THESIS
2
2009

LIBRARY
Michigan State
University

This is to certify that the
dissertation entitled

STATISTICAL MODELS FOR FINGERPRINT INDIVIDUALITY

presented by

YONGFANG ZHU

has been accepted towards fulfillment
of the requirements for the

Ph.D. degree in Statistics and Probability

Major Professor's Signature

09/29/08

Date

MSU is an affirmative-action, equal-opportunity employer

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

5/08 K:/Proj/Acc&Pres/CIRC/DateDue.indd

STATISTICAL MODELS FOR FINGERPRINT INDIVIDUALITY

By

Yongfang Zhu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics and Probability

2008

ABSTRACT

STATISTICAL MODELS FOR FINGERPRINT INDIVIDUALITY

By

Yongfang Zhu

The U.S. Supreme Court in the 1993 case of Daubert vs. Merrell Dow Pharmaceuticals ruled that scientific evidence presented in a court is subject to the principles of scientific validation that include whether (i) the particular technique or methodology has been subject to statistical hypothesis testing, (ii) its error rate has been established, (iii) standards controlling the technique's operation exist and have been maintained, (iv) it has been peer reviewed, and (v) it has a general widespread acceptance. Following Daubert, forensic evidence based on fingerprints was first challenged in the 1999 case of USA vs. Byron Mitchell based on the "known error rate" condition mentioned above, and subsequently, in 20 other cases involving fingerprint evidence. The main concern with the admissibility of fingerprint evidence is the problem of individualization, namely, that the fundamental premise for asserting the uniqueness of fingerprints has not been objectively tested. In other words, the fingerprint matching error rate is unknown. The problem of fingerprint individuality can be formulated as follows: Given a query fingerprint, what is the probability of finding a fingerprint in a target population having features similar to that of the query? To answer this question, the variability of fingerprint features, namely, minutiae in the target population needs to be understood and quantified in terms of statistical models.

For minutiae interclass variability, a family of mixture models is developed to model the minutiae variability of individual fingers, including minutiae clustering tendencies and dependencies in different regions of the fingerprint image domain. For a heterogeneous population of fingers, a hyper-mixture model is proposed to cluster the population into homogeneous groups having similar distribution given by the mixture models. The group

densities and weights are acquired by clustering the mixture models fitted to individual fingers from a sample of the population

Whereas mixture models take into account the minutiae interclass variability, a compound stochastic model is developed for more sources of minutiae variability, i.e., besides interclass variability, the model also considers intraclass variability, such as nonlinear deformation and variability due to partial prints.

The proposed models are shown to better describe the observed variability in the minutiae compared to the model by Pankanti et al. [35]. To quantify fingerprint individuality, a mathematical model that computes the probability of a random correspondence (PRC) between minutiae sets of two randomly selected different fingers is derived. Comparison of PRCs with empirical matching probability shows that the PRCs from the proposed model are closer to the empirical matching probability than those calculated by the model of Pankanti et al. [35].

Copyright by

Yongfang Zhu

2008

This dissertation is dedicated to my professors and my loving family.

ACKNOWLEDGMENTS

Being a doctoral student in statistics, it has been a great opportunity for me to work in the field of biometrics. With the help from my professors and colleagues, my graduate study at Michigan State University has been fruitful.

I would like to express my gratitude to my advisor, Dr. Sarat C. Dass, for introducing me into this very challenging and yet very interesting field of biometrics. I would like to thank him for his continuous guidance for both my research and study, as well as his precious advice for my professional career. Being a non-native English speaker, I received a lot of help from Dr. Dass. He read and corrected my research articles, which led to my improvement on scientific writing. He listened to my talk every time before I gave a presentation, and offered substantial suggestions.

I am grateful to Dr. Lijian Yang, Dr. Connie Page, Dr. Dennis Gilliland from the Department of Statistics and Probability, and Dr. Sasha Kravchenko from the Department of Crop and Soil Sciences, for serving in my thesis committee. Besides the help for my thesis research, Dr. Yang gave me good suggestions on my professional development. I accumulated practical experience from the consulting work at the Center for Statistical Training and Consulting. Dr. Page, Dr. Gilliland, and Dr. Sandra Herman shared with me their experience and knowledge in dealing with practical problems, and guided me through my projects.

I would like to thank Dr. James Stapleton for his help and encouragement on my teaching and study. He has been very supportive to me, as well as to all the other students. When we have problems with our study, or any problem in everyday life, he would spend time with us, and gave very kind suggestions.

I would like to especially thank Dr. Mark M. Meerschaert and the Department of Statistics and Probability, for being very considerate and kind to me and many other students, helping us going through hard times, and trying the best to solve any possible problems.

I want to thank my colleagues in the PRIP lab at the Department of Computer Science

and Engineering, Michigan State University, for their help and being supportive for my research: Hong Chen, Yi Chen, Dirk Colbry, Steve Krawczyk, Pavan K. Mallapragada, Karthik Nandakumar, and Unsang Park. I thank Dr. Anil K. Jain for letting me use the PRIP lab for my research and for reading my thesis. I would like to thank Ms. Wenmei Huang for discussions on course work and research. I thank Dr. Peter Xinya Zhang for spending a lot of time with this thesis work, and for continuously being very supportive for my entire study. As one of the first readers, he gave me a lot of comments on the structure and narration of the thesis from a perspective outside the field. I also received from him many helpful suggestions for my presentations.

I benefited from the comments and suggestions from Dr. Sharath Pankanti and Dr. Salil Prabhakar on some of my publications. I thank the National Science Foundation for supporting this research.

Last but not least, I would like to thank my family for their continuous support, understanding and encouragement for everything in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction to Fingerprint Based Recognition

1.1 Overview of Fingerprint Recognition

Biometric recognition refers to the automatic authentication of humans using his/her anatomical or behavioral characteristics. Biometric recognition is more reliable compared to traditional approaches, such as password-based or token-based approaches, as biometric traits cannot be easily stolen or forgotten. Some examples of biometric traits include fingerprint, face, signature, voice and hand geometry (see Figure 1.1). Biometric recognition systems have been deployed and implemented for human recognition (e.g., US-VISIT program and the e-biometric passport which stores the owner's biometric information in a chip inside a passport). With increasing applications involving human-computer interactions, there is a growing need for fast authentication techniques that are reliable and secure. Biometric recognition meets such demand.

For a particular biometric trait to be considered for human authentication, several requirements need to be met, namely, (i) universality, (ii) distinctiveness, (iii) permanence, and (iv) collectability ([42], [27]). Universality requires that every human being possesses

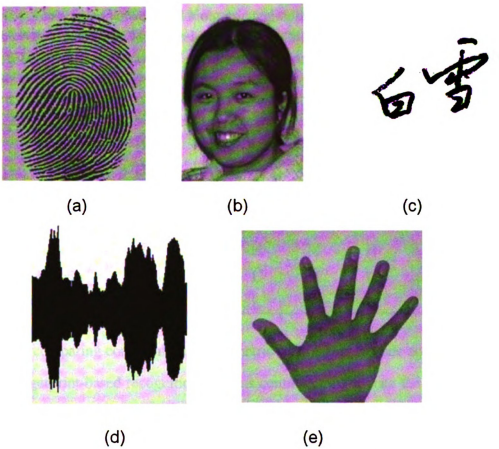


Figure 1.1: Some examples of biometric traits: (a) fingerprint [26], (b) face, (c) signature, (d) voice and (e) hand geometry

the biometric trait; distinctiveness requires that two persons have sufficiently different trait characteristics; permanence requires that trait characteristics remain unchanged over time; collectability requires that traits are quantitatively measurable. There are further considerations for practical biometric systems, such as: (i) performance and authentication rates, measured in terms of speed and recognition accuracy, (ii) public acceptance for use in our daily lives, (iii) the extent of which the biometric recognition system can be attacked or spoofed, and (iv) cost efficiency.

Among biometric traits currently used, fingerprint has the longest history, and has been widely adopted in both forensic and civilian applications. This is because fingerprint meets the previously discussed requirements of a successful biometric trait, (see Table 1.1 taken

Table 1.1: Comparison of selected biometric technologies adapted from Maltoni et al. [27]. UVSL = Universality, DSTC = Distinctiveness, PRMN = Permanence, CLTB = Collectability, PRFM = Performance, ACPT = Acceptability, and CRVN = Circumvention.The symbols H, M and L denote High, Medium and Low, respectively [27].

Biometric Trait	UVSL	DSTC	PRMN	CLTB	PRFM	ACPT	CRVN
DNA	H	H	H	L	H	L	L
Face	H	L	M	H	L	H	H
Fingerprint	M	H	H	M	H	M	M
Hand Geometry	M	M	M	H	M	M	M
Iris	H	H	H	M	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

from [27] comparing commonly-used biometric traits). Due to wide acceptance of fingerprints, fingerprint-based recognition systems continue to dominate the biometrics market by accounting for 52% of current authentication systems [27].

Rapid development of mobile commerce and mobile banking in recent years has created new demands for biometric authentication. Some biometric systems, such as fingerprint, voice and face, have appeared in some high-end mobile phones. Miniaturized fingerprint sensors, capable of being embedded in a cell phone, have been recently developed. Different from commonly-used two-dimensional sensors, new line-scan sensors enable a fingerprint impression to be captured by swiping a finger across a line. Thus biometric authentication, with increasing deployment in various applications, is here to stay.

There are two modes of biometric recognition, namely, verification and identification. In the verification mode, the recognition task is to verify the claimed identity, I_c, of a user based on an input fingerprint, Q. A biometric system retrieves the template, T, of I_c that is stored in its database, and extracts features from Q. The extracted and retrieved features are tested by a fingerprint matcher, and a similarity measure $S(Q,T)$ is obtained (1:1 match). When the similarity measure is above (respectively, below) a pre-determined threshold λ, the claimed identity is accepted (respectively, rejected).

On the other hand, when the biometric system is in the identification mode, the extracted features of an input image Q are tested against the extracted features of every stored template and a decision is made on whether a match is found or not. Compared to verification, identification is much more difficult because no claimed identity is available. For a system with M templates, the system searches through the entire database to recognize an individual (1 to M matches).

1.2 General Procedures in Fingerprint Recognition

After establishing a system database through fingerprint enrollment, both the verification and identification tasks can be divided into three different components, namely fingerprint acquisition, feature extraction, and fingerprint matching. Figure 1.2 shows the basic tasks in enrollment, verification and identification of a fingerprint-based recognition system.

1.2.1 Fingerprint Enrollment

Enrollment is the procedure of sensing and storing fingerprint templates to establish a system database. First, a fingerprint image is captured by a sensor. After that, a quality checker is applied to the image. If the quality is good, the image is retained. Otherwise, the captured image is deleted and a new fingerprint image is acquired. The process is continued till the quality of the acquired image is sufficiently good. Image quality is determined by the clarity of the ridge and valley structures and measured based on many different parameters, e.g., image resolution, sensed area, image contrast and the extent of deformation in the enrolled finger. The quality checker ensures that the acquired images have low noise. Noisy images can create problems for later processing (for example, during the fingerprint matching stage). Figure 1.3 shows three images with good, medium and poor quality from FVC 2002 DB1 database [26]. Poor quality images produce spurious features which lead to fingerprints from the same finger looking different, or fingerprints from different fingers

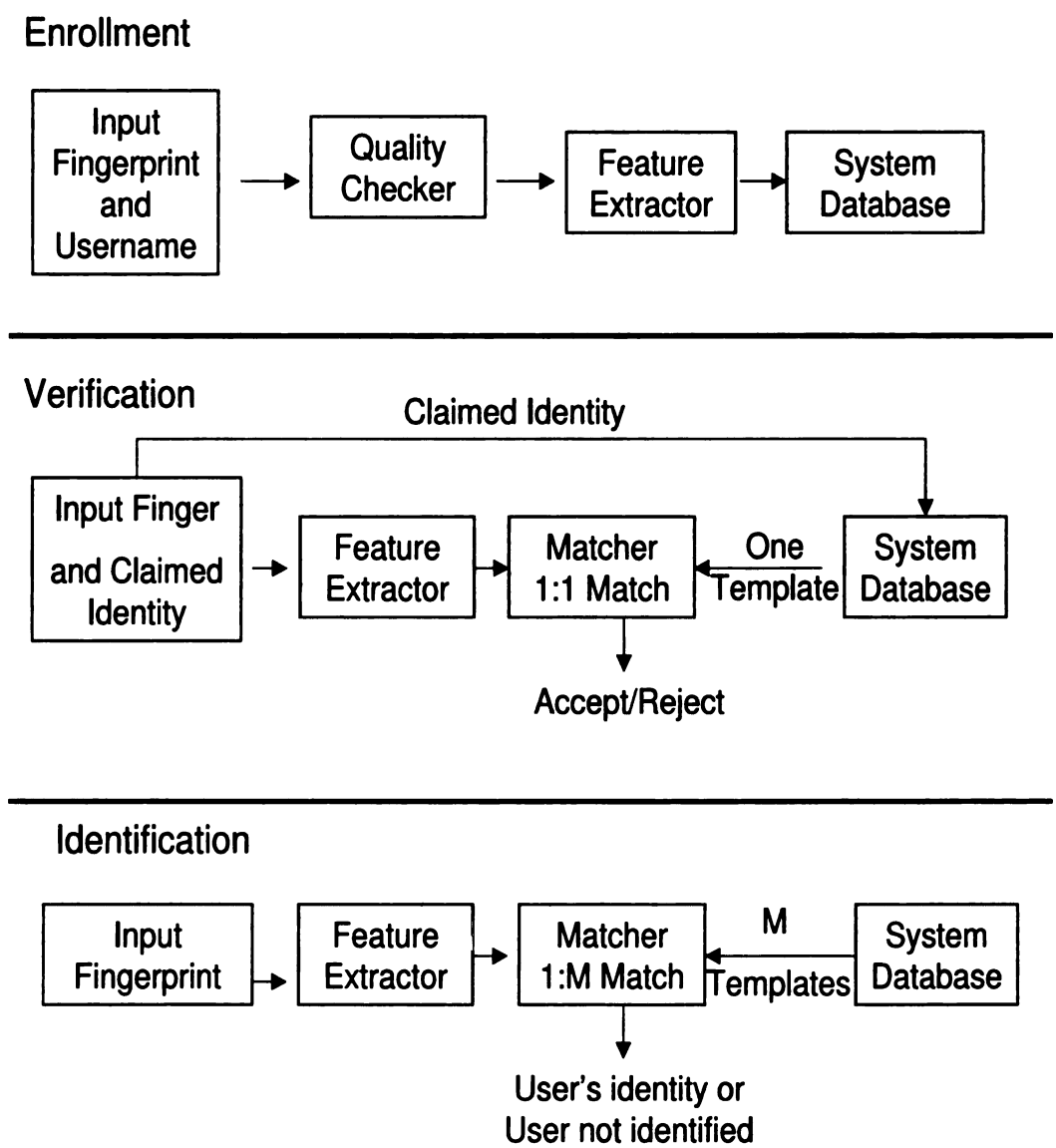


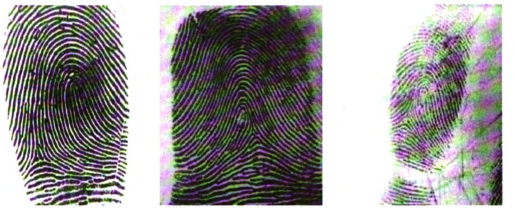
Figure 1.2: Schematic diagram showing the processing tasks involved in the enrollment, verification and identification modes of a fingerprint-based authentication system [27].

Figure 1.3: Fingerprints of different quality based on the clarity of ridge and valley structures: (a) good, (b) medium, and (c) bad. Extracting features and establishing matches based on (c) can be difficult. Images are from FVC 2002 DB1 [26]

looking alike, and thus should be avoided. Finally, a feature extractor is applied to the enrolled image and the extracted features are stored in the system database.

1.2.2 Fingerprint Acquisition

Fingerprint acquisition is to capture fingerprint images during the recognition phase and enrollment phases. There are two types of capture procedure, namely live scan and off-line scan. Currently both types of fingerprints are used in applications. For example, live fingerprints are used in the Automated Fingerprint Identification System (AFIS) [27] whereas off-line fingerprints are still used in the forensic applications. In a live scan, fingerprints are acquired directly from the sensors. In an off-line scan, preliminary fingerprint images are obtained first and the final fingerprint images are obtained by digitizing the preliminary images. Two examples of off-line fingerprint scans are ink-based and latent scans. During the ink scans, fingers are first spread with ink and rolled from nail edge to nail edge against a paper fingerprint card. The rolled images can be digitalized through either a paper scanner or a high quality camera. Latent fingerprint is a film of moisture or grease from fingerprint ridges deposited on the surface of touched objects. Due to poor quality of the lifted image,



(a) (b) (c)

Figure 1.4: Fingerprints of different sources: (a) live [26], (b) ink [33], and (c) latent fingerprint scans [15]

later enhancement methods, such as powder dusting, ninhydrin spraying, iodine fuming, and silver nitrate soaking, are involved for better fingerprint detection [6]. Figures 1.4 (a-c) show examples of a live fingerprint, an ink fingerprint and a latent fingerprint.

1.2.3 Fingerprint Feature Extraction

Fingerprint Features

After the fingerprint acquisition, salient features need to be extracted for later matching. As previously discussed, there is also feature extraction in the enrollment procedure. These two extractions are similar with one main difference: Feature extraction in the enrollment stage is used to establish a database and the quality of the acquired image can be controlled, whereas during the testing phase, especially in a latent scan, we can not guarantee such a good quality image. Two groups of features, namely global features and local features, are critical in fingerprint matching. Local features, which are details that are believed to be unique to an individual, are used for fingerprint matching.

