterministic algorithm is CART with both measures and the second best one is the RBF neural network model.

The most important remarks taking into account both experiments are stated as follows. The fine tuning of the parameters has been very useful due to: i) From the non-stochastic algorithm perspective the best accuracy results has passed 73.5% with C4.5 classifier and are very close to 0.595 for Cohen's kappa with 1-NN Manhattan, ii) The performance of the best stochastic algorithm has reached 68.22 and 0.5576 for accuracy and Cohen's kappa, respectively. The problem tackled can be considered very difficult since, up the best of our knowledge, it is not possible, as this chapter showed, to classify the test instances with an accuracy level a 75%. A possible future research line of this chapter could try to study some pre-processing data mining techniques in order to act on the features, instances or values of the attributes.

**Acknowledgements.** This chapter has been partially subsidized by TIN2011-28956-C02-02 project of the Spanish Inter-Ministerial Commission of Science and Technology (MICYT), FEDER funds and the P11-TIC-7528 project of the "Junta de Andalucía" (Spain).

## References

1. Caddy, B.: *Forensic Examination of Glass and Paint: Analysis and Interpretation*. Taylor & Francis, London (2011)
2. Mumford, C.L., Jain, L.C. (eds.): *Computational Intelligence*. ISRL, vol. 1. Springer, Heidelberg (2009)
3. Popescu, A.C., Farid, H.: *Statistical Tools for Digital Forensics*. In: Fridrich, J. (ed.) *IH 2004*. LNCS, vol. 3200, pp. 128–147. Springer, Heidelberg (2004)
4. Kessler, G.C.: *Advancing the Science of Digital Forensics*. *Computer* 45(12), 25–27 (2012)
5. Stuart, B.H.: *Forensic Analytical Techniques*. John Wiley & Sons, West Sussex (2013)
6. Curran, J.M., Hicks, T.N., Buckleton, J.S.: *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton (2000)
7. Newton, A.W.N., Kitto, L., Buckleton, J.S.: *A study of the performance and utility of annealing in forensic glass analysis*. *Forensic Science International* 155, 119–125 (2005)
8. Winstanley, R., Rydeard, C.: *Concepts of annealing applied to small glass fragments*. *Forensic Science International* 29, 1–10 (1985)
9. Terry, K.W., van Riessen, A., Lynch, B.F., Vowles, D.J.: *Quantitative analysis of glasses used within Australia*. *Forensic Science International* 25, 19–34 (1984)
10. Zadora, G.: *Classification of Glass Fragments Based on Elemental Composition and Refractive Index*. *Journal of Forensic Science* 54(1), 49–59 (2009)
11. Ahmad, U.K., Asmuje, N.F., Ibrahim, R., Kamaruzamanc, N.U.: *Forensic Classification of Glass Employing Refractive Index Measurement*. *Malaysian Journal of Forensic Sciences* 3(1), 1–4 (2012)
12. Zadora, G., Brozek-Mucha, Z., Parczewski, A.: *A classification of glass microtraces*. *Problems of Forensic Sciences XLVII*, 137–143 (2001)
13. Grainger, M.N.C., Manley-Harris, M., Coulson, S.: *Classification and discrimination of automotive glass using LA-ICP-MS*. *Journal of Analytical Atomic Spectrometry* 27, 1413–1422 (2012)
14. Uz Kent, B., Barkana, B.D., Cevikalp, H.: *Non-speech environmental sound classification using SVMs with a new set of features*. *International Journal of Innovative Computing, Information and Control* 8(5B), 3511–3524 (2012)
15. Bottrell, M.C.: *Forensic Glass Comparison: Background Information Used in Data Interpretation*. *Forensic Science Communications* 11(2) (2009)
16. Koons, R.D., Buscaglia, J., Bottrell, M., Miller, E.T.: *Forensic glass comparisons*. In: Saferstein, R. (ed.) *Forensic Science Handbook*, 2nd edn., vol. I, pp. 161–213. Prentice Hall, Upper Saddle River (2002)
17. Evett, I.W., Spiehler, E.J.: *Rule induction in forensic science*. In: *Knowledge Based Systems in Government*, pp. 152–160. Halsted Press, London (1988)
18. Frank, A., Asuncion, A.: *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, Irvine, CA (2010), <http://archive.ics.uci.edu/ml>
19. Buscema, M.: *Artificial Adaptive Systems in Data Visualization: Proactive Data*. In: Buscema, M., Tastle, W. (eds.) *Intelligent Data Mining in Law Enforcement Analytics: New Neural Networks Applied to Real Problems*, pp. 51–88 (2013)
20. Parvin, H., Minaei-Bidgoli, B., Shahpar, H.: *Classifier Selection by Clustering*. In: Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ben-Youssef Brants, C., Hancock, E.R. (eds.) *MCPR 2011*. LNCS, vol. 6718, pp. 60–66. Springer, Heidelberg (2011)

21. Murty, M.N., Devi, V.S.: Pattern Recognition. An Algorithmic Approach. Universities Press (India), Pvt. Ltd., London (2011)
22. Dougherty, G.: Pattern Recognition and Classification: An Introduction. Springer, New York (2013)
23. Murthy, S.K.: Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery* 2, 345–389 (1998)
24. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)
25. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth Int. Group, Belmont (1984)
26. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers, San Francisco (1998)
27. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Waltham (2011)
28. Cohen, W.: Fast effective rule induction. In: Proc. of the 12th Int. ICML Conf., pp. 115–123 (1995)
29. Michie, D., Spiegelhalter, D.J.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York (1994)
30. Haykin, S.O.: Neural Networks and Learning Machines. Prentice Hall, Upper Saddle River (2009)
31. Bishop, M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
32. Howlett, R.J., Jain, L.C.: Radial Basis Function Networks 1: Recent Developments in Theory and Applications. Springer, Heidelberg (2001)
33. Fix, E., Hodges, J.: Discriminatory analysis, nonparametric discrimination: consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine, Randolph Field, Texas (1951)
34. Cover, T., Hart, P.: Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1), 21–27 (1967)
35. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Boston (2005)
36. Boularias, A., Chaib-draa, B.: Apprenticeship learning with few examples. *Neurocomputing* 104, 83–96 (2013)
37. Bargiela, A., Pedrycz, W.: A model of granular data: a design problem with the Tchebyshev FCM. *Soft Computing* 9(3), 155–163 (2005)
38. Hjorth, J.S.U.: Computer intensive statistical methods: Validation model selection and bootstrap. Chapman and Hall, London (1994)
39. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI 1995), Montreal, Quebec, Canada, vol. 2, pp. 1137–1145 (1995)
40. Flach, P.: Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, United Kingdom (2012)
41. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, USA (2011)
42. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46 (1960)

43. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* 11(1), 10–18 (2009)
44. Tallón-Ballesteros, A.J., Hervás-Martínez, C., Riquelme, J.C., Ruiz, R.: Feature selection to enhance a two-stage evolutionary algorithm in product unit neural networks for complex classification problems. *Neurocomputing* 114, 107–117 (2013)
45. Nisbet, R., Elder, J.F., Miner, G.: *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, Canada (2009)
46. Silva, J.A., Hruschka, E.R.: An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks. *Data & Knowledge Engineering* 84, 47–58 (2013)
47. Wang, Y., Cao, F., Yuan, Y.: A study on effectiveness of extreme learning machine. *Neurocomputing* 74, 2483–2490 (2011)

# Speech Quality Enhancement in Digital Forensic Voice Analysis

Moses Ekpenyong<sup>1</sup> and Okure Obot<sup>2</sup>

<sup>1</sup> Department of Computer Science, University of Uyo, Uyo  
Centre for Speech Technology Research (CSTR),  
University of Edinburgh, Edinburgh, UK

mosesekpenyong@uniuyo.edu.ng, mosesekpenyong@gmail.com

<sup>2</sup> Department of Computer Science, University of Uyo, Uyo  
okureobot@uniuyo.edu.ng, abatakure@yahoo.com