Global features are used for fingerprint classification and for ruling out erroneous types of fingerprints prior to matching. Fingerprint classification is the problem of binning finger-

print images into different classes. Figure 1.5 shows examples of major fingerprint classes that include arch, loop (includes left loop and right loop), and whorl. Different fingerprint classes are differentiated by the global ridge structures. For example, the fingerprints of the left loop have ridges initiated from the left side of the fingerprints and continue to the center area and finally come back to the left side.

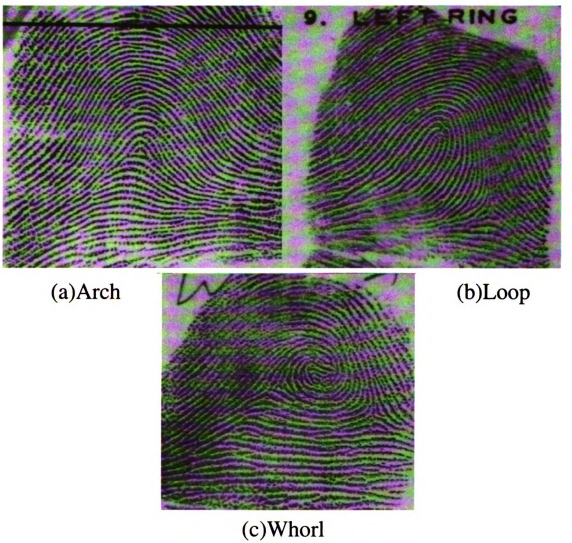


Figure 1.5: Examples of fingerprint images from the major classes. Images are from NIST 2000 SD 4 [33].

Commonly-used global features include (i) singular points and (ii) the directional field of

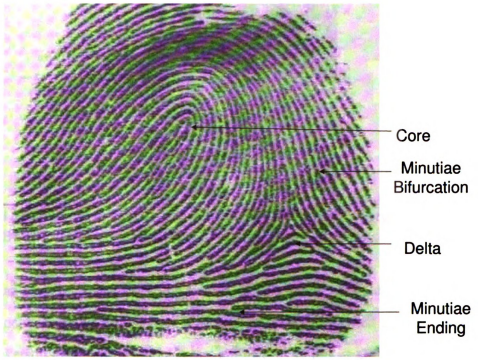


Figure 1.6: A fingerprint image showing the salient features [33]

ridge flow. Singular points are the discontinuities of fingerprints which consist of delta and core. A delta is a point in a fingerprint image which is the confluence of ridge flows in three different directions and a core is the point of inner most ridge with maximum curvature. The core and delta are labeled in Figure 1.6.

Detection of singularity (i.e., core and delta) has been the focus of many previous studies. In Nakamura et al. [32], and Srinivasan and Murthy [46], singularities were detected by first finding high curvature regions and then classifying the regions into core, delta and non-singular regions. In Rao and Jain [37], a method based on geometric theory of differential equations was used to detect cores and deltas.

The directional field reveals direction of the ridge flow for each pixel or a block of pixels in fingerprint images. Ridge flow direction can be described by an angle θ with respect to the x-axis. Opposite ridge flow directions are equivalent, and therefore θ can only be determined in $[0, \pi]$. The main challenge in estimation of directional field is that gradient with opposite directions should not cancel each other, but rather reinforce them (see Figure

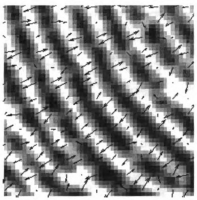


Figure 1.7: Gradient directions and magnitudes of a partial fingerprint region from FVC 2002 DB1 [26] indicated by arrow heads and lengths, respectively.

1.7). Many studies have focused on estimating the directional field. These studies include neural- network based approaches (Wilson et al. [51]), filter-based approaches (O'Gorman and Nickerson [34]), and gradient-based approaches (Hong et al. [20], Jain et al. [1], and Bazen and Gerez [5]).

Local features are anomalies along ridge and valley structures, which are usually called minutiae. There are several types of minutiae: ending, bifurcation, island, spur, crossover, lake and others. Figure 1.8 shows examples of different types of minutiae. A minutia ending is a point where a ridge terminates, and a minutia bifurcation is a point where a ridge bifurcates into two almost parallel ridges. This thesis focuses on ridge bifurcation and ending because the other types of minutiae occur much less frequently compared to endings and bifurcations in fingerprint images. Moreover, other types of minutiae can be thought of as combinations of endings and bifurcations. Examples of a minutia ending and bifurcation in a fingerprint image are shown in Figure 1.6 and Figure 1.8 (a) and (b). Direction of a minutia ending is normally represented by the angle between the horizontal axis and the minutiae tangent pointing away from the ridge terminating point. Direction of a minutia bifurcation is represented by the angle between the horizontal axis and the tangent pointing into the ridge prior to the bifurcation. Hence the minutiae direction of an ending or a bifurcation is in range $[0, 2\pi]$. Figure 1.9 shows illustrations of minutiae directions

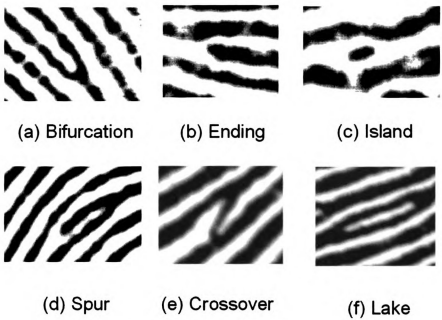
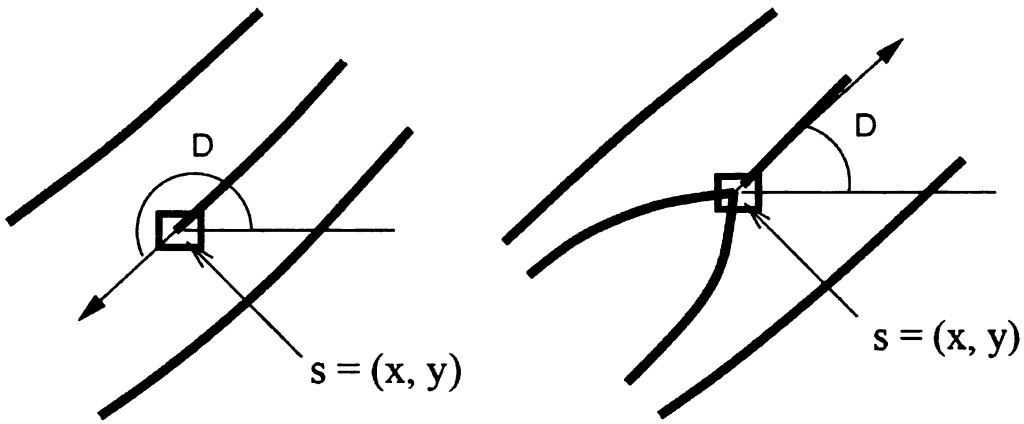


Figure 1.8: Examples of different minutiae. Images are from FVC 2002 DB1 [26].

and locations for both ending and bifurcation, where minutiae location and direction are represented by s and D, respectively.

Minutiae Extraction

Local features or minutiae, are most important for fingerprint matching. To extract minutiae from fingerprint images, various algorithms have been developed. Binarization and direct gray scale approaches are the two most popular methods for extracting minutiae. Both of them use gray scale images, and the difference is that the binarization approach requires gray scale images to be converted into black-and-white images, whereas the direct gray scale method detects minutiae without binarizing gray scale images. The feature extraction algorithm used in this thesis is a binarization method reported in [1]. The enhancement process involves strengthening the clarity of the ridge structures using directional Gabor filters. This is followed by thinning where the enhanced ridges are reduced to connected components of one pixel wide. The minutiae locations are then detected in the thinned image as breaks or bifurcations in the connected components. Figure 1.10 (a) shows a



(a) Ending (b) Bifurcation

Figure 1.9: Locations and directions of bifurcation and ending

typical fingerprint image in the FVC 2002. Figures 1.10 (b) and (c) show, respectively, images after applying enhancement and thinning algorithms to Figure 1.10 (a). Detected minutiae locations and directions are shown in Figure 1.10 (d).

1.2.4 Fingerprint Matching

After features are extracted from the query image, they are matched to the template saved in the system database. There are three approaches for fingerprint matching [27], namely correlation-based, minutiae-based and ridge-feature-based matching. Correlation-based matching computes a correlation between pixel gray values of aligned images. Minutiae-based matching optimizes the alignment by maximizing the number of matched minutiae pairs between aligned minutiae patterns. Minutiae-based matching is the most popular among the commercial fingerprint matchers and it is the basis of fingerprint matching used by forensic fingerprint experts. An example of minutiae-based fingerprint matching is shown in Figure 1.11, where matched minutiae correspondences are marked by connecting lines. Ridge-based matching is usually applied to fingerprint images with low quality. Usually, for these low-quality images, correlation calculation is less consistent and minutiae

(a) (b) (c) (d)

Figure 1.10: Minutiae feature extraction steps. Different fingerprint processing stages for extracting minutiae features: (a) original image from FVC 2002 DB1 [26], (b) enhanced image, (c) thinned image and (d) detected features with minutiae locations indicated by black boxes, and directions represented by solid lines.

features are unavailable, but ridge-pattern features (e.g., local orientation and frequency, texture information, and ridge shape) are more reliable. We used minutiae based matching in this thesis to match two fingerprint images.

1.3 Statistical Test of Hypothesis in Fingerprint Recognition

The task of biometric recognition can be described in terms of a statistical test of hypothesis. Suppose a query image, Q, corresponding to the true but unknown identity, I_t, is acquired. In order to carry out a test to determine whether Q belongs to a claimed identity I_c, template T corresponding to I_c is retrieved from the system database and is matched with Q. The null hypothesis is that I_c is not the owner of the fingerprint Q (i.e., Q is an impostor impression of I_c), and the alternative hypothesis is that I_c is the owner of Q (i.e., Q is a genuine impression of I_c). The null-alternative hypothesis testing scenario is

$$H_0 : I_t \neq I_c \quad \text{vs.} \quad H_1 : I_t = I_c. \tag{1.1}$$

Figure 1.11: Example of fingerprint matching. Two impressions of the same finger from FVC 2002 DB1 [26] with 37 and 38 minutiae, respectively. 25 true correspondences are found here.

Suppose the matching score between Q and T is $S(Q,T)$. Large (respectively, small) values of $S(Q,T)$ indicate that T and Q are similar to (respectively, dissimilar to) each other. If $S(Q,T)$ is lower (respectively, higher) than a pre-specified threshold λ, it leads to rejection (respectively, acceptance) of H_0. Consequently, two types of errors can be made, the probabilities of which are false reject rate (FRR) and false acceptance rate (FAR), respectively. By definition, false reject rate ,which can also be called false non-match rate (FNMR), is the probability of incorrectly rejecting a genuine fingerprint query, and false acceptance rate, which can also be called the false match rate (FMR), is the probability of incorrectly accepting an impostor query. The formulae of FRR and FAR are:

$$\begin{aligned} FRR(\lambda) = FNMR(\lambda) &= P(S(Q,T) \leq \lambda | I_t = I_c), \\ FAR(\lambda) = FMR(\lambda) &= P(S(Q,T) > \lambda | I_t \neq I_c). \end{aligned} \tag{1.2}$$

Two sources of variability, namely large intraclass variability and small interclass variability cause erroneous decisions when testing between H_0 and H_1. Intraclass variability

Figure 1.12: Multiple impressions of the same finger illustrating the intraclass variability [26]

Figure 1.13: Illustrating small interclass variability: A pair of impostor fingerprints with 13 features (minutiae location and direction) in correspondence.

refers to the fact that fingerprints from the same finger look different from one another. Sources for this variability include non-linear deformation due to skin elasticity, partial print, non uniform fingertip pressure, poor finger-condition (e.g., dry finger), and noisy environment, etc. Non-linear deformation is introduced into a fingerprint image when a three dimensional fingertip is projected onto a two dimensional sensing plane. Partial image is due to small sensing surface, capturing only a portion of fingerprint. Poor finger condition and noisy environment can produce images whose features are unable to discriminate between different fingers. In Figure 1.12, multiple impressions of the same finger appear different from each other due to various sources of intraclass variability discussed above. On the other hand, interclass variability is inherent in different fingers in the population. Small interclass variability causes fingerprints from different individuals to look very similar to each other (See Figure 1.13). The research presented in this thesis develops statistically models for both the interclass and intraclass variability of fingerprint minutiae.

1.4 Summary

This chapter gives an overview of fingerprint-based recognition. Among various biometric traits, fingerprints are most commonly used. They are also universal, distinctive, permanent and collectible, and thus have dominated the biometric market over a long time. There are two modes of fingerprint recognition, i.e., verification and identification. In a verification mode, a fingerprint is matched against a claimed identity; whereas in an identification mode, a fingerprint is matched against all fingerprint images in the database.

The general procedure of fingerprint recognition includes four steps, namely fingerprint enrollment, fingerprint acquisition, feature extraction and fingerprint matching. Fingerprint verification can be viewed as a statistical test of hypothesis that involves two types of errors, namely, false acceptance rate and false reject rate. This thesis focuses on false acceptance rate since it gives an estimate of the probability of random correspondence (that is, the probability that a pair of different fingers will match with each other), which is the measure of fingerprint individuality.

CHAPTER 2

Fingerprint Individuality

2.1 Importance of Fingerprint Individuality

There are two premises underlying fingerprint based recognition. The first premise is that fingerprints are permanent, i.e., fingerprints do not change over a person's life-time. The second premise is that fingerprints are unique, i.e., the characteristics of fingerprint features of different fingers are different. The first premise has been widely investigated and proven to be valid based on the anatomy and morphogenesis of friction ridge skin [27]. However, the second premise of uniqueness has not been thoroughly studied. In particular, when fingerprints are matched, statistical measures of the confidence associated with the match have not been thoroughly investigated. Fingerprint individuality is the study of the extent of uniqueness of fingerprints and this is the focus of the research presented in this thesis.

Investigation of fingerprint individuality is most important in the legal setting. Expert testimony based on fingerprint evidence is delivered in a courtroom by comparing salient features of a latent print query lifted from a crime scene with that of the defendant. Thus, we are in the hypothesis testing scenario of Equation 1.1 where the court has to decide whether the defendant is truly the criminal (reject H_0) or otherwise (accept H_0). A reasonably high degree of match between the query and template fingerprints from the defendant leads the

experts to testify irrefutably that the owner of the latent print and the defendant are the same person. For decades, the testimony of forensic fingerprint experts was almost never excluded from these cases, and on cross-examination, the foundation of this testimony was rarely questioned. A prerequisite to establishing an identity based on fingerprint evidence is the assumption of discernible uniqueness, i.e., salient features of fingerprints from different individuals are different. Only when this is true can the experts conclude that the owners of two different prints with reasonably high degree of similarity are one and the same person. However, in reality, forensic experts are never questioned on the uncertainty associated with their testimony (that is, how frequently an observable match between a pair of prints would lead to errors in identification of individuals). Thus, discernible uniqueness precludes the opportunity to determine error rates of fingerprint matching from analyzing inherent feature variability and calculating the probability of two different persons sharing a set of common features.

A significant event that broke this trend occurred in 1993 in the case of Daubert v. Merrell Dow Pharmaceuticals [13] where the U.S. Supreme Court ruled that in order for an expert forensic testimony to be allowed in courts, it had to be subject to five main criteria of scientific validation, that is, whether (i) the particular technique or methodology has been scientifically tested, (ii) its error rates have been established, and (iii) known standards of the technique have been developed and well maintained, (iv) the technique has been peer-reviewed, and (v) the technique has gained broad public acceptance [35]. Forensic evidence based on fingerprints was first challenged under Daubert's ruling in 1999 in the case of USA v. Byron Mitchell [49], stating that the fundamental premise for asserting the uniqueness of fingerprints had not been objectively tested and its matching error rates were unknown. The Brandon Mayfield Case [36] is another case that challenges the reliability of fingerprint. In this case, Brandon Mayfield, a lawyer from Oregon, was mistakenly identified by FBI as the terrorist who attacked the commuter trains in Madrid in March, 2004. In October, 2007, fingerprint evidence was excluded in the case of Bryan Rose for

the conviction of death penalty by a Baltimore County judge [10]. The judge challenged the reliability of fingerprints based on the error made from fingerprint evidence in the Brandon Mayfield case [10]. Fingerprint-based identification has been challenged in more than 20 court cases in the United States (see [9] for details). It is clear that there is a need to study fingerprint individuality which is the basis for the admissibility of fingerprint evidence in court cases.