**Abstract.** The influence of noise and reverberation in Digital Forensic voice evidence can conceal the identification, verification and processing of crime data. Computationally, the efficiency in processing speech signals largely depends on the integrity and authenticity of audio/voice recordings. Our interest is on improving integrity, vis-à-vis the intelligibility of speech signals. We achieved this in four folds. First, a speech quality enhancement technique that cleans and rebuilds defective speech data for quality Forensic analysis is proposed by exploring an optimal estimator for the magnitude spectrum, where the Discrete Fourier Transform (DFT) coefficients of clean speech are modelled by a Laplacian distribution and the noise DFT coefficients are modelled using a Gaussian distribution. Second, an automatic speech pre-processing algorithm for phoneme segmentation of raw speech data, capable of iteratively refining Hidden Markov Model (HMM) speech labels for improved intelligibility is introduced. Third, a simulation of the distortion from a quantised R-bit and computation of the Signal-to-Noise Ratio (SNR) for the signal to quantisation noise is carried out for the purpose of managing speech signal distortions. Fourth, an investigation of the effect of confused phonemic and tone bearing unit features on the intelligibility of speech is presented to assist Forensic experts decode voice disguise or language “barriers” that may impede proper Forensic voice analysis. Results obtained in this investigation reveal a future of prospects in the field of Forensic intelligence and is most likely to reduce unnecessary setbacks during Forensic analysis.

**Keywords:** Forensic science, intelligent system, speech quality evaluation, speech synthesis, voice adaptation.

## 1 Introduction

One of the most essential tools in combating crime is information. Information is required by law enforcement agencies about the nature of the crime, the time and place the crime was committed and the crime suspect(s). Recently, there has been an explosion of information such that it becomes pertinent to process information with

sophisticated tools other than the erstwhile manual and mechanical methods. Digital Forensic is becoming a potential information technology application for investigating crimes and related offences. It is the process of uncovering and interpreting electronic data for use in the law court or related arena. The main goal is to preserve any evidence in its original form while performing a structured investigation of collecting, identifying and validating encoded evidence for the purpose of reconstructing past events to aid apprehend the culprits.

The field of Digital Forensic permeates several topics (in Forensic science) including computer Forensics, peripheral/network Forensics, online/real-time Forensics, database Forensics, wireless/mobile Forensics and software Forensics. This field can be described as precisely expert-driven advances which are developed and subsequently applied [1]. Forensic investigation of digital evidence is predominantly employed as a post-incident response to an action that cannot be described certainly as legal or to an incident defying organizational standards and policies. Digital Forensic is however challenged by both technical and social issues. These include information explosion, information security techniques, intelligence of modern software tools, proliferation of mobile devices, lack of Forensic tools and experts, poor legislation and low comprehension of computing techniques by jurists.

Forensic science is however undergoing a paradigm shift, characterised by data mining and agent-based approaches [2-3] and database implementing the likelihood ratio framework [4], as well as the evaluation of the validity and reliability of results [5]. Hence, the implementation of Forensic science is most likely to

- prevent further malicious attempts against intended “targets”
- provide successful recall of past events that triggered a crime and assist in the identification and prosecution of culprits
- establish the needed mechanisms for improving and preventing unusual events from reoccurring
- enhance standards on corporate networks and security
- serve as a ‘plug-in’ to the digital environment and improve awareness in a bid to prevent future vulnerabilities

### **1.1 Gathering, Storing and Presenting Digital Evidence**

Collection and storage of evidence is done after identifying the information needed. The evidence must be stored in a manner that only unavoidable minimal alteration is made. Alteration to data that is of evidential value must be accounted for and justified [6]. For instance, a picture retrieved in a particular file format might require a slight modification before its present format could be transformed easily into a readable format. Examining the evidence is important to determine if the evidence would on conversion preserve the structure and integrity of the original copy, which prompts for the analysis of the evidence after close examination. This analysis involves processing what has been collected into an interpretable form understood by the common user. A

case of extracting an encrypted file, which requires decryption before being presented in the law court as evidence is a typical example of such analysis. Another example is where the contents of a hard drive image are processed into human understandable formats before being presented to a jurist.

In the presentation of evidence, the method and manner must render it legally acceptable and within the confines of the law. The expertise of the presenter must therefore be brought to bear. Also, the tools used in the extraction and processing of the evidence should have some legal acceptable status. For instance, a pirated copy of software used in the extraction of evidence might become a technical issue used to dismiss a case even when the evidence seems convincing. The evidence presented must as well satisfy the legal requirements as conventional evidence and must be [7]:

- *Authentic*: the evidence must be original and relate to the alleged crime under investigation
- *Reliable*: the evidence must have been collected using reliable procedures and could be repeated by an independent party where necessary to achieve the same result
- *Complete*: the evidence must be used to prove guilt as well as innocence
- *Believable*: the evidence should be convincing to juries and presented in an unambiguous way that preserves its content
- *Admissible*: the evidence should be collected using the procedures that conform to common law and legislative rules

## 1.2 Investigation Media and Image Analysis

The following media may be used to comb for evidences:

- *Data communication media* could be monitored for some unruly behaviour of users such as unauthorized access and policy violations using intrusion detection systems. Such violations or malicious activities when detected should be reported to the appropriate quarters. Evidence gathered during such operations could be properly examined, analysed for presentation in a law court.
- *Storage media* such as hard disk, removable disks, CD-ROMs, Digital Versatile Disk (DVD) etc. are easily examined and analysed with a view to gathering or recovering latent data stored. In the process, evidence of criminality could be gathered.
- *Mobile devices* including personal digital assistants (PDAs) with memory cards are also likely to store incriminating information that could form evidence in a law court.

Images on storage media could be physically or logically analysed. Logical analysis uses graphical tools like file managers and file viewers, while physical analysis is done with a physical viewpoint, not from the perspective of a file system, but using the *hex* editor, for instance. Files *hashing* may also be carried out to distinguish files



with and without evidential values. Images can also be analysed on the strengths of their format and the digital signature is one of such images. Although the format of the file extension could be changed by a culprit to hide the image identity, a forensic expert should first compare the file's extension with its corresponding file's signature and if a mismatch is found further examination is undertaken.

### 1.3 Speaker Identification and Verification

When investigating voice evidence (gathered or stored), one of the unique features used to identify a suspect is his/her voice. Conventionally, there are voice identification specialists capable of deciphering different voices and the most common method used to accomplish this is by comparing the distinctive speech features of the speakers. Speech technologies (speech synthesis, speech recognition and machine translation, etc.) have provided excellent approaches for adapting and recognising certain voices among varied speech clusters. Speech technology explores specific features of the human voice such as the formant frequency, pitch and pitch contours. But some constraints such as the speaker's mood, mimicry, environment and noise must be considered in order to obtain accuracy. Forensic experts require devices/software such as speech recognisers to assist them in obtaining specific features of the suspect's speech and verify same with what the exhibited device(s) carry. Speaker identification plays an important role in Forensic science and embodies the various tasks of discriminating persons based on the sound of their voices. It typically requires two audio recordings: a questionable recording and a voice sample. The questionable recording in most cases constitutes the intercepted or recorded communication (e.g. phone call), while the sample recording is usually taken from the suspect's original statements. Lots of approaches to speaker identification have been proposed [8-9]. Except for the linguistically-based ones (auditory, psychological, etc.), others involve spectral analysis (voiceprint, formant matching etc.).

Speaker verification then aims at accepting or rejecting an identity claim based on a voice sample [10]. Investigations on the problems of imposture as regards speaker verification have been reported over the years and methods to prevent these problems proffered in [11]. Furthermore, the vulnerability of speaker verification to voice impersonation by humans has been examined in [12-13]. However, advancements in speech technologies have presented potential challenges for both offline confirmation of acquired speech with the original speaker, and online authentication of voices [11].