2.2 Early Studies on Fingerprint Individuality

There have been a few previous studies that addressed the problem of fingerprint individuality using mathematical models on fingerprint features. All these studies utilized minutiae features (both location and direction information) to assess individuality. In 1892, Galton [16] raised the topic of fingerprint individuality for the first time. He assumed that a full fingerprint is a combination of 24 disjoint and independent square regions, each of which consists of six ridges. He found that the probability of correctly re-building each of the 24 regions by only looking at its neighboring ridges is $1/2$. Thus the probability of correctly re-building all of the 24 regions is $(1/2)^{24}$. Further, since the probability of finding a specific fingerprint class (i.e., arch, whorl, left loop, right loop, double loop) is $1/16$, and the probability of finding the correct number of ridges entering and exiting each region is $1/256$, the probability of finding a set of given neighboring ridges is $(1/16) \times (1/256)$. Based on these assumptions, Galton estimated that the probability of finding each fingerprint configuration in a given population is

$$P(\texttt{Fingerprint configuration}) = \frac{1}{16} \times \frac{1}{256} \times (\frac{1}{2})^{24} = 1.45 \times 10^{-11}. \quad (2.1)$$

Later, Pearson [41] and Kingston [23] disputed Galton's assumptions and suggested that, given the fact that there are 36 possible minutiae locations within a six-ridge block, the probability of observing a given ridge configuration in one block is actually $1/36$, instead of $1/2$ as proposed by Galton. Therefore, the probability of fingerprint configuration should

be

$$P(\texttt{Fingerprint configuration}) = \frac{1}{36} \times \frac{1}{256} \times (\frac{1}{36})^{24} = 1.09 \times 10^{-41}. \quad (2.2)$$

After Pearson and Kingston, several subsequent models (e.g., Henry [18], Balthazard [3], Bose [47], Wentworth and Wilder [50], and Cummins and Midlo [11]) of fingerprint individuality were proposed based on the number of minutiae in a fingerprint and the probability (p) of occurrence of a minutia. The probability of a fingerprint configuration with N minutiae is given by the general formula

$$P(\texttt{Fingerprint configuration}) = p^N, \quad (2.3)$$

with different values for p used in different studies. For example, Henry estimated p as $1/4$, and thus, for a given fingerprint type and a given number of ridges between core and delta, Henry determined the probability of fingerprint configuration to be p^{N+2}. By contrast, Wentworth and Wilder [50] chose p to be $1/50$, and Cummin and Midlo [11] chose p to be $1/31$.

The above investigations as well as many other later studies made assumptions that were not validated on actual fingerprint databases. This is a serious drawback since estimates of fingerprint individuality obtained by these studies were also never validated. The first attempt to validate a fingerprint individuality model on an actual database was carried out by Pankanti et al. [35].

2.3 A Stochastic Model of Fingerprint Individuality

A significant improvement on earlier models of fingerprint individuality was reported by Pankanti et al. [35]. Since this thesis makes an effort to improve this model, the model by Pankanti et al. [35] is presented in this section in detail .

Pankanti et al. [35] estimated fingerprint individuality via probability of random correspondence (PRC), which is defined as the probability that two different fingerprints from a

target population randomly match each other. Suppose the query fingerprint Q has n "effective" minutiae and the template T has m "effective" minutiae, where "effective" minutiae indicates minutiae in the overlapping region of two fingerprints after alignment. A more natural definition of PRC which is utilized in this thesis is the probability of match when Q and T have n and m minutiae, respectively, in the whole fingerprint instead of in the overlapping region. Recall the hypothesis testing scenario of Equation 1.1 for biometric authentication. When the similarity measure $S(Q,T)$ is above the threshold λ, the claimed identity (I_c) is accepted as the true identity I_t. Based on the statistical hypothesis in Equation 1.1 in Chapter 1, the PRC is defined as the false acceptance rate, which is

$$PRC(\lambda) = P(S(Q,T) \geq \lambda | m, n, H_0), \tag{2.4}$$

where the probability is computed under the assumption that H_0 is true.

To estimate PRC, the following assumptions were made:

1. Only minutia ending and bifurcation are considered as salient fingerprint features for matching. No distinction was made between minutia ending and bifurcation. Other types of minutiae, such as islands, spur, crossover, lake, etc., rarely appear and can be thought of as combination of endings and bifurcations.

2. Minutiae location and direction are uniformly distributed and independent of each other. Further, minutiae locations can not occur very close to each other.

3. Different minutiae correspondences between Q and T are independent of each other, and any two correspondences are equally important.

4. All minutiae are assumed true, that is there are no missed or spurious minutiae.

5. Ridge width is unchanged across the whole fingerprint.

6. Alignment between Q and T exists, and can be uniquely determined. To align two sets of minutiae, corresponding minutiae were first determined by a matching algorithm [1]. A rigid transformation consisting of rotation and translation was then

determined with a least square approximation between the corresponding minutiae pairs [35].

2.3.1 Definition of a Random Minutiae Correspondence

For a query fingerprint Q with n minutiae, Pankanti et al. [35] estimated the probability of Q sharing q minutiae out of the m minutiae in template T. Use the same letters Q and T to denote the minutiae sets in fingerprint Q and template T. These minutiae can be then expressed as

$$Q \doteq \{\{S_1^Q, D_1^Q\}, \{S_2^Q, D_2^Q\}, ..., \{S_n^Q, D_n^Q\}\} \tag{2.5}$$

$$T \doteq \{\{S_1^T, D_1^T\}, \{S_2^T, D_2^T\}, ..., \{S_m^T, D_m^T\}\}, \tag{2.6}$$

where S and D refer to a generic minutia location and direction pair. Assume that the minutiae in Q have been aligned with minutiae in T. To assess fingerprint individuality, a random minutiae correspondence between Q and T needs to be defined: a minutia in Q, (S^Q, D^Q), is said to match or be in correspondence with a minutia in T, (S^T, D^T), if for fixed positive numbers r_0 and d_0, the following inequalities are valid:

$$|S^Q - S^T|_s \leq r_0 \quad \text{and} \quad |D^Q - D^T|_d \leq d_0, \tag{2.7}$$

where

$$|S^Q - S^T|_s \equiv \sqrt{(x^Q - x^T)^2 + (y^Q - y^T)} \tag{2.8}$$

is the Euclidean distance between the minutiae locations $S^Q \equiv (x^Q, y^Q)$ and $S^T \equiv (x^T, y^T)$,

and

$$|D^Q - D^T|_d \equiv \min(|D^Q - D^T|, 2\pi - |D^Q - D^T|) \tag{2.9}$$

is the angular distance between the minutiae directions D^Q and D^T.

The choice of parameters r_0 and d_0 defines a tolerance region (see Figure 2.1), which is critical in determining a match according to Equation 2.7. Large (respectively, small)

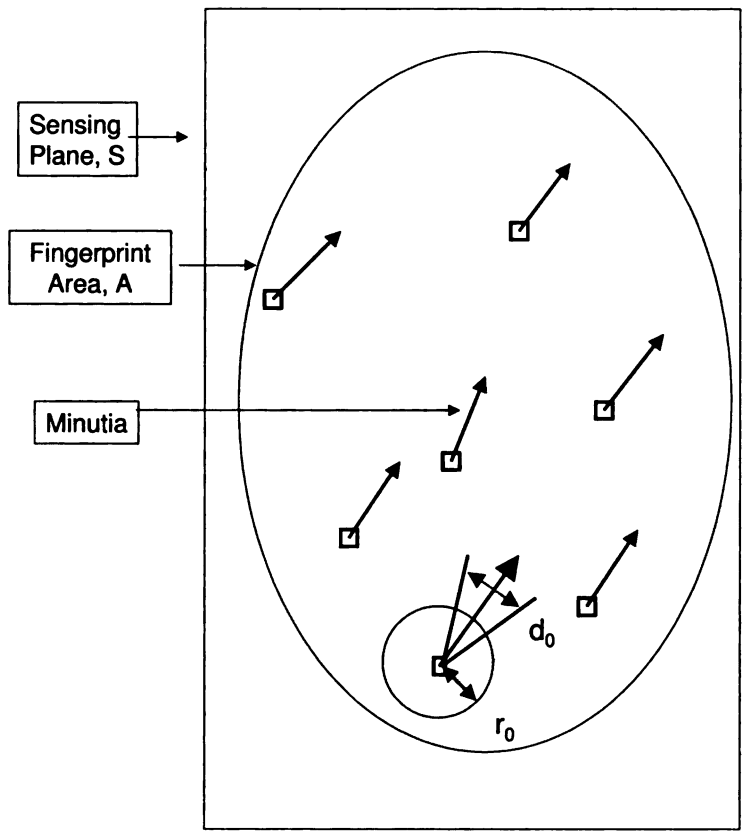


Figure 2.1: Identifying the tolerance region for a query minutia.

values of the pair (r_0, d_0) will lead to spurious (missed) minutiae matches. Thus, it is necessary to select (r_0, d_0) judiciously so that both kinds of matching errors are minimized. A discussion on how to select (r_0, d_0) is given subsequently.

Parameters (r_0, d_0) determine the matching region for a query minutia. In the ideal situation, a genuine pair of matching minutiae in the query and template will correspond exactly, which leads to the choice of (r_0, d_0) as $(0, 0)$. However, factors such as skin elasticity and non-uniform fingertip pressure can cause the minutiae pair that is supposed to perfectly match, to slightly deviate from one another. To avoid rejecting such pairs as non-matches, non-zero values of r_0 and d_0 need to be specified for matching pairs of genuine minutiae.

The value of r_0 is determined based on the distribution of the Euclidean distance between every pair of matched minutiae in the genuine case. To find the corresponding pairs of minutiae, pairs of genuine fingerprints were aligned, and Euclidean distance between each of the genuine minutiae pairs was then calculated. The value of r_0 was selected so that only the upper 5% of the genuine matching location distances (corresponding to large values of r) were rejected. In a similar fashion, the value of d_0 was determined to be the 95th percentile of the distribution of genuine matching angular distances (i.e., the upper 5% of the genuine matching angular distances were rejected).

To find the actual r_0 and d_0, Pankanti et al. [35] used a database of 450 mated fingerprint pairs from IBM ground truth database [38]. The true minutiae locations in this database and the minutiae correspondences between each pair of genuine fingerprints in the database were determined by a fingerprint expert. Using the ground truth correspondences, r_0 and d_0 were estimated to be 15 and 22.5, respectively. These same values will be used to estimate the PRC in the experiments presented in this thesis.

2.3.2 Estimation of Fingerprint Individuality by Uniform Model

In the model by Pankanti et al. [35], the similarity measure between the query Q and the template T, namely $S(Q,T)$, is taken as the number of common minutiae between Q and T. When $S(Q,T)$ is above a threshold w, the claimed identity (I_c) is accepted as true identity I_t. Given Q with n minutiae and T with m minutiae, Pankanti et al. [35] evaluated fingerprint individuality by measuring the probability of finding exactly q matches between them.

Let A be the overlapping area between Q and T, and r_0 and d_0 be the estimated threshold for the tolerance region. Moreover, define $C = \pi {r_0}^2$. Pankanti et al. [35] claimed that considering any subset of ρ minutiae in Q, the probability that all of these ρ minutiae and

only these ρ minutiae have correspondences in location (in whatever direction) in T is

$$\frac{mC}{A} \cdot \frac{(m-1)C}{A-C} \cdot \ldots \cdot \frac{(m-\rho+1)C}{A-(\rho-1)C} \times \frac{A-mC}{A-\rho C} \cdot \frac{A-(m-1)C}{A-(\rho+1)C} \cdot \ldots \cdot \frac{A-(m-(n-\rho+1))C}{A-(n-1)C}. \tag{2.10}$$

The probability in Equation 2.10 can be derived sequentially as follows. First, let $\{S_1, S_2, ..., S_\rho\}$ be locations of the ρ selected minutiae in Q. Then, the probability that S_1 is matched with one of the m minutiae in the template T is

$$\frac{mC}{A}.$$

Moreover, given that S_1 is matched with a minutia in T, the probability that S_2 is matched with one of the remaining $m-1$ minutiae in T is

$$\frac{(m-1)C}{A-C}.$$

Recall that the second assumption of their model is that minutiae can not be very close to each other. After introduction of the tolerance region by r_0, this assumption can be restated as follows: The minimum distance between any pair of minutia locations in a fingerprint is larger than $2r_0$. Therefore, S_2 cannot correspond with the same minutiae of T as that of S_1. Similarly, given that all of $\{S_1, S_2, ..., S_{k-1}\}$ find their matches in T, the probability that S_k ($k \leq \rho$) finds a match in T is

$$\frac{(m-k+1)C}{A-(k-1)C}.$$

Furthermore, given that all of the ρ minutiae have found their correspondences in T, the probability for minutiae S_k ($\rho < k \leq n$) not to match with any minutiae in T is (recall that only these ρ minutiae have correspondences in T)

$$\frac{A-(mC-(k-\rho+1))C}{A-(k-1)C}.$$

Combining these steps gives Equation 2.10.

Equation 2.10 gives the probability of ρ matches between Q and T for a given set of ρ minutiae, $\{\{S_1, D_1\}, \{S_2, D_2\}, ..., \{S_\rho, D_\rho\}\}$, in Q. However, in practice, it is more important to find the probability for any set of ρ minutiae in Q. Since there are $\binom{n}{\rho}$ ways to select ρ minutiae from Q, the probability of finding exactly ρ minutiae pairs between Q and T matched in location is

$$\begin{aligned} P(A, C, m, n, \rho) &= \binom{n}{\rho} \cdot \frac{mC}{A} \cdot \frac{(m-1)C}{A-C} \cdot ... \frac{(m-\rho+1)C}{A-(\rho-1)C} \\ &\times \frac{A-mC}{A-\rho C} \cdot \frac{A-(m-1)C}{A-(\rho+1)C} \cdot ... \cdot \frac{A-(m-(n-\rho+1))C}{A-(n-1)C}. \end{aligned} \tag{2.11}$$

Equation 2.11 can be further simplified. Define $M = \frac{A}{C}$. Since M is large (this is because C is much smaller compared to the fingerprint area, A), it is realistic to take M as an integer. Thus Equation 2.11 approaches the probability of a hyper-geometric distribution:

$$P(M, m, n, \rho) = \frac{\binom{m}{\rho} \times \binom{M-m}{n-\rho}}{\binom{M}{n}}. \tag{2.12}$$

The model introduced above considers only minutiae matches in location, and ignores minutiae direction. To account for direction, Pankanti et al. [35] introduced a binomial model. They assumed that minutiae direction is independent of minutiae location, and therefore matching minutiae in location and in direction are also independent. They further assumed that minutiae direction is uniformly distributed in $[0, 2\pi]$. Defining

$$l = P(|D^Q - D^T| \leq d_0) = \frac{\theta_0}{\pi},$$

the probability that there are q pairs of minutiae matched in direction follows a binomial distribution:

$$\binom{\rho}{q} l^q (1-l)^{\rho-q}.$$

Based on the hyper-geometric distribution for minutiae match in location and the binomial distribution for minutiae match in direction, the probability that there are ρ matches in location and q matches in both location and direction ($q \leq \rho$) between Q and T is

$$P(M, m, n, l, \rho) = \frac{\binom{m}{\rho}\binom{M-m}{n-\rho}}{\binom{M}{n}} \times \binom{\rho}{q} l^q (1-l)^{\rho-q}. \tag{2.13}$$

Therefore, the probability of q minutiae matches in both location and direction is the sum of Equation 2.13 over ρ, i.e.,

$$P(M,m,n,q) = \sum_{\rho=q}^{min\{m,n\}} \frac{\binom{m}{\rho}\binom{M-m}{n-\rho}}{\binom{M}{n}} \times \binom{\rho}{q} l^q (1-l)^{\rho-q}. \tag{2.14}$$

2.3.3 Corrected Uniform Model

Pankanti et al. [35] validated their stochastic model on various databases. However, the model predictions deviated significantly from empirical results obtained through an automatic fingerprint matching system [35]. This is mainly because the assumption of uniform distribution on minutiae location and direction does not hold true in practice. For example, it is known that fingerprint minutiae tend to form clusters [44], and minutiae only occur on fingerprint ridges instead of valleys. Therefore minutiae locations are not uniformly distributed. Moreover, minutiae in different regions of a fingerprint are observed to be associated with different region-specific minutiae directions. Hence minutiae directions are neither uniformly distributed nor independent of the location. Furthermore, minutiae points that are spatially close to each other tend to have similar directions. These observations on the distribution of fingerprint minutiae need to be accounted for in eliciting reliable statistical models for fingerprint individuality.

Pankanti et al. [35] improved their model to better fit the empirical results. First, to account for non-homogeneity of minutiae location, they redefined the parameter M in their model as

$$\frac{A}{2r_0\omega},$$

where ω is the ridge width. This definition deviates from their uniform assumption of minutiae location.

Second, when evaluating the parameter l, Pankanti et al. [35] found that instead of using

$$l = \frac{2 \times 22.5}{360} = 0.125,$$

as derived based on the uniform distribution, they instead used

$$l = 0.267,$$

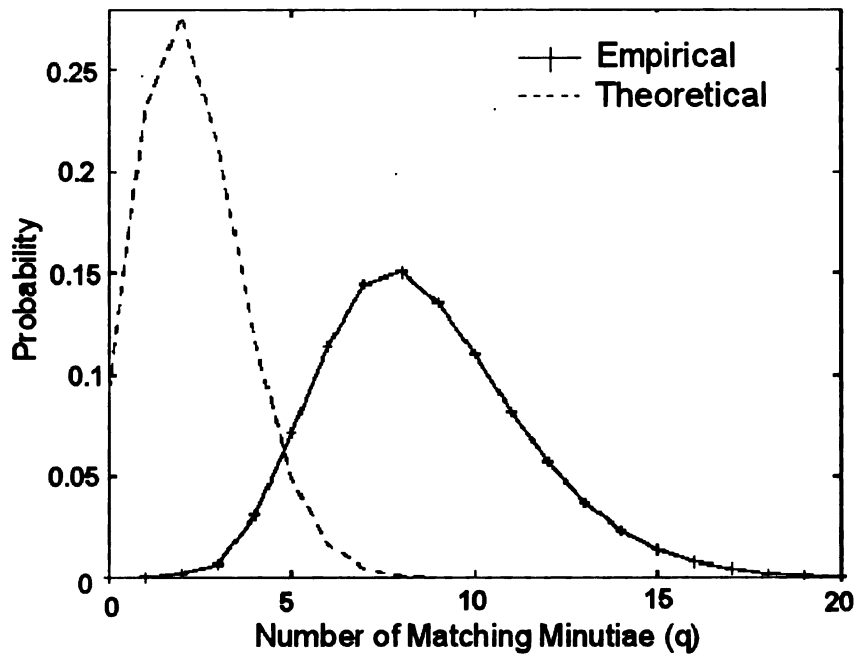
based on empirical results on real databases which again deviates from their uniform assumption minutiae directions.

2.3.4 Limitations of Corrected Uniform Model

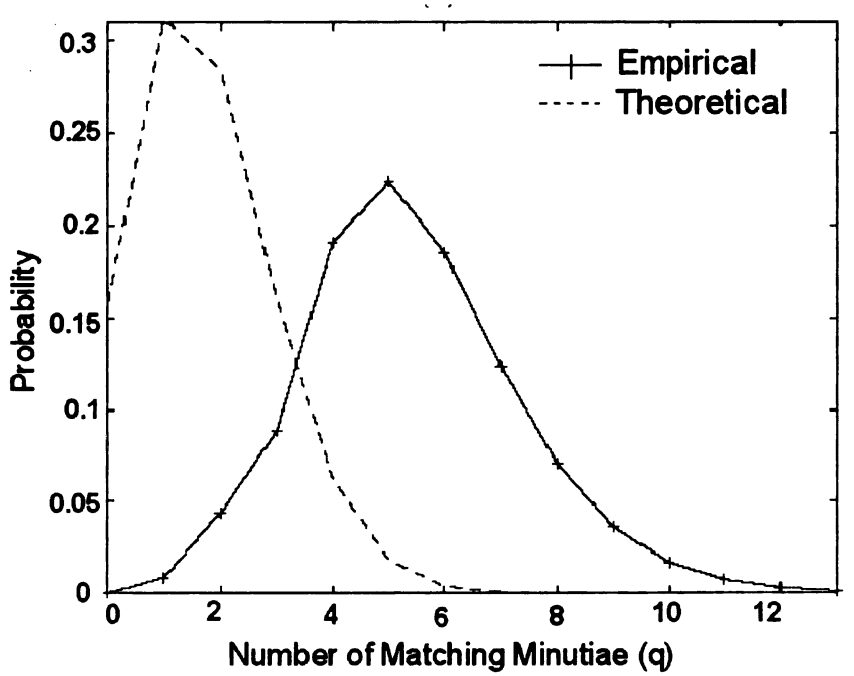
A comparison between model predictions and empirical observations in Figure 2.2 [35] based on two databases, MSU DBI and MSU VERIDICOM [35], showed that the corrected uniform model grossly underestimated the probabilities. In this figure, there are two different probability distributions of the number of matched minutiae pairs for each database, namely the empirical distribution and the theoretical distribution. The empirical distribution is obtained through an automatic fingerprint matching system (AFMS) [1] and theoretical distribution is computed from the corrected uniform model when M is taken as the average value estimated from the database. As seen from the figures, the empirical distributions are to the right of their corresponding theoretical distributions which indicates that the corrected uniform model grossly underestimates the PRCs. This is mainly because the corrected uniform model didn't model the minutiae clustering tendency and the dependence between minutiae locations and directions.

A comparison of PRCs derived from their model with empirical PRCs based on NIST 2002 SD4 reached the same conclusion. For example, for a query fingerprint Q and a template fingerprint T, each with 52 minutiae, the empirical probability that they share 12 or more minutiae is 3.9×10^{-3}, differing greatly from the model estimation of 4.3×10^{-8} (See Table 7.11 based on the estimates from NIST 2000 SD 4 [30]).

The inherent limitation of the corrected uniform model motivated the research presented in this thesis. The thesis develops statistical models that are significant improvements over the model by Pankanti et al. [35].



(a)



(b)

Figure 2.2: Comparison of experimental and theoretical probabilities for the number of matching minutiae: (a) MSU DBI database, and (b) MSU VERIDICOM database. Figure is the reproduction of Figure 9 in Pankanti et al. [35]

2.4 Contributions of the Thesis

The uncertainty involved in assessing fingerprint individuality can be quantified as the probability of finding a fingerprint in a target population having minutiae similar to that of a given query fingerprint. This probability is also known as the probability of random correspondence (PRC). To compute this probability, fingerprint samples from a target population are collected first. Then the variability of the minutiae from various fingerprints should be analyzed. After that, a notion of similarity between a pair of fingerprints and the probability that two different individuals share a set of similar minutiae should be defined. In this thesis, it is assumed that a sample of prints is available from a target population and a notion of similarity is given, and thus it does not address the issues and challenges involved in sampling from a target population, which is worthy of separate investigations. Instead, this thesis demonstrates how the proposed methodology can be used to obtain estimates of fingerprint individuality based on the database that is assumed to be available. If the available database is representative of the target population, the proposed estimates will generalize to the target population. Figure 2.3 shows the structure of the thesis.

To address the issue of individuality, candidate models must meet two important requirements: (i) flexibility, i.e., the elicited model can represent the minutiae distribution from a variety of fingerprints in the target population, and (ii) computational efficiency, i.e., associated measures of fingerprint individuality can be easily obtained from the model. In practice, a forensic expert uses many fingerprint features, such as minutiae location and direction, fingerprint class, inter-ridge distance, to match a pair of fingerprints. In this thesis, we only use minutiae locations and directions for simplicity. Although Pankanti et al. [35] provided a stochastic model based on the same set of features (as discussed in Section 2.3), their model cannot satisfactorily represent minutiae variability because the uniformity assumption of minutiae location and direction disagrees with observations from empirical studies.

Empirical studies suggest that minutiae tend to cluster together, and minutiae close to

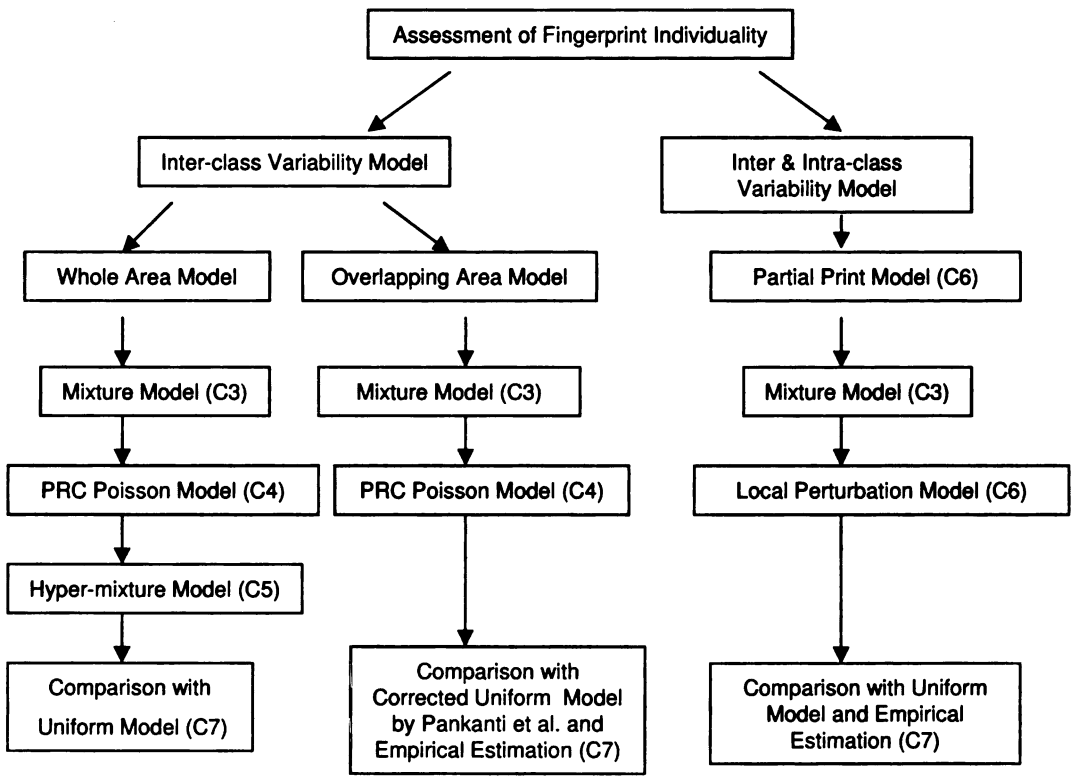


Figure 2.3: Outline of the thesis contributions where the labels Cks indicate the chapters where the corresponding contribution is made.

each other share similar orientations (in similar or almost opposite directions), which, in turn, implies that minutiae locations and directions are dependent. To account for these minutiae characteristics, a family of mixture models is proposed to represent the observed distribution of minutiae (both location and direction) in Chapter 3. A mathematical model for the PRC is derived in Chapter 4 and an approximation formula is derived to improve computational efficiency. The approximate PRC follows a Poisson distribution, and thus the corresponding model is called Poisson model.

The PRCs obtained in this thesis and those from Pankanti et al. [35] are estimated in different scenarios, which makes it difficult to compare these two models. To carry out this model-comparison, an overlapping area model was also developed, which adopts the definition of PRC proposed by Pankanti et al. [35], and which still estimates the PRC with the mixture models developed in this thesis.

Estimation of fingerprint individuality for a target population involves averaging over all pairwise impostor fingerprints in a sample database. For a reasonably large database, to calculate PRCs for all impostor pairs is infeasible. To solve this problem, a hyper-mixture model is developed in Chapter 5, assuming that the target population is composed of clusters of mixture distributions. The model conveniently estimates the PRC by the clusters without dealing with the individual mixture distributions, which greatly improves the computational efficiency. Finally, experimental results that compare hyper-mixture model and the uniform model are given in Chapter 7.

Between the two sources of minutiae variability, namely interclass and intraclass variabilities, the intraclass variability has not been thoroughly investigated by previous research on fingerprint individuality. While the mixture and hyper-mixture models still only deal with minutiae interclass variability, a compound stochastic model is developed in Chapter 6 to account for both interclass variability and intraclass variability. In particular, we address the two most important types of intraclass variability: (i) variability due to local perturbations arising from non-linear distortions in multiple impressions of the same finger,

and (ii) variability due to partial prints in multiple acquisitions of the same finger. Therefore, compared with the mixture model and hyper-mixture model, the compound stochastic model is a more realistic model, accounting for both interclass and intraclass variability.

2.5 Summary

Fingerprint individuality studies uniqueness of fingerprints which is the premise for fingerprint evidence. This problem has not been extensively studied and as a consequence, fingerprint evidence has been challenged in multiple court cases in the last two decades. Although there is some existing research on fingerprint individuality, almost all early researchers developed their models based on assumptions that were not validated on actual databases.

The first study validating model assumptions on an actual database was presented by Pankanti et al. [35]. In their model, fingerprint individuality was assessed by probability of random correspondence, which is the false acceptance rate of fingerprint recognition. Their model made two simplifying assumptions, (i) the uniformity of minutiae locations and directions and, (ii) independence between minutiae location and direction. Pankanti et al. [35] did modify their hyper-geometric model to account for the non-uniformity of minutiae locations and directions. However, fundamentally, this model is still based on the uniform assumptions. Furthermore, it does not consider the dependence between minutiae location and direction. As a result, experimental results showed that the hyper-geometric model grossly under-estimates fingerprint individuality even after the empirical modification (see figure 2.2).

Section 2.4 summarizes the contributions by this thesis.The proposed model gives a better estimate of the fingerprint individuality as supported by experimental results.

CHAPTER 3

Mixture Models for Fingerprint Minutiae

In this chapter, a mixture model is proposed to model the minutiae variability of a fingerprint. For each fingerprint, this mixture model captures the minutiae clustering tendencies in different fingerprint regions and the dependence between minutiae locations and directions.

3.1 Features for Fingerprint Individuality Model

Minutiae are utilized as the features for fingerprint matching by forensic experts and most automatic fingerprint matching systems. In this research, only the two dominant types, minutia bifurcation and minutia ending are considered. Minutia bifurcation and ending are not distinguished since it is often not easy to distinguish between them by automatic systems and they can convert between each other under noisy environment during the capturing procedure. Each minutia is characterized by its location and direction. Subsequently, the term minutiae features will be used to refer to the location and direction of a minutia in a fingerprint impression.

3.2 Mixture Models

Let S denote a generic random minutiae location and D denote its corresponding direction. Let $\Lambda \subseteq \mathbf{R^2}$ denote the subset of the plane representing the fingerprint domain. It follows that $S \in \Lambda$ and $D \in [0, 2\pi)$. Further, denote the total number of minutiae in the fingerprint region by k. A joint distribution model for the k minutiae features $\{ (S_j, D_j),\ j = 1, 2, \ldots k \}$ is proposed to account for (i) clustering tendencies (i.e., non-uniformity) of minutiae, and (ii) dependence between minutiae location (S_j) and direction (D_j) in different regions of Λ.

The joint distribution model proposed is a mixture model consisting of G components. Let c_j, $c_j \in \{1, 2, \ldots, G\}$, be the cluster label of the j-th minutia, $j = 1, 2, \ldots, k$. The labels c_j's are identically and independently distributed according to a multinomial distribution with G classes, and the probabilities of the G classes (i.e., $\tau_1, \tau_2, \ldots, \tau_G$) satisfy $\tau_j \geq 0$ and $\sum_{j=1}^{G} \tau_j = 1$. Given label $c_j = g$, we assume the minutiae location S_j is distributed according to the density

$$f_g^S(\, s \,|\, \mu_g,\, \Sigma_g \,) = \phi_2(\, s \,|\, \mu_g,\, \Sigma_g \,), \tag{3.1}$$

where ϕ_2 is the bivariate Gaussian density with mean μ_g and covariance matrix Σ_g. Equation 3.1 states that the minutiae locations arising from the g-th cluster follow a two-dimensional Gaussian distribution with mean μ_g and covariance matrix Σ_g.

The Von-Mises distribution [28] is a typical distribution used to model angular random variables, and we adopt it to model minutiae directions. We elicit the distribution D_j given $c_j = g$ to be the density

$$f_g^D(\theta \,|\, \nu_g, \kappa_g,\, p_g) = p_g\, v(\theta) \cdot I\{0 \leq \theta < \pi\} + (1 - p_g)\, v(\theta - \pi) \cdot I\{\pi \leq \theta < 2\pi\}, \tag{3.2}$$

where $I\{A\}$ is the indicator function of A (i.e., $I\{A\} = 1$ if A is true; and $I\{A\} = 0$, otherwise), and $v(\theta)$ is the Von-Mises distribution given by

$$v(\theta) \equiv v(\, \theta \,|\, \nu_g,\, \kappa_g) = \frac{2}{I_0(\kappa_g)} \exp\{\kappa_g \cos 2(\theta - \nu_g)\}, \tag{3.3}$$

with $I_0(\kappa_g)$ defined as

$$I_0(\kappa_g) = \int_0^{2\pi} \exp\{\kappa_g \cos(\theta - \nu_g)\}\, d\theta. \tag{3.4}$$

In Equation 3.3, ν_g and κ_g represent the mean angle and the precision (inverse of the variance) of the Von-Mises distribution, respectively. Figure 3.1 plots two density functions associated with Von-Mises distributions with common means ν_g but with two different precisions $\kappa_g < \kappa_g^*$. The figure shows that ν_g represents the "center" (or modal value) while κ_g controls the degree of spread around the center (thus, the density with precision κ_g^* has higher concentration around ν_g). The density f_g^D in Equation 3.2 can be interpreted in the following way: The ridge flow orientation, ω, is assumed to follow the Von-Mises distribution in Equation 3.3 with mean ν_g and precision κ_g. Subsequently, minutiae arising from the g-th component have directions that are either ω or $\omega + \pi$ with probabilities p_g and $1 - p_g$, respectively.

Combining the distributions of the minutiae location (S) and direction (D), it follows that each (S, D) is distributed according to the mixture density

$$f(s, \theta \mid \Theta_G) = \sum_{g=1}^{G} \tau_g\, f_g^S(s \mid \mu_g, \Sigma_g) \cdot f_g^D(\theta \mid \nu_g, \kappa_g), \tag{3.5}$$

where $f_g^S(\cdot)$ and $f_g^D(\cdot)$ are defined in Equations 3.1 and 3.2, respectively.