Generally, to obtain high quality voice database for Forensic analysis, usable phonetic units must be annotated. These annotations are required to map the distinctive linguistic features of an utterance to its acoustic properties. The goal here is to guarantee quality and improved accuracy when comparing Forensic voices. Before the advent of state-of-the-art approaches to automatic annotation and linguistic features alignment, annotations were done manually. This made Forensic analysis of voice data laborious and difficult to accomplish. Initially, Forensic Voice Comparison (FVC) systems employed the acoustic-phonetic approach, characterised by the identification and marking of functional tokens of phonetic units (sufficiently) by trained phoneticians in both the suspect and offender recordings. The annotations were then

analysed statistical. But, a plethora of research works have proved the efficacy of automatic approaches to FVC. Specifically, recent FVC research works [14-15] present the potentials of hybrid approaches – a combination of automatic systems with acoustic-phonetic systems, to improve both validity and reliability.

#### 1.4 Forensic Intelligence

It is assumed that during an investigation some level of reasoning is employed. These can be classified as basic inferential steps which vary both in structure and levels of detail, thus, requiring a broad variety of specialised knowledge, scientific attitude as well as the integration of automated tools [16]. Within the last two decades, security strategies have drifted toward more intelligence-led and proactive frameworks [17]. Beyond the notable achievements of identification databases such as DNA or AFIS, there are indications that Forensic case data could provide intelligence [17]. This pivotal instrument is efficient for taking informed decisions at the strategic and tactical levels with the application of suitable strategies to fight crimes, or the appropriate deployment of resources for effective security. Also, criminal intelligence is now broadly implemented within law enforcement organisations through the leverage of emerging database technologies such as Geographical Information Systems (GISs), data mining, biometrics, etc.

In [17], the seven primitive inferences that form a basic framework for Forensic intelligence have been discussed. These inferences which include identity; information source: source to trace and trace to source; unanticipated “side effects”; source profile; source classification; list of possible sources; list of possible relatives (to the source) are mainly based on analogical reasoning and combines a greater part of existing approaches and databases. They do not however represent an exhaustive inventory, as certain information regarding Forensic computing and intelligence provided by mobile phone analysis and inferences relating to crime reconstruction and process modelling are missing.

In [18], the use of Hidden Markov Models (HMMs) in Digital Forensics is explored using Embedded Bayesian Network HMMs to investigate interactions between multiple suspects in Forensic cases. They compare the output of a coupled HMM to a similarly trained single-chain HMMs and to expert knowledge in a simulated digital Forensic case with two suspects. Their results demonstrate that there exists some form of interaction between suspects of a digital Forensic case and the interaction can be effectively modelled using HMMs.

In [19], a framework for an intelligent natural language interface (NLI) that suits the need of an embedded platform using agent-based approach is proposed. The architecture is based on various forms of action representations and a sequence of transformations that converts users’ inputs into suitable sets of agent-based actions that produce response to the input. The approach incrementally eliminates the complexity and ambiguity of the input by using predefined sets of interim actions at different levels, thereby, increasing the robustness and reliability of the NLI. To optimise this architecture for practical dialog systems, Ekpenyong [20] provides a tonal framework for building the synthesis component of a voice user interface (VUI) system.

The design exploits the HMM framework to achieve speaker-dependent voices for Ibibio (ISO 693-3: nic; Ethnologue: IBB), and is adaptable to other tone languages with similar structure.

## 1.5 Our Approach

This chapter focuses at enhancing the intelligibility of speech, by proposing a speech enhancement model useful for quality Forensic voice analysis. Our model is capable of rebuilding defective voices without losing much signal information (e.g., due to clipping and reverberation). We also deal with parameters that affect the perceived quality of speech using the Hidden Markov Model (HMM) which is known to yield high quality synthetic speech suitable for voice adaptation and simulation. The effectiveness of HMM-based speech synthesis approach for Forensic voice analysis is also evaluated in this chapter to guarantee comprehensibility, accuracy and precision.

## 1.6 Chapter Organisation

The rest of this chapter offers the following: (i) a discussion on speech signal degradation and assessment, (ii) a presentation of the proposed speech enhancement architecture and system model for Digital Forensic analysis and an analysis of a defective speech sample, (iii) an implementation of the proposed model and evaluation of some voices.

## 2 Speech Signal Degradation and Assessment

A speech signal comprises of different parameters which can roughly be split into three groups. The first group pertains to quantity and include parameters such as duration of the voice sample. The second group has to do with quality and include parameters such as Signal to Noise Ratio (SNR), clipping, frequency range, etc. The third group refers to comparability and the speaker's emotional state falls into this category. Speech signals are corrupted by various types of degradation. The most common types of degradation include background noise, reverberation and speech from competing speakers. Degraded speech is poor, both in terms of perceptual quality (naturalness) and ability to comprehend the perceived speech (intelligibility). Poor speech naturalness often results in listener's fatigue, while poor intelligibility ultimately degrades performance in speech technology applications. We shall categorize these degradations according to how they alter the resultant speech signal as follows:

1. *Uncorrelated additive noise* co-existing with the wanted speech signal may arise in the acoustic and electronic domain. Its perceived effect is to degrade perceptual quality and in extreme cases may completely mask the wanted signal. For some types of additive noise, the spectral characteristics are stationary and gradually transform over time. This is typically true of amplifier noise as well as of some environmental acoustic noise sources. Other forms of additive noise are intermittent or highly non-stationary; such non-stationary noise sources include media interference, unwanted co-talkers and some forms of electrical interference

2. *Convolutive effects* are perceived as reverberation and poor spectral balance because the added noise is strongly correlated with the wanted signal. Reverberation and echo normally arise from acoustic reflections and can seriously degrade intelligibility. The increasing use of distant microphones in hands-free telephony has prompted extensive research into mitigating the effects of reverberation. Bandwidth restrictions and uneven spectral response may also arise from microphone placement and characteristics as well as amplifier limitations.
3. *Non-linear distortions* frequently arise from amplitude limiting the microphone and amplifier. This is perceived as harsh distortion which varies with the signal amplitude. A similar perceptual effect can result from high bit errors in the coded signal used by some amplifiers.

The assessment of speech quality (intelligibility and naturalness) is therefore of utmost importance when proposing the deployment of speech products. For intelligibility, there is need to choose an appropriate linguistic level at which to make measurements. This problem arises from the fact that linguistic units possess increased redundancy (i.e., not all phonetic sequences form words in the language and not all word sequences form meaningful sentences). While the assessment of channel utilization to convey meanings of real spoken utterances is necessary, listeners vary widely in their perception/understanding of the speech based on their own linguistic competence. In assessing speech naturalness, the reliability of listeners becomes an issue. The mean opinion score test [21], in which listeners assign absolute ratings to individual speech stimulus and diagnostic tests [22], where listeners exhibit preference on speech stimulus over the other, are the most widely used tests. The advantage of the former is that a system can be assigned some absolute score, but this requires a large sample size of listeners to achieve satisfactory sensitivity. The advantage of the latter is that statistically significant results can be obtained by comparing two systems with relatively few listeners.

In assessing the perceived speech quality, we concentrate on improving speech intelligibility and investigate further into the results of a Modified Rhyme Test (MRT) obtained in [23]. Specifically, we compare the degree of confused phonemes (vowels and consonants) and tone bearing units of two synthesis systems (with and without tone evidence) in a *discriminant* analysis. The reason for comparing both systems is to measure their perceived intelligibility levels and suitability for FVC.

## 3 Enhancing Speech Quality for Digital Forensics

### 3.1 Phoneme Segmentation Algorithm

Initially, speech annotation required skilful phoneticians and was laborious. Now, there exist numerous automatic segmentation approaches to speech annotation. The approach (used in this chapter) for phoneme segmentation or speech labelling of the HMMs employs machine learning techniques for optimised refinements, and relies on cheap features of the intended language. The algorithm is succinctly described as

follows: (i) map the utterance text of a given language to the respective sound files, (ii) allocate each phone of the utterance mapped same duration. This initial assumption is used to pre-annotate the speech corpus and can be achieved by dividing the total utterance's duration by the total number of phonemes in that utterance. This gives each phone an equal range, with a fixed incremental factor. The resultant voice from this annotation sounds poor, but the quality was improved through a re-alignment process implementing the Viterbi algorithm. Viterbi algorithm is the most common algorithm for implementing n-gram search. It is an efficient dynamic search technique that avoids the polynomial expansion of a breath first search, by 'trimming' the search tree at each level using the best  $n$  Maximum Likelihood Estimates (MLEs). The realignment uses a HMM edit (HLEd) tool of the HTS toolkit and can be performed until the desired intelligibility is obtained.