In Equation 3.5, Θ_G denotes all the unknown parameters in the mixture model which includes the total number of mixture components, G, the mixture probabilities τ_g, $g = 1, 2, \ldots, G$, the component means and covariance matrices of f_g^S's given by $\boldsymbol{\mu}_G \equiv \{\mu_1, \mu_2, ..., \mu_G\}$ and $\boldsymbol{\Sigma}_G \equiv \{\Sigma_1, \Sigma_2, ..., \Sigma_G\}$, the component mean angles and precisions of f_g^D's given by $\boldsymbol{\nu}_G \equiv \{\nu_1, \nu_2, \ldots, \nu_G\}$ and $\boldsymbol{\kappa}_G \equiv \{\kappa_1, \kappa_2, \ldots, \kappa_G\}$, and the mixing probabilities $\boldsymbol{p}_G \equiv \{p_1, p_2, \ldots, p_G\}$. The model described by Equation 3.5 has three advantages: (i) it allows for clustering tendencies in minutiae locations and directions via G different clusters, (ii) it incorporates dependence between minutiae location and direction since if S_j is known to come from the g-th component, the direction D_j also comes

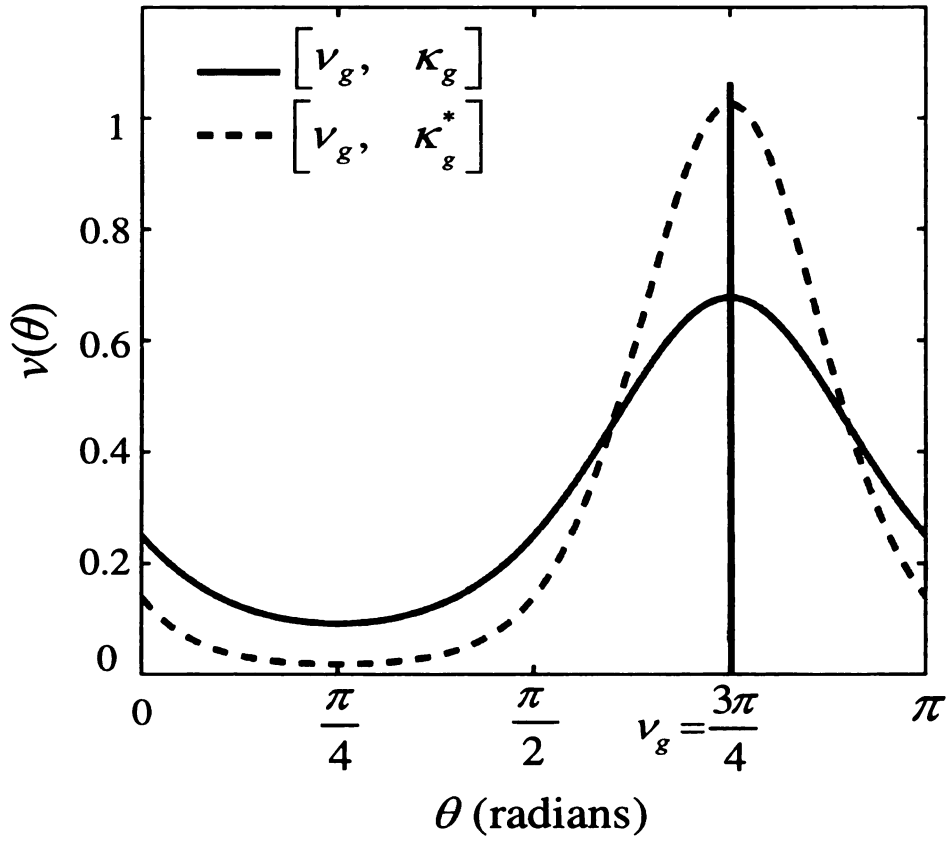


Figure 3.1: Probability distribution plots of the Von-Mises distribution with center $\nu_g = 3\pi/4$, and with two different precisions, κ_g and κ_g^*, with $\kappa_g < \kappa_g^*$. The values of $v(\theta)$ at 0 and π are equal to each other due to the cyclical nature of the cosine function.

from the g-th component, and (iii) it is flexible because it allows for two sub-components with almost opposite minutiae directions in each component which is the novel part of this mixture model.

The mixture density given in Equation 3.5 is defined on the entire plane $\mathbf{R}^2$, and is not restricted to the fingerprint domain Λ. The constrained mixture model on the fingerprint domain is

$$f_\Lambda(s, \theta \mid \Theta_G) = \frac{f(s, \theta \mid \Theta_G)}{\int_{s \in \Lambda} \int_{\theta=0}^{2\pi} f(s, \theta \mid \Theta_G)\, d\theta\, ds}. \tag{3.6}$$

If the fingerprint area Λ is large, it follows that $\Lambda \approx R^2$ and,

$$f_\Lambda(s, \theta \mid \Theta_G) \approx f(s, \theta \mid \Theta_G) \tag{3.7}$$

because the denominator in (3.6)

$$\int_{s \in \Lambda} \int_{\theta=0}^{2\pi} f(s, \theta \mid \Theta_G)\, d\theta\, ds \approx 1. \tag{3.8}$$

To estimate the unknown parameters in the model, an algorithm is developed based on the expectation maximization (EM) algorithm [14, 31]. The optimal number of components, G^*, is selected using the Bayes Information Criteria (BIC). The BIC [43] has been widely used in various model selection problems, and has the property that it selects a model that is most parsimonious (with the least number of model parameters). Details of the EM algorithm and the BIC are given in the next section.

The extension of mixture models to minutiae direction is a novel contribution. In each component in the mixture model, as shown in Equation 3.5, $f_g^D(\theta \mid \nu_g, \kappa_g, p_g)$ is, according to Equation 3.2, a sum of two weighted Von-mises distributions for minutiae direction. The mean angles of the two Von-Mises distributions in Equation 3.2, i.e., $v(\theta)$ and $v(\theta - \pi)$, are different by π, capturing two sub-groups of minutiae in each component with almost opposite minutiae directions. This is motivated by the fact that neighboring minutiae tend to have similar orientations, i.e., with similar or almost opposite directions.

3.3 EM algorithm for Estimating Θ_G

The Expectation Maximization (EM) algorithm [14] is a well-known iterative method for finding the maximum likelihood estimate of parameters either in the presence of missing data or when the model can be simplified by adding latent variables. In such cases, the original (or observed) likelihood can be obtained by marginalizing a complete likelihood over the missing or latent variables. The EM algorithm consists of two main steps:

1) the E-step, where the expectation of the logarithm of the complete likelihood is obtained conditional on the observed data and parameter estimates at the current iteration step, and

2) the M-step, where a maximization is performed to update the parameter estimates for the subsequent iteration step.

The E- and M-steps are cycled until the parameter estimates converge. For a more detailed introduction of the EM algorithm, refer to [31].

The k minutiae features (S_j, D_j), $j = 1, 2, \ldots, k$, are assumed to be independent of each other and distributed according to the mixture density in Equation 3.5. In this case, the missing component for the EM algorithm consists of the class labels, c_j, $j = 1, 2, \ldots, k$, corresponding to each of the k features. The transformation

$$\omega_j = \begin{cases} D_j & \text{if } D_j \in [0, \pi) \\ D_j - \pi & \text{if } D_j \in [\pi, 2\pi) \end{cases} \tag{3.9}$$

converts the minutiae directions into orientations which take values in $[0, \pi)$. The corresponding distribution for each (S_j, ω_j) then becomes

$$\sum_{g=1}^{G} \tau_g \, f_g^S(S_j \mid \mu_g, \Sigma_g) \cdot f_g^O(\omega_j \mid \nu_g, \kappa_g), \tag{3.10}$$

where $f_g^O(\omega_j \mid \nu_g, \kappa_g)$ is as given in Equation 3.3. Note that the expression in Equation 3.10 is now in the standard form for mixture models (see, for example, Section 2.7 of [31]),

and can be solved using general formulas for the E- and M-steps. In this case, the E- and M-steps can be combined into a single updating equation for each parameter linking the current estimates to subsequent ones. The estimate of τ_g at the $(n+1)$-th iteration, $\tau_g^{(n+1)}$, is given by

$$\tau_g^{(n+1)} = \frac{1}{k}\sum_{j=1}^{k} z_{gj}^{(n)}, \tag{3.11}$$

where

$$z_{gj}^{(n)} \equiv P(c_j = g \mid S_j, \omega_j, \Theta_G^{(n)}) \tag{3.12}$$

is the posterior probability that the j-th observation is from the g-th class, conditioned on S_j, ω_j and the parameter estimates at the n-th iteration $\Theta_G^{(n)}$.

The estimates of μ_g and Σ_g at the $(n+1)$-th iteration, $\mu_g^{(n+1)}$ and $\Sigma_g^{(n+1)}$, respectively, are given by the equations

$$\mu_g^{(n+1)} = \frac{\sum_{j=1}^{k} z_{gj}^{(n)} S_j}{\sum_{j=1}^{k} z_{gj}^{(n)}} \tag{3.13}$$

and

$$\Sigma_g^{(n+1)} = \frac{\sum_{j=1}^{k} z_{gj}^{(n)} (S_j - \mu_g^{(n+1)})(S_j - \mu_g^{(n+1)})^T}{\sum_{j=1}^{k} z_{gj}^{(n)}}, \tag{3.14}$$

where $z_{gj}^{(n)}$ is as defined in Equation 3.12.

We next proceed to estimate the parameters ν_g and κ_g at the $(n+1)$-th iteration. The component of the complete log-likelihood function (after the E-step of the $(n+1)$-th iteration) involving only the parameters ν_g and κ_g is given by

$$\sum_{j=1}^{k} z_{gj}^{(n)} \log\{f_g^O(\omega_j \mid \nu_g, \kappa_g)\} = \sum_{j=1}^{k} z_{gj}^{(n)} \left\{ \kappa_g \cos 2(\omega_j - \nu_g) + \log\left\{ \frac{2}{I_0(\kappa_g)} \right\} \right\}. \tag{3.15}$$

Differentiating with respect to ν_g and setting the derivative to zero, the estimate of ν_g at the $(n+1)$-th step, $\nu_g^{(n+1)}$, satisfies the equation

$$\sum_{j=1}^{k} z_{gj}^{(n)} \sin 2(\omega_j - \nu_g^{(n+1)}) = 0, \tag{3.16}$$

which can be solved to give the closed form solution

$$\nu_g^{(n+1)} = \frac{1}{2} \tan^{-1} \left\{ \frac{\sum_{j=1}^{k} z_{gj}^{(n)} \sin 2\omega_j}{\sum_{j=1}^{k} z_{gj}^{(n)} \cos 2\omega_j} \right\}. \tag{3.17}$$

Substituting Equation 3.17 in Equation 3.15, differentiating with respect to κ_g and setting the derivative to zero, we note that the $(n+1)$-th step estimate of κ_g, $\kappa_g^{(n+1)}$, satisfies the equation

$$\frac{I_0'(\kappa_g^{(n+1)})}{I_0(\kappa_g^{(n+1)})} = \frac{\sum_{i=1}^{k} z_{gj}^{(n)} \cos 2(\omega_j - \nu_g^{(n+1)})}{\sum_{i=1}^{k} z_{gj}^{(n)}}. \tag{3.18}$$

The numerical method outlined in [19] is then used to compute $\kappa_g^{(n+1)}$ from Equation 3.18.

The cluster label for observation (S_j, D_j) at the $(n+1)$-th step, $c_j^{(n+1)}$, is determined as

$$c_j^{(n+1)} = \arg\max_g z_{gj}^{(n)}. \tag{3.19}$$

and the estimate of p_g is obtained as

$$p_g^{(n+1)} = \frac{\sum_{j=1}^{k} I\{c_j^{(n+1)} = g\,,\, D_j \in [0, \pi)\}}{\sum_{j=1}^{k} I\{c_j^{(n+1)} = g\}}. \tag{3.20}$$

The E- and M-steps are repeated till the parameter estimates converge. To find the optimal number of clusters (G^*), the EM algorithm was first implemented to estimate the

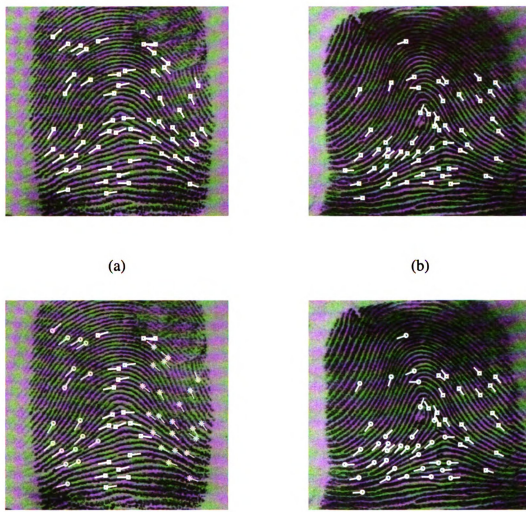
model parameters for different values of G, and the BIC criteria

$$BIC(G) = 2 * \sum_{j=1}^{k} \log f(S_j, D_j \mid \Theta_G) - |\Theta_G| \log(k), \tag{3.21}$$

is used to select G^*, where $|\Theta_G|$ is the cardinality of Θ_G, i.e., the number of unknown parameters in Θ_G, and f is the mixture density as defined in Equation 3.5. For the databases used in this thesis, G was chosen to be less than or equal to 5. Based on the number of minutiae typically observed in the database used here, choosing a larger value of G may lead to model over-fitting. The value of G^* is selected as the value of G that maximizes $BIC(G)$.

Figure 3.2 illustrates the fit of the mixture model to two different fingerprint images from the NIST 2000 SD 4 . Observed minutiae locations (white boxes) and directions (white lines) are shown in panels (a) and (b). Panels (c) and (d), respectively, give the cluster assignment for each minutia feature in (a) and (b). The cluster label of the j^{th} minutiae (S_j, D_j) is estimated according to Equation 3.19 after the EM algorithm has converged. Panels (a) and (b) in Figure 3.3 shows the BIC values for different values of G. When $G = 3$ (or $G = 2$), BIC is maximum for fingerprint in Figure 3.3 (a) (or b). Figures 3.4 (a) and (b) plot the minutiae features in the 3-D (S, D) space for easy visualization of the clusters (in both location and direction). The BIC criteria yields G^* to be 3 and 2 for panels (a) and (b), respectively. Minutiae from the same cluster are labeled with the same shape and number.

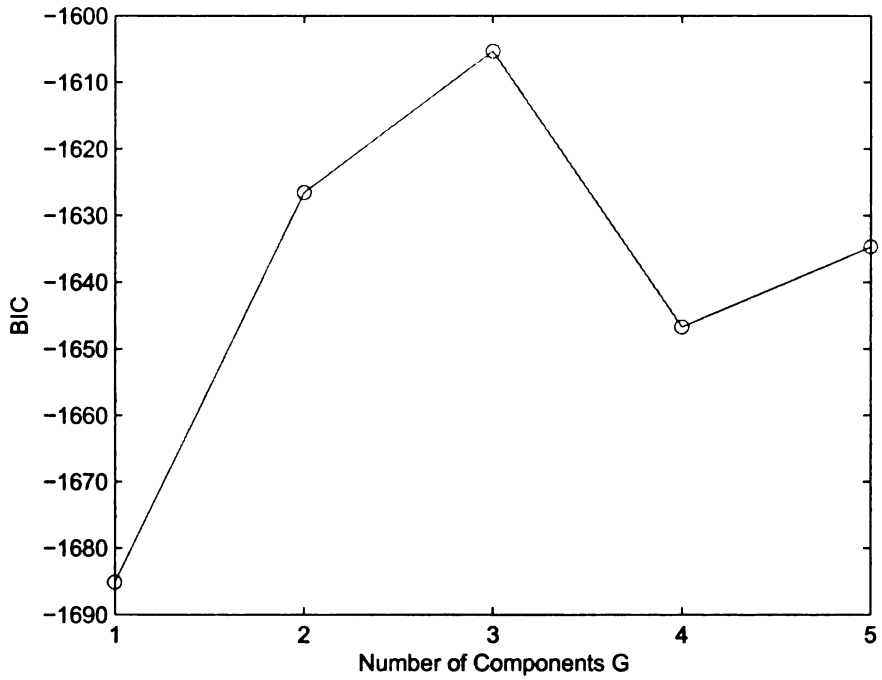
Another way to show the effectiveness of the fit of the models to the observed data is to simulate a minutiae realization from the fitted models. Figures 3.5 (a) and (b) show two fingerprints whose minutiae features were fitted with the mixture distribution in Equation 3.6. Figures (c) and (d) show a simulated realization when both S and D are assumed to have the mixture distributions fitted to (a) and (b), respectively. Figures 3.5 (e) and (f) show a simulated realization when both S and D are assumed to be uniformly distributed and independent of each other. Note that there is a good agreement, in the distributional



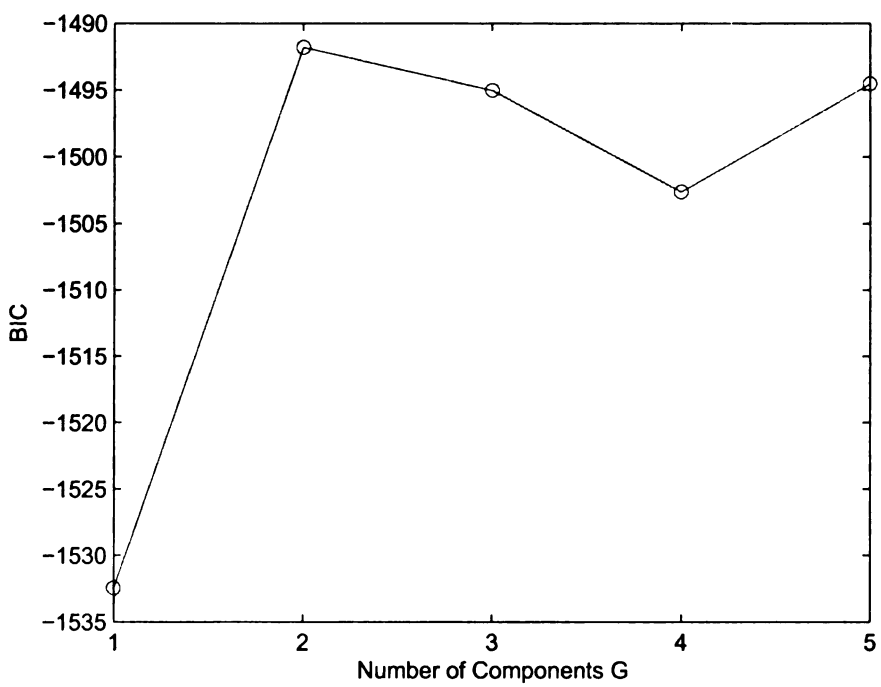
(a) (b)

(c) (d)

Figure 3.2: Assessing the fit of the mixture models to minutiae location and direction: Observed minutiae locations (white boxes) and directions (white lines) are shown in panels (a) and (b) for two different fingerprints from the NIST 2000 SD 4. Panels (c) and (d), respectively, show the cluster labels for each minutia in (a) and (b).