### 3.2 System Architecture

The system architecture adopted in this chapter for enhancing defective speech is shown in Fig. 1. The noisy signal  $n(s)$  is sampled at regular time intervals and comprises of the clean speech  $c(s)$  and additive noise  $g(s)$ . The DFT coefficient of the frame and frequency bin is computed in the Fast Fourier Transform (FFT) block after widowing, to mitigate the disturbing effects of cyclic convolution. The SNR estimation block computes apriori SNR  $\xi$  and posteriori SNR  $V$  for each DFT bin  $k$ . The task of the speech estimation block is to compute the *spectral weights* for the noisy spectral components  $N$ , such that the estimated DFT coefficient  $\hat{c}$  is derived. Thereafter, an Inverse Fast Fourier Transform (IFFT) and overlap-add method application produces the enhanced (desired) signal  $\hat{c}(s)$ .

### 3.3 System Model

A noisy signal can be expressed as the sum of the clean speech signal and the additive noise, and is represented mathematically as:

$$n(s) = c(s) + g(s) \quad (1)$$

Taking the Fourier Transform of  $n(s)$ , we obtain

$$N(\gamma_k) = C(\gamma_k) + G(\gamma_k) \quad (2)$$

where

$$\gamma_k = \frac{2\pi k}{L} \{k = 1, 2, \dots, L-1\}$$

and  $L$  is the frame length.

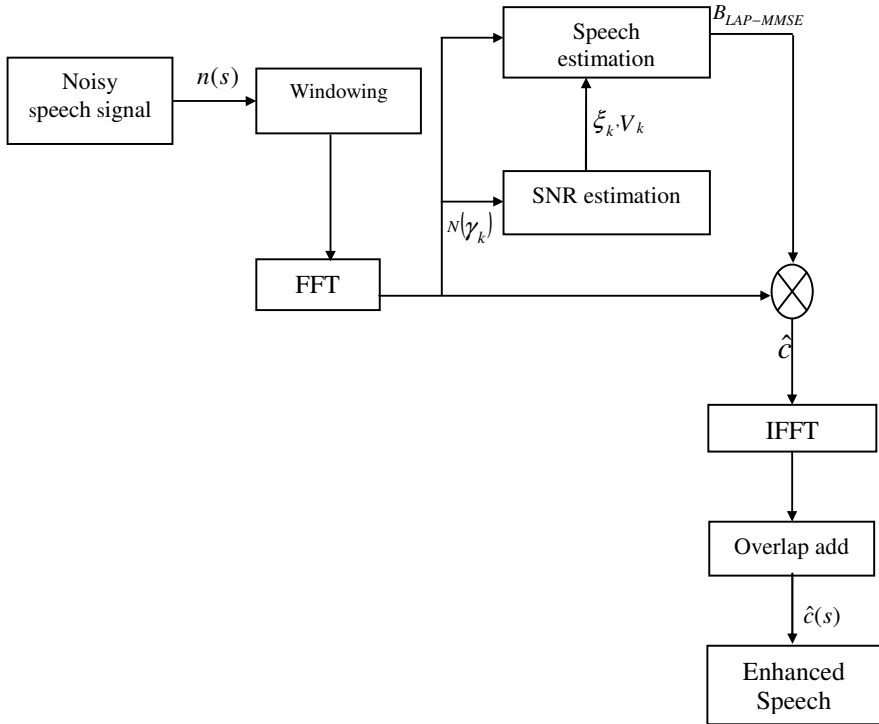


Fig. 1. Speech enhancement system architecture

Equation (2) can also be expressed in the form:

$$N_k e^{i\theta_n(k)} = C_k e^{i\theta_c(k)} + G_k e^{i\theta_g(k)} \tag{3}$$

where,  $N_k$ ,  $C_k$  and  $G_k$  are magnitude spectra corresponding to the noisy, clean and additive noise signals, respectively.  $\theta_n(k)$ ,  $\theta_c(k)$  and  $\theta_g(k)$  are the phase spectra corresponding to the noisy, clean and additive noise signals. Thus, the MMSE estimator for the magnitude spectrum  $C_k$  can be obtained as [24]:

$$\begin{aligned} \hat{C}_k &= E\{C_k | N(\gamma_k)\}, \quad k = 0,1,2,\dots,L-1 \\ &= \frac{\int_0^\infty \int_0^{2\pi} C_k f(N(\gamma_k) | C_k, \theta_k) f(C_k, \theta_k) d\theta_k dC_k}{\int_0^\infty \int_0^{2\pi} f(N(\gamma_k) | C_k, \theta_k) f(C_k, \theta_k) d\theta_k dC_k} \end{aligned} \tag{4}$$

where

$E\{\cdot\}$  denotes the expectation operator,  $\theta_k \overset{\Delta}{=} \theta_n(k)$

$f(C_k, \theta_k)$  is the joint probability density function of the magnitude and phase spectra

$f(N(\gamma_k) | C_k, \theta_k)$  is given as [24]:

$$f(N(\gamma_k) | C_k, \theta_k) = \frac{1}{\pi \lambda_g(k)} e^{-\frac{1}{\lambda_g(k)} |N(\gamma_k) - C(\gamma_k)|^2} \tag{5}$$

where

$\lambda_g(k)$  represents the variance of the additive noise.

Following the derivations in [25],  $f(C_k, \theta_k)$  becomes:

$$f(C_k, \theta_k) = \frac{C_k}{\sqrt{\lambda_c(k)}} e^{-\frac{C_k}{\sqrt{\lambda_c(k)}} (|\cos \theta_k| + |\sin \theta_k|)} \tag{6}$$

where  $\lambda_c(k)$  is the variance of the  $k^{\text{th}}$  DFT coefficients. Substituting equations (5) and (6) into equation (4) we obtain the following estimator:

$$\hat{C}_k = \frac{\int_0^{2\pi} \int_0^{2\pi} \left\{ \frac{C_k}{\pi \lambda_g(k)} e^{-\frac{1}{\lambda_g(k)} |N(\gamma_k) - C(\gamma_k)|^2} \right\} \left\{ \frac{C_k}{\sqrt{\lambda_c(k)}} e^{-\frac{C_k}{\sqrt{\lambda_c(k)}} (|\cos \theta_k| + |\sin \theta_k|)} \right\} d\theta_k C_k}{\int_0^{2\pi} \int_0^{2\pi} \left\{ \frac{1}{\lambda_g(k)} e^{-\frac{1}{\lambda_g(k)} |N(\gamma_k) - C(\gamma_k)|^2} \right\} \left\{ \frac{C_k}{\sqrt{\lambda_c(k)}} e^{-\frac{C_k}{\sqrt{\lambda_c(k)}} (|\cos \theta_k| + |\sin \theta_k|)} \right\} d\theta_k C_k} \tag{7}$$

Now, let  $\xi_k \overset{\Delta}{=} \frac{\lambda_c}{\lambda_g}$  and  $V_k \overset{\Delta}{=} \frac{N_k^2}{\lambda_g}$  denote a priori and posteriori SNRs, respectively. Substituting these into equation (7), yields a more compact form, thus:

$$\hat{C}_k = \frac{\int_0^\infty C_k^2 e^{-\frac{V_k C_k^2}{N_k^2}} \int_0^{2\pi} e^{-\left(\frac{2C_k V_k \cos \theta_k}{N_k} - \frac{C_k \sqrt{V_k}}{N_k \sqrt{\xi_k}} (|\cos \theta_k| + |\sin \theta_k|)\right)} d\theta_k C_k}{\int_0^\infty C_k e^{-\frac{V_k C_k^2}{N_k^2}} \int_0^{2\pi} e^{-\left(\frac{2C_k V_k \cos \theta_k}{N_k} - \frac{C_k \sqrt{V_k}}{N_k \sqrt{\xi_k}} (|\cos \theta_k| + |\sin \theta_k|)\right)} d\theta_k C_k} \tag{8}$$

Equation (8) represents the Laplacian Minimum Mean Square Error (MMSE) estimator of the spectral magnitude and is a modified form of Chen and Loizuo's [26] formulation for the MMSE estimator. Since the derived Laplacian MMSE estimator is computationally complex, some approximations can be applied to this equation to yield a computationally-feasible estimator. Therefore, applying the derivations in

[27], we expressly obtain a better approximation of  $f_c(c)$  (the PDF of the spectral amplitude and Bessel function) as:

$$B_{LAP-MMSE} = \frac{1}{\sqrt{2\nu}} \frac{\Gamma\left(\frac{5}{2}\right) G_{\frac{3}{2}}(P)}{\Gamma\left(\frac{3}{2}\right) G_{\frac{5}{2}}(P)} \tag{9}$$

where

$$P = \left( \sqrt{\frac{1}{\xi}} - \sqrt{2V} \right)$$

Therefore

$$\hat{C} = B_{LAP-MMSE} \cdot N \tag{10}$$

and the clean speech component is finally extracted using ISTFT and weighted overlap-add method.