(a)



(b)

Figure 3.3: Assessing the fit of the mixture models to minutiae location and direction observed for fingerprint images (a) and (b) in Figure 3.2. Panels (a) and (b) show the BICs when $1 \leq G \leq 5$.

sense, between the observed minutiae locations and directions [Figures 3.5 (a) and (b)] and those simulated from the proposed models [Figures 3.5 (c) and (d)], but no such agreement exists for the uniform model. [Figures 3.5 (e) and (f)]

3.4 Goodness-of-fit Tests for the Mixture Models

To test the goodness-of-fit of the mixture models to the observed minutiae, the following null and alternative hypotheses have been considered:

$$H_0 : f_T(s,\theta) = f_S(s,\theta \mid \Theta_G) \text{ for some } G \text{ and } \Theta_G, \quad \text{versus} \quad H_1 : \text{not } H_0, \qquad (3.22)$$

where $f_T(s,\theta)$ is the true distribution of minutiae location and direction. For a fingerprint with k minutiae, the above goodness-of-fit test can be carried out by partitioning the (S, D) space into $W^2 \times V$ non-overlapping blocks, which means the space S is partitioned into W equal-size rows by W equal-size columns and the space D is partitioned into v equal-size blocks, and computing

$$\begin{aligned} o_{(w_r,w_c,v)} &= \text{observed number of } (S_j, D_j)\text{'s that fall in the } (w_r, w_c, v)\text{-th block, and} \\ e_{(w_r,w_c,v)} &= k \cdot P((S,D) \in (w_r, w_c, v)\text{-th block} \mid \hat{\Theta}_{G^*}) \\ &= \text{expected number of } (S, D)\text{'s that fall in the } (w_r, w_c, v)\text{-th block under the fitted} \\ &\quad \text{mixture model.} \end{aligned}$$

The two tests discussed below require $e_{w_r,w_c,v}$ to be large for each block (w_r, w_c, v). Let $n_{w_r,w_c,v}$ denote the number of minutiae in block (w_r, w_c, v). For the tests to be valid, the expected frequency for all the blocks should be at least 5 [45]. Consequently, the total number of minutiae in the finger should be at least 5 times the number of blocks. Therefore, a threshold $\tau = 5$ was selected, so that blocks with $e_{w_r,w_c,v} < \tau$ for either the mixture models or the uniform model are combined with neighboring blocks that have $e_{w_r,w_c,v}$ greater than or equal to τ. The set of blocks resulting from this merger is denoted by $\mathcal{B}$.

Two non-parametric test statistics are considered:

(i) the Freeman-Tukey statistic [24] given by

$$\sum_{(w_r,w_c,v)\in\mathcal{B}} \{o_{w_r,w_c,v}^{1/2} + (o_{w_r,w_c,v}+1)^{1/2} - (4\times e_{w_r,w_c,v}+1)^{1/2}\}^2, \quad (3.23)$$

and

(ii) the Chi-square statistic given by

$$\sum_{(w_r,w_c,v)\in\mathcal{B}} \frac{(o_{w_r,w_c,v} - e_{w_r,w_2,v})^2}{e_{w_r,w_c,v}}. \quad (3.24)$$

Both the Freeman-Tukey and the Chi-square statistics have asymptotic chi-square distributions (corresponding to the total number of minutiae being large) with $|\mathcal{B}| - 1$ degrees of freedom under H_0, where $|\mathcal{B}|$ is the total number of blocks in $\mathcal{B}$. The chi-square distribution can be used to obtain a p-value to either accept or reject H_0. Small (respectively, large) p-values, typically below (respectively, above) 0.05, lead to rejection (respectively, acceptance) of H_0, which in the case of Equation 3.22, leads to a conclusion that a mixture model is inadequate (respectively, adequate) as a model for minutiae.

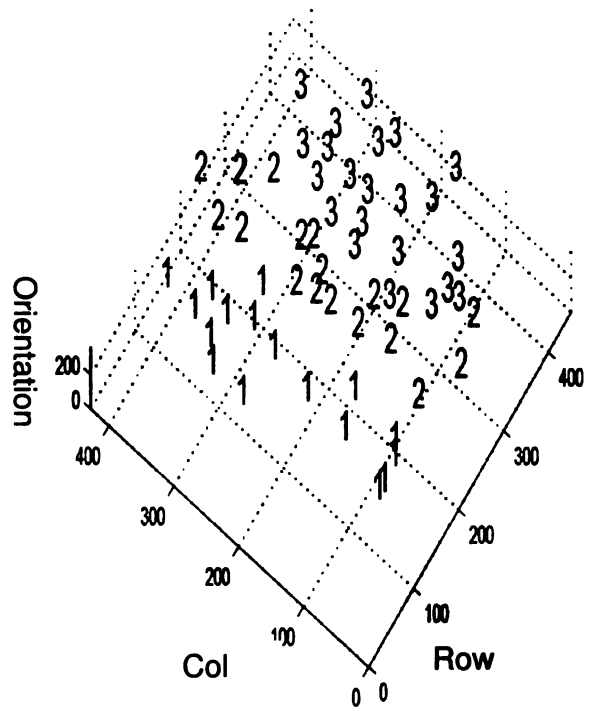
To perform the good-ness-of-fit test, the parameters W and V are taken to be $W = 10$ and $V = 4$, resulting in $W^2V = 400$ blocks. However, since many of these blocks contain less than or equal to 5 minutiae, the merging procedure discussed earlier results in a smaller number of blocks. For example, the fingerprint image in Figure 3.5 (a) gives $|\mathcal{B}| = 8$ blocks, with observed and expected frequencies of $(3, 9, 5, 7, 8, 8, 10, 7)$ and $(5.1, 5.8, 6.3, 6.9, 9.0, 6.5, 10.9, 6.6)$, respectively. The Freeman-Tukey and Chi-square tests give p-values of 0.88 and 0.84, respectively, based on a chi-square distribution with 7 degrees of freedom, resulting in the acceptance of H_0.

In order to compare the adequateness of the mixture and uniform models as candidate models on minutiae, it is necessary to perform the goodness-of-fit test for the uniform model as well. If a larger number of H_0's are rejected for the uniform model compared to the mixture model, it can be concluded that the mixture model is more adequate for representing the distribution of minutiae. To obtain the goodness-of-fit test for the uniform

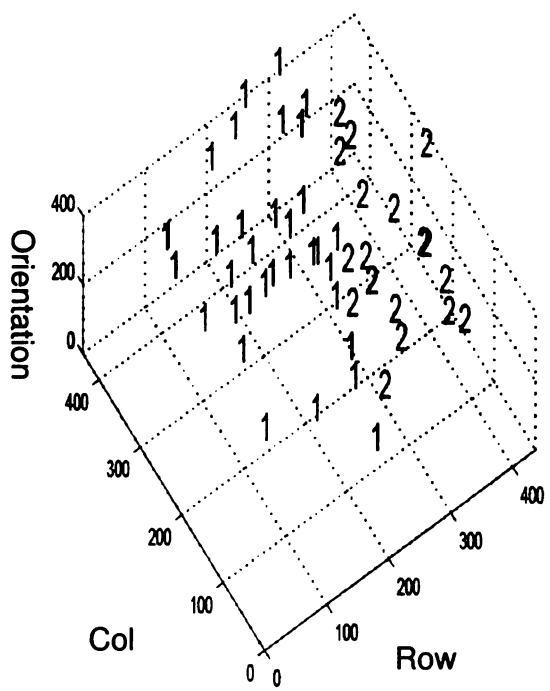
model, we simply substitute $f_\Lambda(s, \theta \mid \Theta_G)$ in Equation 3.22 by the uniform distribution $1/(2\pi A)$, where A is the fingerprint area. The p-values for the Freeman-Tukey and Chi-square tests for the uniform models were calculated in the same way as for the mixture models. The results of p-values can then be used to decide either in favor, or against, H_0. For the fingerprint image in Figure 3.5 (a), the expected frequencies under the uniform model are $(14.6, 5.1, 5.1, 5.1, 5.1, 5.1, 5.1, 12.9)$. The p-values for the Freeman-Tukey and Chi-square tests are 2×10^{-4} and 1.2×10^{-4}, leading to the rejection of the uniform model. Results of the model fit on several fingerprint databases are given in Section 7.2 based on which we can conclude that the mixture model is a far superior model to describe the distribution of minutiae compared to the uniform model.

3.5 Summary

In order to model minutiae distribution, a G-component mixture model was developed in this chapter. The model takes into account clustering tendency of minutiae and dependence between minutiae location and direction. For each of the G components, minutiae location was modeled by a bivariate Gaussian distribution, and minutiae direction was modeled by a mixture of two Von-Mises distributions. To compare the mixture model with the uniform model, goodness-of-fit tests based on Freeman-Tukey test and Chi-square test were performed for both models. The test results showed the superiority of the mixture model over the uniform model.

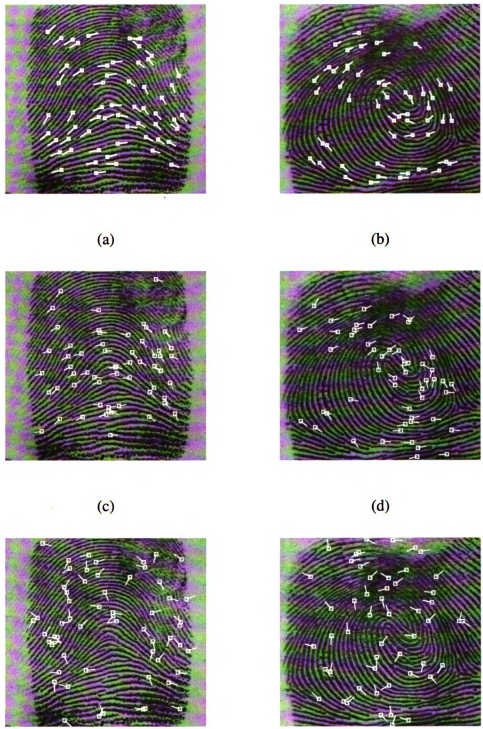


(a)



(b)

Figure 3.4: The clusters in 3-D space for fingerprint images in Figure 3.2 (a-b) are shown in panels (a) and (b) with x, y, z as the row, column, and the direction of the minutiae.



(a) (b)

(c) (d)

(e) (f)

Figure 3.5: Minutiae locations and directions simulated from the proposed model ((c) and (d)), and from the uniform distribution ((e) and (f)) for two different images ((a) and (b)). The true minutiae locations and directions are marked in (a) and (b).

CHAPTER 4

Fingerprint Individuality for a Pair of Fingerprints

4.1 Assumptions for Estimating PRC

To estimate fingerprint individuality, a similarity measure between a pair of fingerprints is required. In this thesis, the similarity measure, $S(Q,T)$, is defined as the number of minutiae matches between a query fingerprint Q and a template fingerprint T. Thus, estimation of PRC is equivalent to finding the probability distribution of the number of matched minutiae pairs for every impostor (Q,T) pair from the target population. This estimation is achieved by a newly developed mathematical model introduced in this chapter.

Suppose the query, Q has n minutiae and template T has m minutiae. Let (S_i^Q, D_i^Q), $i = 1,2,\ldots,n$, and (S_j^T, D_j^T), $j = 1,2,\ldots,m$, be the minutiae in query fingerprint Q and template fingerprint T, respectively. A fingerprint recognition system accepts or rejects the query fingerprint Q based on a threshold w, which is a positive integer. The PRC is the probability of obtaining w or more matched minutiae pairs between Q and T and can be

expressed as

$$PRC(w|m,n) = P(S(Q,T) \geq w \,|\, n, m, I_c \neq I_t)$$
$$= \sum_{w \leq i \leq min\{n,m\}} P(S(Q,T) = i \,|\, n, m, I_c \neq I_t). \quad (4.1)$$

The right-hand side of Equation 4.1, i.e., the probability distribution of $S(Q,T)$, will be carefully investigated in this chapter.

Query and template minutiae are assumed to be independently distributed according to the following mixture densities:

$$f_Q(S^Q, D^Q) = f(S^Q, D^Q \,|\, \Theta_Q) \quad (4.2)$$

and

$$f_T(S^T, D^T) = f(S^T, D^T \,|\, \Theta_T). \quad (4.3)$$

In this chapter, a match is defined in the same way as used by Pankanti et al. [35] (Equation 2.7), and it depends on the two parameters r_0 and d_0. Using the method to estimate r_0 and d_0 introduced in Chapter 2, the values of r_0 and d_0 are found to be 15 and 22.5, based on the ground truth database [38], respectively. These values will be used in the subsequent experiments to estimate the PRC.

As in corrected uniform model by Pankanti et al. [35], minutiae in Q are taken to be at least a distance of $2r_0$ apart from each other. The spatial region within a distance r_0 of S_i^Q is defined by

$$B(S_i^Q; r_0) = \{(x,y) : \sqrt{(x - x_i^Q)^2 + (y - y_i^Q)} \leq r_0\}, \quad (4.4)$$

where $i = 1, 2, \ldots, n$.

It follows from our assumption that the sets $B(s_i^Q; r_0), i = 1, 2, 3, ..., m$ are non-overlapping. A similar condition is imposed on the template minutiae set. Subsequently, the sets $B(S_j^T; r_0)$ for $j = 1, 2, \ldots, m$, are also pairwise non-overlapping. It follows from this assumption that there can be at most one match for each query minutiae point

(S_i^Q, D_i^Q). If (S_i^Q, D_i^Q) matches with (S_j^T, D_j^T) for some j, then (S_j^T, D_j^T) cannot match with the other minutiae points of Q.

4.2 Model for Estimating Fingerprint Individuality

An analytical model for the PRC is obtained in this section. Initially, let the query minutiae set, (S_i^Q, D_i^Q), $i = 1, 2, \ldots, n$, be fixed. Define

$$u_i = P\{|S^T - S_i^Q| \leq r_0 \text{ and } |D^T - D_i^Q| \leq d_0\} \tag{4.5}$$

to be the probability that a random minutiae from T, (S^T, D^T), distributed according to Equation 4.3, is matched with (S_i^Q, D_i^Q). Similarly, let

$$v_j = P\{|S^T - S_j^T| \leq r_0 \text{ and } |D^T - D_j^T| \leq d_0\} \tag{4.6}$$

denote the probability that (S^T, D^T) is matched with (S_j^T, D_j^T). The dependence of u_i and v_j on Q and T via the mixture distribution is implicit and subsequently suppressed.

We first compute the probability that there is exactly one match between Q and T, without loss of generality, between (S_1^Q, D_1^Q) and (S_1^T, D_1^T). This probability is given by

$$u_1 \times \left(\frac{1 - \sum_{i=1}^{n} u_i}{1 - u_1} \right) \times \left(\frac{1 - \sum_{i=1}^{n} u_i - v_2}{1 - u_1 - v_2} \right) \times \left(\frac{1 - \sum_{i=1}^{n} u_i - v_2 - v_3}{1 - u_1 - v_2 - v_3} \right) \times \cdots \times \left(\frac{1 - \sum_{i=1}^{n} u_i - \sum_{j=2}^{m} v_j}{1 - u_1 - \sum_{j=2}^{m} v_j} \right). \tag{4.7}$$

The first term in Equation 4.7, namely u_1, corresponds to the event that there is a match between the minutiae points (S_1^Q, D_1^Q) and (S_1^T, D_1^T). The second term is the probability that (S_2^T, D_2^T) does not match with any of the other minutiae (S_i^Q, D_i^Q), $i \geq 2$, given

that there is already a match for (S_1^Q, D_1^Q). Given (S_1^T, D_1^T) and (S_2^T, D_2^T), the third template minutiae (S_3^T, D_3^T) can be positioned anywhere in the region outside $B(S_1^Q, D_1^Q)$ and $B(S_2^T, D_2^T)$; $B(S_2^T, D_2^T)$ has to be considered as well, due to the imposed condition that the template minutiae should not be close to each other. The requirement that (S_3^T, D_3^T) should not match with any other query minutiae (S_i^Q, D_i^Q), $i \geq 2$, gives

$$\frac{1 - \sum_{i=1}^{n} u_i - v_2}{1 - u_1 - v_2} \tag{4.8}$$

as the required probability, which is the third term in Equation 4.7. Proceeding in this way, the last term in Equation 4.7 is the probability that (S_m^T, D_m^T) does not match any of the previous template minutiae (S_j^T, D_j^T), $j \geq 2$, due to the same condition that the template minutiae should not be close to each other, or any of the query minutiae points.