### 3.4 Analysis of Recorded Speech Sample

Speech sounds of a male speaker were recorded in noisy and non-noisy backgrounds. Fig. 2 shows the waveform and spectrogram of the recorded speech in a noisy background. The waveform explains the physical signal data, with respect to time. The spectrogram reveals that every spoken word form a band of formant and its frequency are harmonically ordered neatly into striations or clear bands. From the waveform, the silent period of the noisy speech has random energy fluctuations and since noise is an error or undesired random disturbance, the energy is placed in frequency and amplitude more randomly rather than being organized neatly into clear bands.

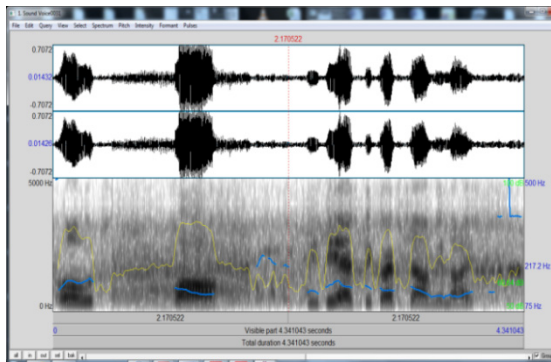


Fig. 2. Waveform and spectrogram of speech in noisy background



Fig. 3 shows the waveform and spectrogram of speech in a non-noisy background. The spectrogram representation also forms bands of formant for each word spoken and its frequency harmonically organized. The silent period does not have random energy because it appears white in the spectrogram, which shows that the noise level is negligible. As can be seen, the pitch plot in the spectrogram section is low, and the speech output sounds more intelligible compared to the one in noisy background.

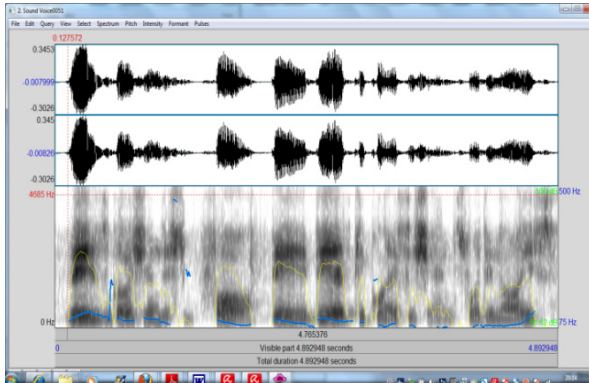


Fig. 3. Waveform and spectrogram of speech in non-noisy background

A Fast Fourier Transform (FFT) was applied to the respective wave files. The FFT analysis essentially separates the frequencies and amplitudes of its component waves. The results as can be seen in Figs. 4 and 5 are displayed with degrees of amplitude (represented light-to-dark) at various frequencies by time. From the graphs we observed that the FFT magnitude of speech in noisy background doubles that of speech in non-noisy background.

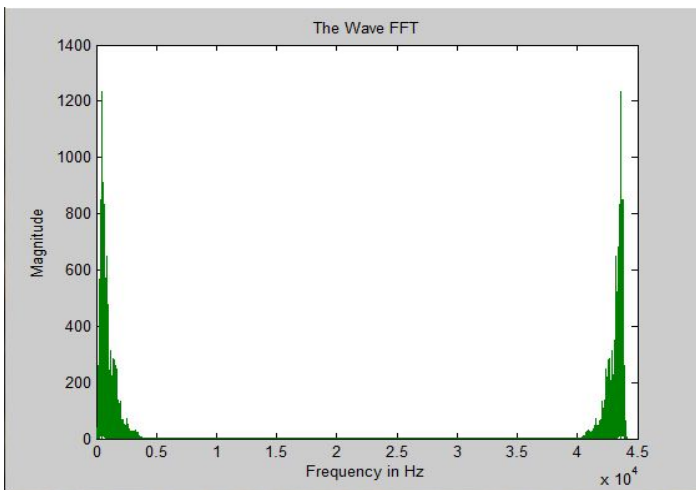


Fig. 4. Extracted FFT of speech in noisy background

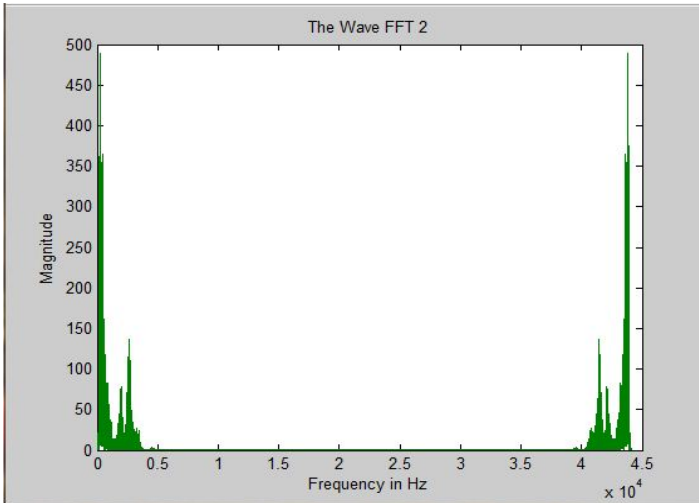


Fig. 5. Extracted FFT of speech in non-noisy environment

## 4 Implementation and Performance Evaluation

### 4.1 Speech Quality Enhancement

The *fundamental frequency* (F0) of both speech signals were extracted using *Praat*, a speech processing and analysis software. In Fig. 6, we discovered that due to the noisy background, the fundamental frequency was raised by 65Hz (on the average) above the speech signal in non-noisy background. The voiced speech of a typical adult male has a fundamental frequency between 85 to 180Hz, while that of a typical adult female falls within 165 to 255 Hz [28-29]. Hence, the fundamental frequency of most speech is expected to fall below the voice frequency band as just defined. This indicates that the presence of noise calibrates upward the F0 of speech and contributes to reducing its overall intelligibility, thus, resulting in the poor speech quality.

Next, we simulate the distortion from a quantised R-bit and compute the signal to Noise Ratio (SNR) for the signal to quantisation noise. A *quantiser* maps amplitude values into a set of discrete values  $kQ$ , where  $Q$  is the quantisation interval or stepsize and  $k$ , a constant. Quantisation is applied here to retain as much signal fidelity as possible while eliminating unnecessary precision and to maintain the dynamic range of the signal within practical limits (i.e., avoid signal clipping or arithmetic overflow). Our simulation ensures negligible loss of signal fidelity and centres on managing the approximation error to ensure that very little distortion is introduced. Simulation results in Figs. 7 and 8 revealed that noisy backgrounds contribute to raising the signal to quantisation noise.