Carrying this argument for w minutiae matches, the probability of obtaining matches between (S_l^Q, D_l^Q) and (S_l^T, D_l^T) for $l = 1, 2, \ldots, w$, and no matches between all remaining minutiae is shown in Equation 4.9 and is denoted as $g(\{(S_l^Q, D_l^Q) : l = 1, 2, \cdots w\}, \{(S_l^T, D_l^T) : l = 1, 2, \cdots, w\}$.

$$\begin{aligned} & g(\{(S_l^Q, D_l^Q) : l = 1, 2, \cdots w\}, \{(S_l^T, D_l^T) : l = 1, 2, \cdots, w\}) \\ = \; & \underbrace{u_1 \times \left(\frac{u_2}{1 - u_1}\right) \times \left(\frac{u_3}{1 - u_1 - u_2}\right) \times \cdots \times \left(\frac{u_w}{1 - \sum_{l=1}^{w-1} u_l}\right)}_{(1)} \\ & \underbrace{\times \left(\frac{1 - \sum_{i=1}^{n} u_i}{1 - \sum_{i=1}^{w} u_i}\right) \times \left(\frac{1 - \sum_{i=1}^{n} u_i - v_{w+1}}{1 - \sum_{i=1}^{w} u_i - v_{w+1}}\right) \times \cdots \times \left(\frac{1 - \sum_{i=1}^{n} u_i - \sum_{j=w+1}^{m-1} v_j}{1 - \sum_{l=1}^{w} u_l - \sum_{j=w+1}^{m-1} v_j}\right)}_{(2)}, \end{aligned} \tag{4.9}$$

where the term (1) denotes the probability that minutiae $\{(S_l^Q, D_l^Q) : l =$

$3,4,\cdots w\}, \{(S_l^T, D_l^T) : l = 3,4,\cdots,w\}$ are corresponded and term (2) denotes the probability that there are no other matches except the ρ matches between Q and T.

Equation 4.9 is derived assuming that matches occurred between the first w minutiae of the query and template, and no other matches were found for the remaining minutiae. It is the first step towards estimation of probability of exactly w matches between Q and T. In general, the match between Q and T can happen between any w minutiae from Q and T respectively. Let $\{S_{l^Q}^Q, D_{l^Q}^Q\}$ be the w minutiae from Q that are matched to minutiae in T, where

$$l^Q = \{(i_1, i_2, \cdots, i_w) : 1 \leq i_1 < i_2 < \cdots < i_w \leq n\}.$$

Let $\{S_{l^T}^T, D_{l^T}^T\}$ be the w minutiae from T that are matched to minutiae in $\{S_{l^Q}^Q, D_{l^Q}^Q\}$, where

$$l^T = \{(j_1, j_2, \cdots, j_w) : 1 \leq j_1, j_2, \cdots j_w \leq m\}.$$

Denote the minutiae in T that fail to match with any minutiae in Q as $\{S_{l^0}^T, D_{l^0}^T\}$, where

$$l^0 = \{j_{w+1}, j_{w+2}, \cdots, j_m : 1 \leq j_1, j_2, \cdots j_w \leq m\}.$$

The probability that minutiae $\{S_{l^Q}^Q, D_{l^Q}^Q\}$ are matched to minutiae $\{S_{l^T}^T, D_{l^T}^T\}$, and no other minutiae are matched, is

$$\begin{aligned} & f(\{(S_{l^Q}^Q, D_{l^Q}^Q)\}, \{(S_{l^T}^T, D_{l^T}^T)\}) \\ = \; & u_1 \times \left(\frac{u_{i_2}}{1-u_{i_1}}\right) \times \left(\frac{u_{i_3}}{1-u_{i_1}-u_{i_2}}\right) \times \cdots \times \left(\frac{u_{i_w}}{1-\sum_{k=1}^{w-1} u_{i_k}}\right) \end{aligned}$$

$$\times \left(\frac{1 - \sum_{i=1}^{n} u_i}{1 - \sum_{k=1}^{w} u_{i_k}} \right) \times \left(\frac{1 - \sum_{i=1}^{n} u_i - v_{j_{w+1}}}{1 - \sum_{k=1}^{w} u_{i_k} - v_{j_{w+1}}} \right) \times \cdots$$

$$\times \left(\frac{1 - \sum_{i=1}^{n} u_i - \sum_{k=w+1}^{m-1} v_{j_k}}{1 - \sum_{l=1}^{w} u_l - \sum_{k=w+1}^{m-1} v_{j_k}} \right). \tag{4.10}$$

In order to calculate the probability of obtaining exactly w matches, all possible w indices out of the total m from the template have to be considered for matching with the first w minutiae of the query. This can be done in $m(m-1)(m-2)\cdots(m-w+1) = m!/(m-w)!$ ways. Furthermore, w query minutiae can be selected for matching in $\binom{n}{w}$ ways. Taking into account the above facts, the probability of obtaining exactly w matches is given by

$$\sum_{\substack{1 \le i_1 < i_2 < \cdots < i_w \le n, \\ 1 \le j_1, j_2, \cdots, j_w \le m}} f(\{(S_{lQ}^{Q}, D_{lQ}^{Q})\}, \{(S_{lT}^{T}, D_{lT}^{T})\}) \tag{4.11}$$

In summary, the probability of exactly w matches between Q and T can be calculated as follows,

(1) For each sequence of w minutiae in Q and T, S_{lQ}^{Q} and S_{lT}^{T}, calculate the probability that minutiae S_{lQ}^{Q} is matched with minutiae S_{lT}^{T} from Equation 4.10.

(2) Calculate the sum of all the $m!/(m-w)!$ probabilities calculated in step (1) and obtain the PRC by Equation 4.11. Thus the probability of exactly w matches between Q and T is achieved.

4.3 Difficulties in Estimating Fingerprint Individuality

There are several difficulties in calculating the PRCs for a given fingerprint database according to Equation 4.11. The main challenge is that it involves a sum over all possible

subset of w distinct indices from $\{1, 2, \ldots, m\}$. Even for moderate values of w and m (such as the values considered in Section 7.1), the number of terms in the summation, i.e., $\binom{n}{w}$, is very large. For example, when $m = 26$ and $w = 12$, the value of $\binom{n}{w}$ is $9,657,700$. Thus, any method that involves simulation for computing the entire summation (or, an estimate of the summation using, for example, bootstrap samples) becomes infeasible in terms of computational time. Another challenge is that the above summation needs to be computed for every pair of impostor fingerprint images in the given database to estimate fingerprint individuality of a target population. For example, the NIST database used in this thesis have, respectively, $3,998,000$ pairs of impostor fingerprint images, making any simulation-based method both infeasible and impractical. In the next section, we show that Equation 4.11 can be approximated by a Poisson distribution which drastically simplifies estimation of fingerprint individuality. The Poisson model solves the problems mentioned above with regard to the summation over different i and j indices. The Poisson model simplifies the estimation of fingerprint individuality for a pair (Q, T). In chapter 5, we consider a population/database from which the pairs (Q, T) are generated. There are problems with computations in this scenario, too, and for this reason, the hyper-mixture models are developed.

4.4 Poisson Model

In this section, we show that Equation 4.11 can be approximated by a Poisson distribution which drastically simplifies assessment of fingerprint individuality and therefore solves the main challenge discussed above.

Recall that u_1, defined in Equation 4.5, is the probability that (S_1^Q, D_1^Q) matches (S^T, D^T) . Let

$$p(Q,T) = E(u_1) = P(|S^T - S^Q| \leq r_0 \text{ and } |D^T - D^Q| \leq d_0) \tag{4.12}$$

denote the probability of a match when (S^Q, D^Q) and (S^T, D^T) are random pair of minu-

tiae from Equations 4.2 and 4.3. For a set of n minutiae in query Q sampled from Equation 4.2, and a set of m minutiae in query T sampled from Equation 4.3,

$$\lambda(Q,T) \equiv m\,n\,p(Q,T) \tag{4.13}$$

is the expected value of the number of matches between Q and T. Similarly, when both Q and T arise from Equation 4.3, let

$$\nu(Q,T) = m\,n\,E(v_1) \tag{4.14}$$

be the expected value of the number of matches between them, where v_1 is defined in Equation 4.6. The term (2) in Equation 4.9 is approximately $e^{-\lambda}$ for large m and n. This is derived as follows. When m and n are large ($m \times n > 100$), the largest summation involving the v_j's in Equation 4.9 (term (2)) is

$$\sum_{j=w+1}^{m-1} v_j \;=\; m \cdot \left(\frac{1}{m} \sum_{j=w+1}^{m-1} v_j \right) \approx mE(v_1) = m\left(\frac{\nu}{nm}\right) = \frac{\nu}{n}. \tag{4.15}$$

Since $\dfrac{\nu}{n}$ is close to zero for large n, each of the terms $v_{w+1}, v_{w+1} + v_{w+2}, \ldots, \displaystyle\sum_{j=w+1}^{m-1} v_j$ is also close to zero by Equation 4.15. Moreover, note that the difference

$$1 - \sum_{i=1}^{w} u_i - \left(1 - \sum_{i=1}^{n} u_i\right) = \sum_{i=w+1}^{n} u_i \approx \frac{\lambda}{m} \tag{4.16}$$

using an argument similar to that in Equation 4.15. Furthermore, since $\displaystyle\sum_{i=1}^{w} u_i \approx 0$ for large m and n, we get the following two equations:

$$1 - \sum_{i=1}^{n} u_i = 1 - n\left(\frac{1}{n}\sum_{i=1}^{n} u_i\right) \approx 1 - n\left(\frac{\lambda}{mn}\right) = 1 - \frac{\lambda}{m}, \tag{4.17}$$

and

$$1 - \sum_{i=1}^{w} u_i \approx 1. \tag{4.18}$$

In order to calculate (term (2)) in Equation 4.9, first calculate the logarithm of ((2)) as follows:

$$\underbrace{(m-w)\log(1-\sum_{i=1}^{n}u_i)-(m-w)\log(1-\sum_{i=1}^{w}u_i)}_{(I)}$$
$$+\underbrace{\sum_{B=w+1}^{m-1}\log\left(1-\frac{\sum_{j=w+1}^{B}v_j}{1-\sum_{i=1}^{n}u_i}\right)-\sum_{B=w+1}^{m-1}\log\left(1-\frac{\sum_{j=w+1}^{B}v_j}{1-\sum_{i=1}^{w}u_i}\right)}_{(II)}.$$

(II) can be simplified as

$$\begin{aligned}(II) &= \frac{\sum_{B=w+1}^{m-1}\sum_{j=w+1}^{B}v_j}{1-\sum_{i=1}^{n}u_i}-\frac{\sum_{B=w+1}^{m-1}\sum_{j=w+1}^{B}v_j}{1-\sum_{i=1}^{w}u_i}\\ &= \frac{\lambda/m}{(1-\lambda/m)}\sum_{B=w+1}^{m-1}\sum_{j=w+1}^{B}v_j\\ &\leq C\frac{\lambda}{m}\left(m\frac{\nu}{n}\right) \qquad (4.19)\end{aligned}$$

for some constant C for large m and n using equations 4.15, 4.16 and 4.17. From the last line in Equation 4.19, it follows that (II) goes to zero as m and n go to to infinity. Thus,

$$\begin{aligned}\log(B) &= (m-w)\log(1-\sum_{i=1}^{n}u_i)-(m-w)\log(1-\sum_{i=1}^{w}u_i)\\ &= -(m-w)\sum_{i=w+1}^{n}u_i\\ &= -(m-w)\frac{\lambda}{m}\rightarrow-\lambda\end{aligned}$$

as m and n go to to infinity. Therefore (term 2) in (4.9) can be approximated by

$$\exp\{-\lambda(Q,T)\} \qquad (4.20)$$

when both m and n are large, and when the number of matches w is moderate (i.e., w/m and w/n are not too close to either 0 or 1). In applications, where $m \times n > 100$ and $m \times n \times p(Q,T) < 10$, m and n are considered sufficiently large and the number of matches are considered as moderate [39].

Each term in the denominator of (term (1)) in equation 4.9 is close to 1 since $u_i \approx 0$ for each $i = 1, 2, \ldots, w-1$. Thus, equation 4.9 can be written approximately as

$$\prod_{i=1}^{w} u_i \exp\{-\lambda(Q,T)\}. \tag{4.21}$$

Similar to Equation 4.21, starting with Equation 4.10 (instead of Equation 4.9), the following holds:

$$f(\{(S_{lQ}^{Q}, D_{lQ}^{Q})\}, \{(S_{lT}^{T}, D_{lT}^{T})\}) = \prod_{k=1}^{w} u_{i_k} \exp\{-\lambda(Q,T)\}. \tag{4.22}$$

Applying Equation 4.22 to Equation 4.11, the probability of obtaining exactly w matches is given by

$$p(w\,;\,Q,T) = \frac{m!}{(m-w)!}\left(\sum_{i_1<i_2<\cdots<i_w}\prod_{k=1}^{w} u_{i_k}\right)\cdot \exp\{-\lambda(Q,T)\}, \tag{4.23}$$

where the summation in (4.23) is over all w distinct indices $i_1 < i_2 < \ldots < i_w$ from $\{1, 2, \ldots, n\}$.

Equation 4.23 can be simplified when n is large. Note that

$$\frac{1}{\binom{n}{w}}\sum_{i_1<i_2<\cdots<i_w}\prod_{k=1}^{w} u_{i_k} \approx (E(u_1))^w, \tag{4.24}$$

where $E(u_1)$ is defined as in Equation 4.12, because u_i's are independent and identically distributed. Substituting Equation 4.24 in Equation 4.23 with $E(u_1) = \lambda(Q,T)/(mn)$ from Equation 4.13, the probability of exactly w matches is

$$\begin{aligned} p(w\,;\,Q,T) &\approx \frac{m!}{(m-w)!}\binom{n}{w}\left(\frac{\lambda(Q,T)}{mn}\right)^w \exp\{-\lambda(Q,T)\} \\ &\approx \frac{e^{-\lambda(Q,T)}\,\lambda(Q,T)^w}{w!} \end{aligned} \tag{4.25}$$

for large m and n and moderate w, which corresponds to a Poisson distribution with mean $\lambda(Q, T)$.

4.5 Justification of the Poisson Model

Equation 4.29 corresponds to a Poisson probability mass function with mean λ (as defined in Equation 4.13), where λ is the expected number of matches from the total number of mn possible pairings between n minutiae in Q and m minutiae in T, with the probability of each match being $p(Q,T)$. Using a Poisson distribution to approximate the number of matched minutiae pairs, which follows a binomial distribution, is valid because of the following three properties. (i) In fingerprint matching, a "success" is defined as a minutia match and the number of trials, mn, is large ($\geq$ 100) [39]. This can be confirmed by Figures 7.2 and 7.3 in Chapter 7, which show that m and n are much greater than 10 for most fingerprints. (ii) The probability of a success, $p(Q,T)$, is small. (iii) The number of impostor matches between Q and T is moderate. The properties (ii) and (iii) can be confirmed by Table 7.7 where the right-most column shows that, for all three databases, the empirical value of the number of matched minutiae pairs between two fingerprints is always less than 10. In summary, it is appropriate to use Poisson model in Equation 4.29.

4.6 Overlapping Area Model: Comparison of Mixture Model with Corrected Uniform Model

During fingerprint matching, an overlapping area is formed after alignment between minutiae regions of Q and T (see Figure 4.1). Assume that n_0 out of the n minutiae in Q, and m_0 out of the m minutiae in T are within the overlapping area. The PRC proposed by Pankanti et al. [35] was based on the parameters within the overlapping area, namely n_0 and m_0, whereas the PRC in this thesis is based on n and m of the entire fingerprint.

Because of this difference, the PRCs from the corrected uniform model and those from the mixture model on the entire fingerprint area cannot be compared. For comparing with the corrected uniform model, an overlapping area is developed, adopting the parameter n_0 and m_0 as used by Pankanti et al. [35], but still estimating the PRC based on the minutiae density from the mixture models.

4.6.1 Determination of n_0, m_0 and the Overlapping Area

The first step of overlapping area model is to find n_0, m_0 and the overlapping area between Q and T. To find the overlapping area, the model has to first determine the entire minutiae region for both Q and T, which is defined as the minimal ellipse that encompasses all the minutiae in the fingerprint, as follows. A convex hull encompassing all minutiae locations is first determined (see Figure 4.2, where minutiae locations are labeled as squares). Then, an ellipse (denoted by a dashed line in Figure 4.2) is obtained by the direct least square fitting method [17]. Since some of the minutiae fall outside the dashed ellipse, the size of the ellipse is increased until it encloses all the minutiae. The resulting ellipse (i.e., the minimal ellipse) is denoted by a solid line in Figure 4.2. After determining the minutiae regions, Q and T are aligned with a Procrustes transformation [29]. This completes the first step of the model construction, namely, finding n_0, m_0 and the overlapping area.