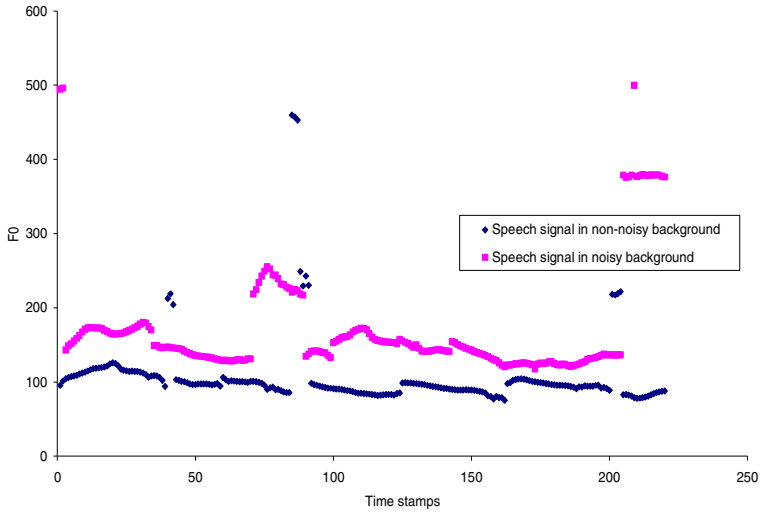


Fig. 6. F0 plots comparing speech signals in noisy and non-noisy background

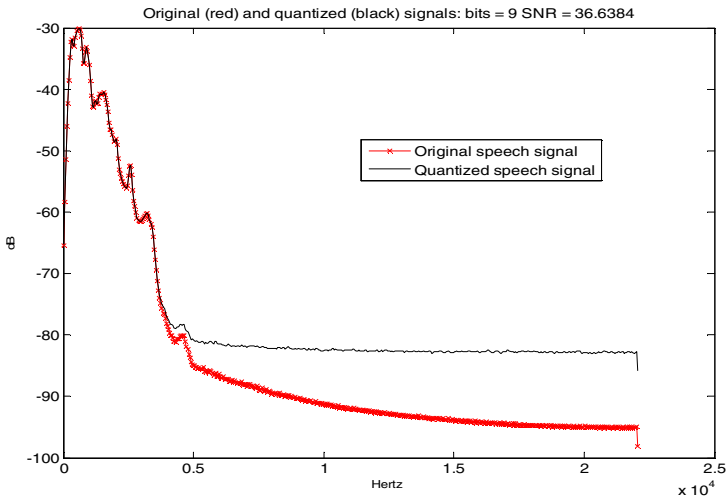
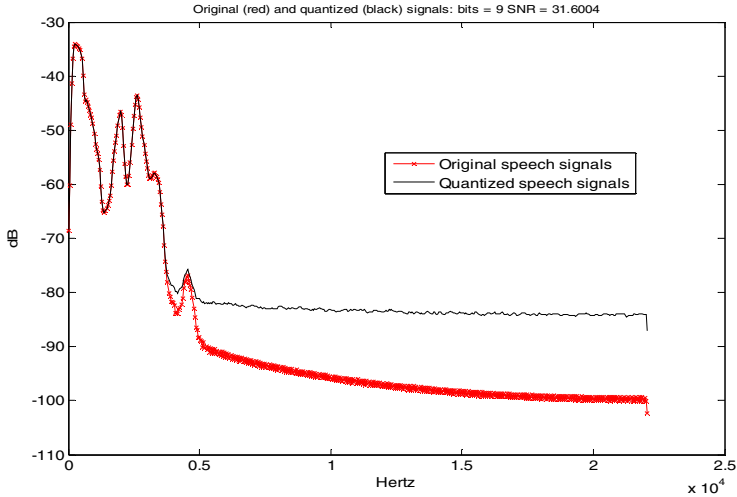


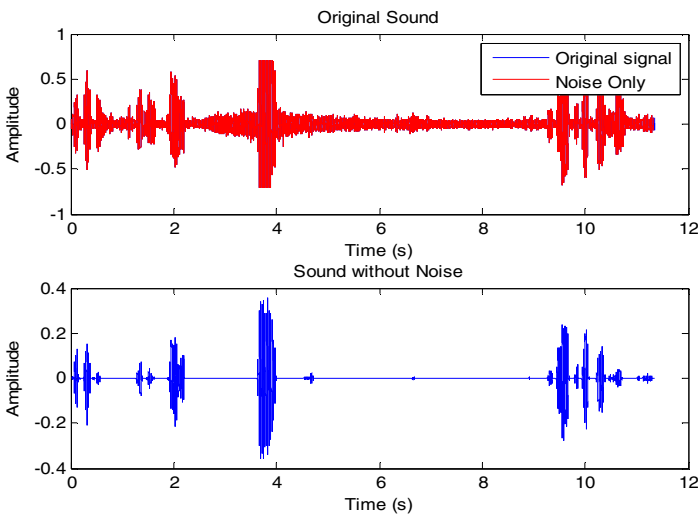
Fig. 7. Original and quantised speech signals in noisy background



**Fig. 8.** Original and quantised speech signals in non-noisy background

Quantisation could also be applied to manage distortion within the limits of the bit rate supported by a communication channel or storage medium. With quantisation, erroneous bits can be corrected and missing bits recovered through the use of a technique known as forward error correction (FEC).

Figs. 9 and 10 illustrate the application of the proposed model on the recorded speech in noisy and non-noisy backgrounds, respectively. In both figures, we separate



**Fig. 9.** Original and cleaned speech signals (in noisy background)

the noisy part of the waveform from the signal. The proposed estimator was applied to about 12 seconds duration of speech using a window function with 50% overlap between frames and a priori SNR. The enhanced speech was then combined using the overlap and add approach and the STFT of the speech signal estimated. The STFT was finally combined with the complex exponential of the noisy phase. The subplots of both speeches confirm the efficacy of our model to eliminate noisy signals. The waveforms of both speeches were re-modelled to improve intelligibility and unnecessary clips as a result of poor recording sessions.

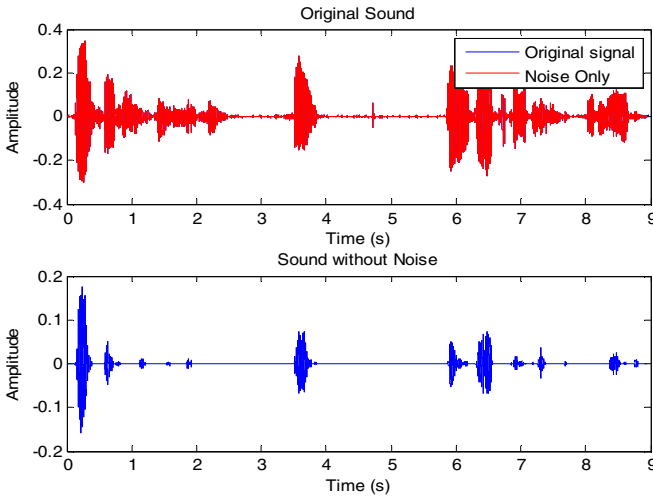


Fig. 10. Original and cleaned speech signals (in non-noisy background)

## 4.2 Voice Synthesis

The field of Forensic speech and audio analysis consist of a wide range of activities of which the most astonishing is speaker identification. Other activities include intelligibility enhancement of recorded speech samples, analysis of disputed utterances, and the examination of audio recordings authenticity. Voice disguise, depending on the extent of use may also pose serious problems in speaker identification. At the extreme end of the spectrum we find electronic manipulations even with speech synthesis data, which could render speaker identification virtually impossible. In Forensic science, however, voice disguise tends to be of a rather unsophisticated nature. Recently, the Centre for Speech Technology Research (CSTR), UK, has adopted voice synthesis to provide personalised communication aids for those who are losing the ability to speak, e.g., persons with the Motor Neurone Disease [30]. The research is a *Voice-bank project* and aims at developing clinical applications of HMM-based speech synthesis such as personalised voices for communication aids. This novel research has

the potentials of giving not only people with speech disorders *a voice*, but could also benefit Forensic science in different ways. One of the potential benefits is in the area of voice adaptation, where a donor's (e.g., a relation) voice may be adapted to assist investigation of physically challenged suspects.

To further our investigation on speech intelligibility of tone language systems, we improve on the research in [23] by performing a *discriminant analysis* (DA) on the MRT results. We study the overall effect of confused linguistic features (vowels, consonants and tone bearing units) of synthesised voices, as perceived by listeners. The essence of this analysis is to compare the overall performance of the two synthesis systems (with and without tone evidence). After cross-validation, we collated the frequency of confusions generated by a statistical tool/software called *xlstat2012* into tables of wrong and correct guesses.

In system A (synthesis system without tone), Tables 1, 2 and 3 reveal the frequency of vowel, consonant and tone confusions, respectively (perceived by listeners). In Table 1, a total of 120 vowels were wrongly perceived. Specifically, rare vowels such as  $\text{\textcircled{a}}$  and  $\text{\textcircled{u}}$ , suffered greatly. Also, the  $\text{\textcircled{A}}$ ,  $\text{\textcircled{e}}$  and  $\text{\textcircled{o}}$  vowels had higher confusion tallies. These vowels amounted to 30% of the overall vowels and 87% of the wrongly perceived vowels.

**Table 1.** Perceived vowel confusion analysis for system A

Vowel	Wrong guess	Correct guess	Total
$\text{\textcircled{a}}$	20	1	21
$\text{\textcircled{i}}$	7	28	35
$\text{\textcircled{o}}$	5	16	21
$\text{\textcircled{u}}$	4	3	7
$\text{\textcircled{A}}$	24	18	42
$\text{\textcircled{a}}$	8	13	21
$\text{\textcircled{e}}$	14	14	28
$\text{\textcircled{i}}$	6	29	35
$\text{\textcircled{o}}$	25	34	59
$\text{\textcircled{u}}$	7	14	21
Total	120	170	290

In Table 2, a total of 100 consonants were wrongly perceived by listeners. The k, m, n, p and w consonants had high tallies, indicating poor system performance. The contribution of these consonants amounted to 17.7% of the overall consonants and 50% of the wrongly perceived consonants.

**Table 2.** Perceived consonant confusion analysis for system A

Consonant	Wrong guess	Correct guess	Total
ñ	9	17	26
ʙ	1	6	7
b	7	21	28
d	2	19	21
f	5	16	21
j	8	6	14
k	12	9	21
kp	9	26	35
m	13	29	42
n	13	2	15
p	9	5	14
s	1	13	14
t	6	15	21
w	5	9	14
Total	100	193	293

The major reasons behind perceived phoneme confusions include swapped pairs – a case where phoneme pairs are closely perceived, hearing difficulties and poor recordings/synthesis. We are however investigating to uncover the exact phoneme pairs wrongly perceived by listeners.

In Table 3, we observed that a greater percentage of tone bearing units was wrongly perceived. The result revealed that synthesis systems require sufficient tone features to synthesise tone languages optimally.

**Table 3.** Perceived tone confusion analysis for system A

Tone	Wrong guess	Correct guess	Total
ˊ	12	8	20
ˋ	15	6	21
ˊˋ	16	13	29
ˊˊ	41	11	52
ˋˋ	6	1	7
ˋˊ	17	11	28
ˊˋˊ	6	7	13
ˋˋˋ	11	0	11
ˊˊˊ	7	0	7
ˋˋˊ	2	12	14
ˊá	24	4	28
á	31	25	56
â	1	6	7

**Table 3.** (continued)

à	13	29	42
ă	6	1	7
é	14	21	35
ê	14	0	14
è	5	9	14
í	12	23	35
ì	9	12	21
!ó	1	5	6
ó	26	17	43
ò	18	17	35
ú	2	6	8
ù	0	10	10
Total	309	254	563

In system B (synthesis system with tone), a total of 31 vowels, 262 consonants and 113 tone bearing units were wrongly perceived by listeners, as shown in Tables 4, 5 and 6, respectively. These results show a remarkable improvement over system A, except for Table 4 (consonant confusion analysis). Also, an investigation to reveal the exact tone pairs confusions is being carried out.

**Table 4.** Perceived vowel confusion analysis for system B

Vowel	Wrong guess	Correct guess	Total
ə	8	13	21
ɪ	4	31	35
ɔ	2	19	21
ʊ	0	7	7
ʌ	3	39	42
a	1	20	21
e	6	22	28
i	1	34	35
o	6	53	59
u	0	21	21
Total	31	259	290

**Table 5.** Perceived consonant confusion analysis for system B

Consonant	Wrong guess	Correct guess	Total
ñ	26	0	26
ʙ	7	0	7
b	24	4	28
d	18	3	21



**Table 5.** (continued)

f	21	0	21
j	13	1	14
k	19	2	21
kp	30	5	35
m	39	3	42
n	10	5	15
p	10	4	14
s	14	0	14
t	19	2	21
w	12	2	14
Total	262	31	293

**Table 6.** Perceived tone confusion analysis for system B

Tone	Wrong guess	Correct guess	Total
í	1	19	20
ì	9	12	21
!ó	10	19	29
ó	12	40	52
ò	1	6	7
ò	7	21	28
ú	2	11	13
ù	3	8	11
á	7	0	7
â	0	14	14
!á	7	21	28
á	2	54	56
â	1	6	7
à	7	35	42
ă	2	5	7
é	3	32	35
è	4	10	14
è	3	11	14
í	8	27	35
ì	0	21	21
!ó	2	4	6
ó	9	34	43
ò	12	23	35
ú	0	8	8
ù	1	9	10
Total	113	450	563

The cross validations were done using random validation sets at 0.05 level of significance with a 0.0001 tolerance. On the average, it was clear that more feature pairs were confused in system A compared to system B, which had less confusion pairs, except for the consonant pairs which requires further investigation. Also both systems showed significant difference in a Wilcoxon signed-rank test with continuity correction at  $\alpha = 0.01$ .

## 5 Conclusion and Future Works

Existing solutions to the field of digital Forensics are largely ad-hoc, and current Forensic systems record numerous false data, thereby, complicating the analysis of digital evidence [31]. The development of standard formalisms useful for performing goal-oriented Forensic modelling is therefore necessary for the purpose of ensuring accuracy and acceptability of the analysis. Although many open questions on the assessment and quality assurance of digital Forensic evidence remain, the application of speech technologies to Forensic analysis should be intensified with better procedure and methods in order to increase the acceptability of processed voices. Also, the future of Forensic science appears challenging and requires the willingness to risk failure [32]. However, this field holds lots of prospects in the future. To enjoy the full potentials of Forensic science, crime laboratories and related agencies should be provided with the needed resources such as sophisticated databases and access to state-of-the-art technology tools.

In future research, we shall integrate the current design into [20], and unify the frameworks so far modelled in [19], bearing components interaction and ‘corporate’ intelligence in mind. The resulting design is most likely to guarantee a robust-intelligent system that is capable of serving as a model for other language systems and future Forensic frameworks.

## References

1. Nance, A., Hay, B., Bishop, M.: Digital Forensics: Defining a Research Agenda. In: 42nd Hawaii International Conference on System Sciences, pp. 1–6 (2009)
2. Ren, W.: Distributed agent-based real time network intrusion Forensics system architecture design. In: 19th International Conference on Advanced Information Networking and Applications, AINA 2005, vol. 1, pp. 177–182 (2005)
3. Sathesh Kumar, S., Thomas, B., Thomas, K.L.: An agent based tool for windows mobile forensics. In: Gladyshev, P., Rogers, M.K. (eds.) ICDF2C 2011. LNICST, vol. 88, pp. 77–88. Springer, Heidelberg (2012)
4. Bhat, V.H., Rao, P.G., Abhilash, R.V., Patnaik, L.M.: A Novel data generation approach for Digital Forensic Application in Data Mining. In: 2nd IEEE International Conference on Machine Learning and Computing, pp. 86–90. IEEE Computer Society (2010)
5. Morrison, G.S.: Measuring the validity and reliability of Forensic likelihood-ratio systems. *Science & Justice* 51, 91–98 (2011)

6. McKenmmish, R.: What is Forensic Computing? In: Trends and Issues in Crime and criminal Justice, pp. 1–6. Australian Institute of Criminology (1999), <http://www.aic.gov.au>
7. Reilly, D., Wren, C., Berry, T.: Cloud Computing: Pros and cons for Computer Forensic Investigations. *Int Journal Multimedia and Image Processing (IJMIP)* 1(1), 26–34 (2011)
8. Rose, P.: Forensic speaker identification. Taylor and Francis, London (2002)
9. Rose, P.: Technical Forensic speaker recognition: evaluation, types and testing of evidence. *Comput Speech Lang* 20(2–3), 159–191 (2006)
10. Raynolds, D.A., Quanteieri, T.F., Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. *Digital signal process* 10, 19–41 (2000)
11. De Leon, P.L., Pucher, M., Yamagishi, J., Inma, H., Saratxaga, I.: Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech and Language Process* 20(8), 2280–2290 (2012)
12. Lau, Y.W., Wagner, M., Tran, D.: Vulnerability of speaker verification to voice mimicking. In: *International Symposium on Intelligent Multimedia, Video, Speech Process*, pp. 145–148 (2004)
13. Sullivan, K.P.H., Pelecanos, J.: Revisiting carl bildt’s impostor: Would a speaker verification system foil him? In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 144–149. Springer, Heidelberg (2001)
14. Zhang, C., Morrison, G.S., Thiruvaran, T.: Forensic voice comparison using Chinese /iau/. In: *17th ICPhS, Hong Kong, China*, pp. 2280–2283 (2011)
15. Huang, C.C., Epps, J.: A study of automatic phonetic segmentation for Forensic voice comparison. In: *IEEE International conference on Acoustic, Speech and Signal Process*, pp. 1853–1856 (2012)
16. Kind, S.: *The Scientific Investigation of Crime*. Forensic Science Services Ltd., Harrogate (1987)
17. Ribaux, O., Walsh, S.J., Margot, P.: The contribution of Forensic science to crime analysis and investigation: Forensic intelligence. *Forensic Science International* 156, 171–181 (2006)
18. Brewer, N., Liu, N., De Vel, O., Caelli, T.: Using Coupled Hidden Markov Models to Model Suspect Interactions in Digital Forensic Analysis. In: *IEEE International Workshop on Integrating AI and Data Mining, AIDM 2006*, pp. 58–64 (2006)
19. Ekpenyong, E., Urua, E.-A.: Agent-based Framework for Intelligent Natural Language Interface. *Telecommunication Systems Journal* (2011a) (First online, September, 2011)
20. Ekpenyong, M.: Optimizing Speech Naturalness in Voice User Interface Design: A Weakly-Supervised Approach. In: *Proceedings of IEEE World Congress on Information and Communication Technologies, Mumbai, India*, pp. 99–105 (2011b)
21. Toda, T., Kawai, H., Tsuzaki, M., Shikano, K.: An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis. *Speech Communication* 48, 45–56 (2006)
22. Nusbaum, H.C., Francis, A.L., Henly, A.S.: Measuring the naturalness of synthetic speech. *International Journal of Speech Technology* 2(1), 7–19 (1997)
23. Ekpenyong, M., Urua, E.-A., Watts, O., King, S., Yamagishi, J.: Statistical Parametric Speech Synthesis for Ibibio. *Speech Communication* (2013), <http://dx.doi.org/10.1016/j.specom.2013.02.003> (First online: February 2013)
24. Ephraim, Y., Malah, D.: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. In: *IEEE Int. Conf. Acoustic., Speech, Signal Processing, ASS*, vol. P-32(6), pp. 1109–1121 (1984)

25. Papoulis, A., Pillai, S.U.: Probability, Random Variables, and Stochastic Processes. McGraw Hill (2001)
26. Chen, B., Loizou, P.C.: A Laplacian-based MMSE estimator for speech enhancement. *Speech Communication* 49, 134–143 (2007)
27. Rashidi-nejad, M., Abutalebi, H.R., Tadaion, A.A.: Speech Enhancement using an Improved MMSE Estimator with Laplacian Prior. In: 5th International Symposium on Tele-Communications, pp. 889–894 (2010)
28. Titze, I.R.: Principles of Voice Production. Prentice Hall (1994)
29. Baken, R.J.: Clinical Measurement of Speech and Voice. Taylor and Francis Ltd, London (1987)
30. Yamagishi, J., Veaux, C., King, S., Renals, S.: Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology* 33(1), 1–5 (2012)
31. Peisert, S., Bishop, M., Karin, S., Marzullo, K.: Toward Models for Forensic Analysis. In: 2nd International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE), Seattle, WA, pp. 3–15 (2007)
32. Shapiro, H.T.: 'The willingness to risk failure. *Science*, Editorial 250(4981), 609 (1990)

# Erratum: Impact of Some Biometric Modalities on Forensic Science

Ali Ismail Awad<sup>1</sup> and Aboul Ella Hassanien<sup>2</sup>

<sup>1</sup> Faculty of Engineering, Al Azhar University, Qena, Egypt  
Member of the Scientific Research Group in Egypt (SRGE)  
aawad@ieee.org

<sup>2</sup> Faculty of Computers & Information,  
Cairo University, Cairo, Egypt  
Chairman of Scientific Research Group in Egypt (SRGE)  
aboitcairo@gmail.com

A.K. Muda et al. (eds.), *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications*, Studies in Computational Intelligence 555, DOI: 10.1007/978-3-319-05885-6\_3, © Springer International Publishing Switzerland 2014

---

**DOI 10.1007/978-3-319-05885-6\_19**

In the original version, the author's Email id and affiliations were incorrect. It should read as:

Ali Ismail Awad<sup>1,2</sup> and Aboul Ella Hassanien<sup>3</sup>

<sup>1</sup>Department of Computer Science, Electrical and Space Engineering  
Lulea University of Technology, Lulea, Sweden  
ali.awad@ltu.se

<sup>2</sup>Faculty of Engineering, Al Azhar University, Qena, Egypt

<sup>3</sup>Faculty of Computers & Information  
Cairo University, Cairo, Egypt  
Chairman of Scientific Research Group in Egypt (SRGE)  
aboitcairo@gmail.com

---

The original online version for this chapter can be found at  
[http://dx.doi.org/10.1007/978-3-319-05885-6\\_3](http://dx.doi.org/10.1007/978-3-319-05885-6_3)

---

# Author Index

- Abraham, Ajith 1  
Akosu, Nicholas 63  
Alahakoon, Dammina 97  
Awad, Ali Ismail 47
- Bagchi, Parama 315  
Basu, Dipak Kumar 315  
Bera, Asish 145  
Bhattacharjee, Debotosh 125, 145, 315  
Bhowmik, Mrinal Kanti 125  
Blumenstein, Michael 285  
Boo, Yee Ling 97  
Bustamante, Isneri Talavera 187
- Choo, Yun-Huoy 385
- Ekpenyong, Moses 429
- Haron, Habibollah 165  
Hasanzadeh, Faranak 17  
Hassanien, Aboul Ella 47  
Hernández, Noslen 187
- Ismail, Widad 211
- Júnior, Agostinho M. Brito 253
- Kadir, Mohammed R.A. 165  
Keyvanpour, MohammadReza 17  
Khor, JingHuey 211  
Kundu, Anirban 79
- Martínez, Lázaro Bustio 187  
Martínez-Díaz, Yoanna 187  
Mata, Francisco José Silva 187  
Medeiros, João P. Souza 253  
Mitra, Arnab 79  
Moradi, Mohammad 17  
Muda, Azah Kamilah 1, 385  
Muda, Noor Azilah 385  
Muñoz, Dania Porro 187  
Muñoz, Diana Porro 187
- Nasipuri, Mita 145, 315  
Neto, João B. Borges 253
- Obot, Okure 429
- Pal, Srikanta 285  
Pal, Umapada 285  
Pires, Paulo S. Motta 253  
Pratama, Satrya Fajri 1, 385  
Pratiwi, Lustiana 1
- Rahman, Mohammad Ghulam 211  
Riquelme, José C. 413
- Saha, Kankan 125  
Sahadun, Nur A. 165  
Selamat, Ali 63  
Srihari, Sargur N. 333
- Tallón-Ballesteros, Antonio J. 413  
Tang, Yi 333