4.6.2 Adaptation of Mixture Model to Overlapping Area Model

The second step of the overlapping area model is to estimate the PRC adopting n_0 and m_0 as used by Pankanti et al. [35], and employing the minutiae density of the mixture model truncated to the overlapping area. Assume that the minutiae densities of fingerprints Q and T are $f_Q(S, D)$ and $f_T(S, D)$ as in Equations 4.2 and 4.3, respectively. The alignment of minutiae region Q with respect to minutiae region T, which is obtained from the Procrustes transformation, is denoted as $TR(Q, T)$. Further assume that after alignment, the overlapping area is $A(Q, T)$. It is obvious the minutiae density from fingerprint Q in the

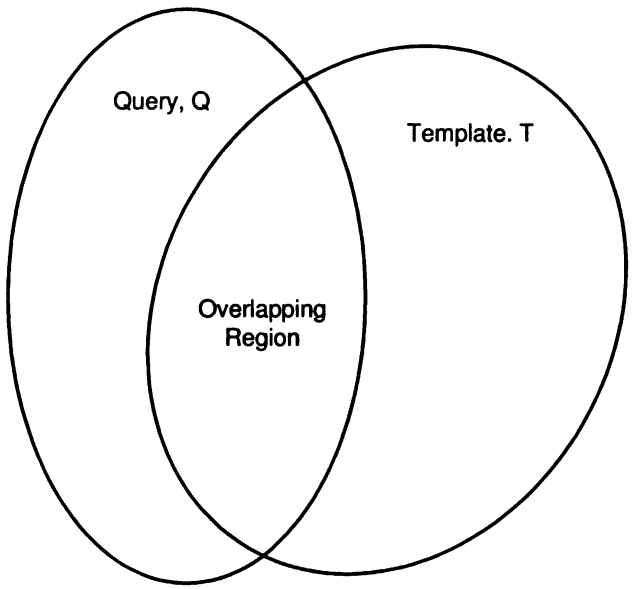


Figure 4.1: Overlapping area of two fingerprints during matching

overlapping area is

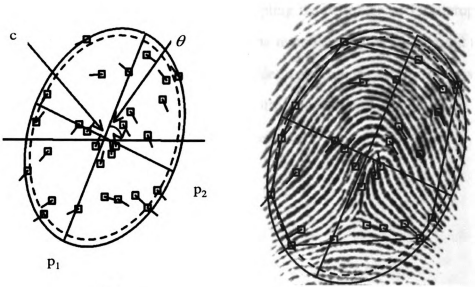
$$\frac{f_Q(TR^{-1}(Q,T)(S,D)) \times |J|}{\int_{A(Q,T)} f_Q(TR^{-1}(Q,T)(S,D)) \times |J| dSdD}. \tag{4.26}$$

where $TR^{-1}(Q,T)(S,D)$ is the transformation to obtain the original placement of query Q without its alignment with T at which placement the minutiae density of Q is estimated and $|J|$ is absolute value of the Jacobian determinate. The minutiae density from T in the overlapping area is

$$\frac{f_T(S,D)}{\int_{A(Q,T)} f_T(S,D) dSdD}. \tag{4.27}$$

Let (S^Q, D^Q) be a random selected minutia from Q in the overlapping area, and let (S^T, D^T) be a random selected minutiae from T in the overlapping area. Furthermore, let the probability of a random match in the overlapping area be

$$p_0(Q,T) = P(|S^Q - S^T| \leq r_0 \text{ and } |D^Q - D^T| \leq d_0 | n_0, m_0). \tag{4.28}$$



(a) (b)

Figure 4.2: Convex hull of minutiae and best fitting ellipses. (a) The minutiae from image (b) and the best fitting ellipse to the minutiae set. p_1, p_2, c, θ are the major axis, minor axis, center, and orientation of the ellipse. (b) Fingerprint image with minutiae and the best fitting ellipse

Applying the Poisson model, the probability of obtaining exactly w minutiae matches is

$$p_0(w\,;\,Q,T) \approx \frac{e^{-\lambda_0(Q,T)}\,\lambda_0(Q,T)^w}{w!}\,, \tag{4.29}$$

where $\lambda_0(Q,T) = n_0 \times m_0 \times p_0(Q,T)$.

4.7 Summary

A mathematical model was developed to estimate PRCs, given a pair of fingerprints Q and T with n and m minutiae, respectively. The minutiae density was estimated from the mixture models developed in Chapter 3. Although the model calculates PRCs in a closed form, its estimation is time-consuming for practical applications. Hence a Poisson approximation was derived to improve the computational efficiency. The Poisson model simplifies the estimation of fingerprint individuality for a pair (Q,T).

To compare PRCs from the mixture model with those from the corrected uniform model,

an overlapping area model was developed, adopting the definition of PRC proposed by Pankanti et al. [35], while still assuming that the minutiae density is estimated from the proposed mixture model. In chapter 5, we consider a population/database from which the pairs (Q,T) are generated. There are problems with computations in this scenario, too, and for this reason, the hypermixture models are developed.

CHAPTER 5

Assessment of Fingerprint Individuality: Target Population

There are two challenges involved in obtaining an assessment of fingerprint individuality. The first challenge was resolved after Poisson model was developed in the last chapter. The second challenge for the assessment of fingerprint individuality is that computation for every pair of impostor fingerprint images in the given database is required to estimate fingerprint individuality of a target population. For example, the NIST 2000 SD4 database used in this thesis have, respectively, $3,998,000$ pairs of impostor fingerprint images, making any simulation-based method both infeasible and impractical. The hyper-mixture density model developed in this chapter, on the other hand, is meant to solve the second challenge.

5.1 The Hyper-mixture Density Model

Assume in a target population there are N^* unknown different minutiae distribution groups with class densities $H_1, H_2, ..., H_{N^*}$ and corresponding proportions $\pi_1, \pi_2, ..., \pi_{N^*}$ (where $\pi_i \geq 0$ for $i = 1, 2, ..., N^*$, and $\sum_{i=1}^{N^*} \pi_i = 1$). Thus, fingerprints in different groups have different distributions (H_i's), whereas those within the same group have similar minutiae

distributions. Therefore, this hyper-mixture model is capable of capturing different minutiae distributions in different fingers in the population. Note that this assumption is needed since it is well-known that fingerprints belong to five different classes (i.e., right-loop, left-loop, whorl, arch and tented arch), and therefore are likely to have different class-specific minutiae distributions. Thus, using only one common minutiae distribution may smooth out different distributions in the fingerprint classes. Moreover, PRCs depend heavily on the composition of each target population and different target populations may have different composition of the fingerprint classes [18]. For example, the proportion of occurrence of the right-loop, left-loop, whorl, arch and tented arch classes of fingerprints is estimated to be 31.7%, 33.8%, 27.9%, 3.7% and 2.9%, respectively, in the general population [27]. Thus, PRCs computed for fingerprints from the general population will be influenced more by the mixture models fitted to the right-loop, left-loop and whorl classes, than to the arch and tented arch classes. In effect, the composition of target population needs to be studied, which is the goal of the hyper-mixture model. Besides the possible uneven proportions, more important is the fact that the class proportions might change across different target populations (for example, if the target population has an equal number of fingerprints in each class, or with class proportions different from the ones given above), which will lead to change of the PRCs. With a hyper-mixture model comprising of N^* clusters of minutiae distributions, the methodology of obtaining PRCs for a pair of fingerprints can be extended to any target population.

To formally obtain the composition of a target population, an agglomerative hierarchical clustering procedure [21] was adopted on the space of all fitted mixture models. The dissimilarity measure between the estimated mixture densities f and g is taken to be the Hellinger distance [25]

$$H(f, g) = \int_{s \in S} \int_{\theta \in [0, 2\pi)} (\sqrt{f(s, \theta)} - \sqrt{g(s, \theta)})^2 \, dx \, d\theta. \tag{5.1}$$

The reasons for using Hellinger distance, $H(f, g)$, is that it is a number bounded between 0 and 2, with $H(f, g) = 0$ (respectively, $H(f, g) = 2$) if and only if $f = g$ (respectively,

f and g have disjoint support). Thus, we avoid distance measures that are arbitrarily large and therefore can focus on thresholds of clustering in $[0, 2]$ only.

For a database of F fingers, a total of $F(F-1)/2$ Hellinger distances were obtained corresponding to the $F(F-1)/2$ mixture pairs. The agglomerative hierarchical clustering methodology with Ward's method [22] gives a dendrogram that can be cut at an appropriate level to form N clusters of mixture densities, $C_1, C_2, \ldots, C_N$. Note that $N = 1$ when $\lambda = 2$, and N increases to $F(F-1)/2$ as λ decreases to 0. When the number of clusters is N, the within cluster dissimilarity is defined as

$$W_N = \sum_{i=1}^{N} \frac{1}{2|C_i|} \mathcal{D}(C_i), \tag{5.2}$$

where

$$\mathcal{D}(C_i) = \sum_{f,g \in C_i} H(f, g) \tag{5.3}$$

is the sum of all distances $H(f, g)$ for f and g in C_i, and $|C_i|$ is the number of mixture densities in C_i. Note that as N increases to F, W_N decreases to 0. To choose the optimal number of clusters, the "Gap Statistic" [48] is applied as follows: Let $G_N = |W_N - W_{N-1}|$ denote the absolute difference between the within cluster dissimilarities W_{N-1} and W_N. N^* is selected as the number of clusters if the values of G_N for $N > N^*$ are insignificant (close to 0) compared to the value of G_{N^*}. Figure 5.1 shows the plot of G_N against N for NIST 2000 SD 4. G_N doesn't change significantly when the number of clusters is more than 33. Hence N^* is chosen as 33. For now, N^* is chosen by visual inspection of Figure 5.1; we tend to prefer larger N^* values so as not to under-represent the interclass variability of the population.

Once the number of clusters N^* has been determined, the mean mixture density for each cluster C_i is determined as

$$\bar{f}(s, \theta) = \frac{1}{|C_i|} \sum_{f \in C_i} f(s, \theta), \tag{5.4}$$

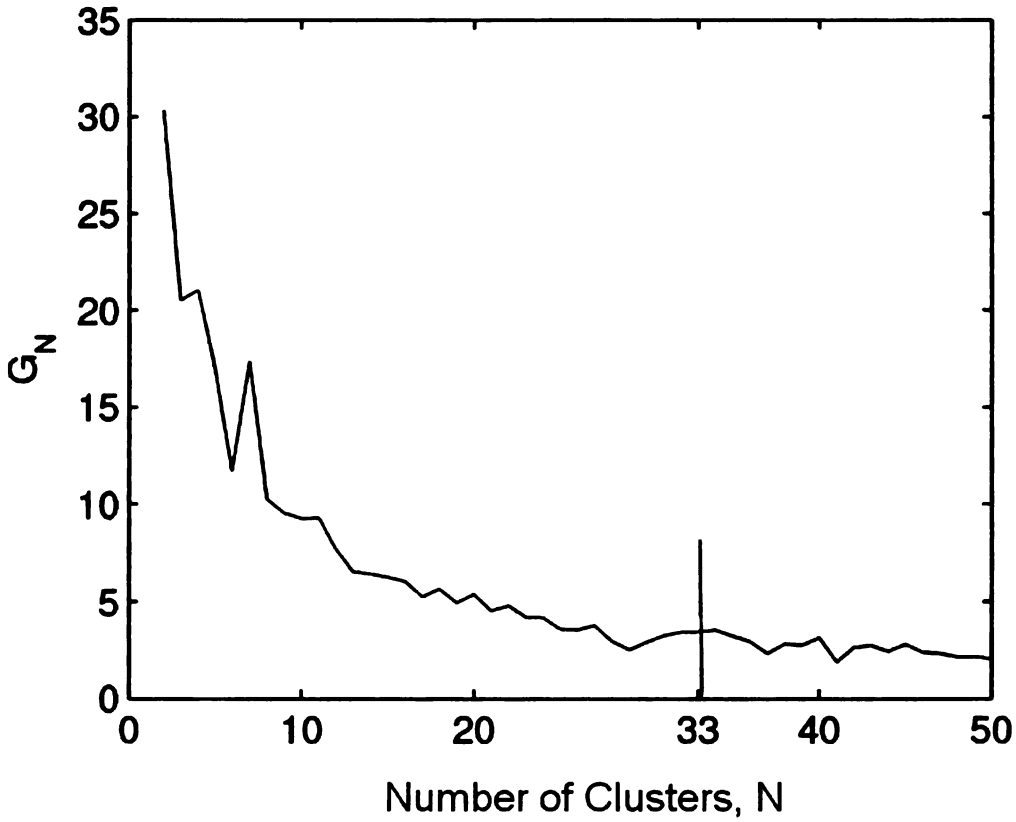


Figure 5.1: Determination of the number of clusters for NIST 2000 SD 4. The number of clusters estimated is 33.

where $f(s, \theta)$ is the mixture distribution from Equation 3.5. The weight π_i of cluster C_i is

$$\pi_i = \frac{|C_i|}{\sum_{i=1}^{N^*} |C_i|}. \tag{5.5}$$

At this point, all the parameters in the model have been estimated. Since the minutiae distribution of each cluster is a mixture (average) of the densities of the mixture models that are fitted to each individual finger, the model is appropriately called a "hyper-mixture model".

5.2 Relationship between Clusters from Hyper-mixture Model and Fingerprint Classes

Fingerprint clusters in the hyper-mixture model are determined by minutiae distribution, whereas the definition of the fingerprint classes is based on global ridge pattern. Thus the relationship between clusters and fingerprint classes is worthy of investigation. Fingerprints in a cluster can belong to different classes. If the clusters and classes are totally uncorrelated, every one of the three classes should be evenly distributed among all the 33 clusters. On the other hand, if a certain fingerprint class tends to concentrate into some clusters, it can be concluded that there is a strong correlation between clusters and classes. Figure 5.2 shows the composition of the three major fingerprint classes (i.e., the number of fingerprints for each class, loop, whorl, arch, where loop includes left loop and right loop, arch includes arch and tented arch) for each of the 33 clusters in the NIST 2000 SD 4. The wide spread on the vertical axis indicates that the classes are not evenly distributed among the 33 clusters (otherwise the plot should be approximately a flat line). More quantitatively, for each fingerprint class, after ranking the clusters according to the number of fingerprints in the class, the total number of fingerprints from the top 16 out of the 33 clusters were counted. If the clusters are evenly distributed, the total number should be close to

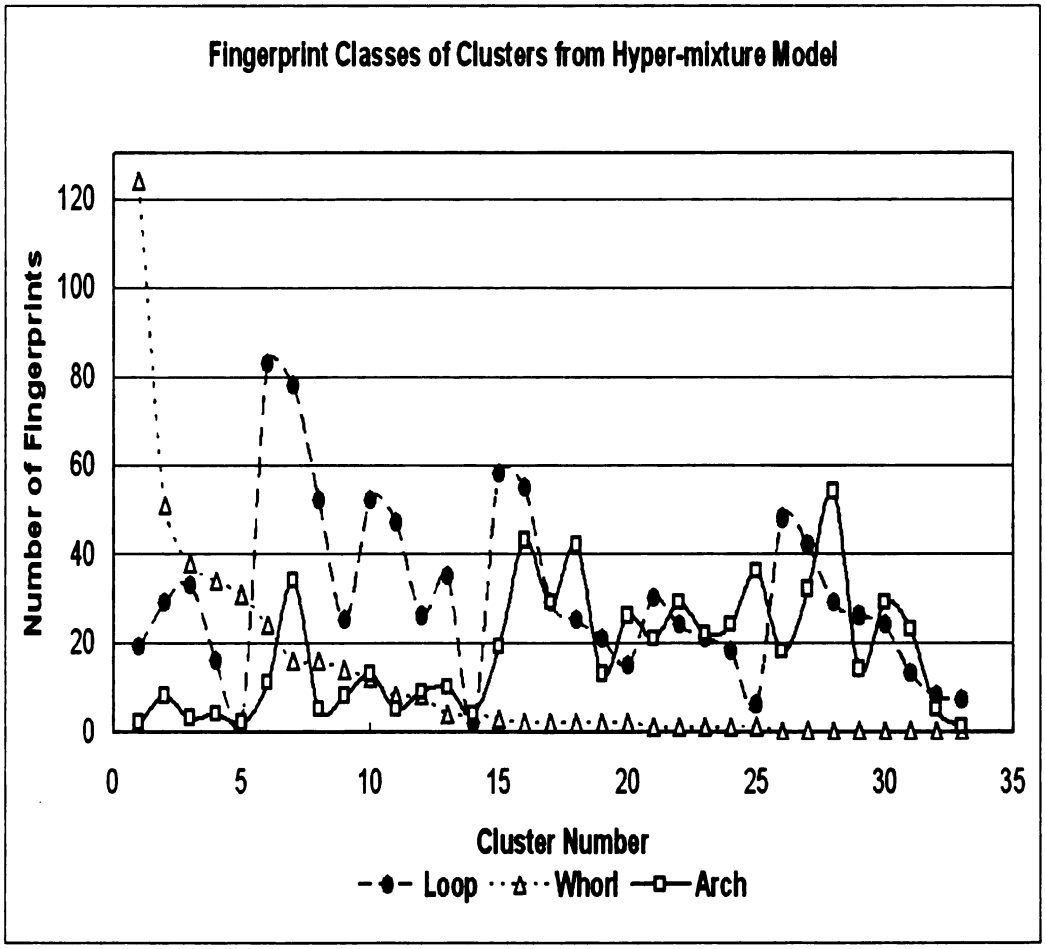


Figure 5.2: The number of fingerprints in the three main classes for the 33 clusters from the hyper-mixture models on NIST 2000 SD 4. For each cluster label i as shown in the horizontal axis, the vertical coordinate of each point shows the number of fingers in loop (labeled with dots), whorl (labeled with triangles) and arch (labeled with squares).

$16/33 \approx 50\%$. However, the top 16 whorl clusters consist of 97% of all whorl fingerprints; for the top 16 loop clusters and the top 16 arch clusters, the percentages are 73% and 80%, respectively. Furthermore, the whorl clusters are so clustered that 69% of the whorl fingerprints belong to the top 5 whorl clusters, and those 5 clusters have very few fingerprints from the other two classes, namely loop and arch. We conclude there is a strong correlation between clusters of hyper-mixture models and the three main fingerprint classes.

5.3 Assessment of Fingerprint Individuality for a Target Population

Given the N^* cluster densities $(H_1, H_2, ..., H_{N^*})$ and cluster weights $\{\pi_1, \pi_2, ...\pi_{N^*}\}$, fingerprint individuality of the target population can be calculated as follows.

The mean parameter $\lambda(Q,T)$ in Equation 4.13 depends on Q and T via the mean mixture densities of the clusters from which Q and T are taken. If Q and T belong to clusters C_i and C_j, respectively, then the mean mixture densities of C_i and C_j can be used in place of the original mixture densities in Equation 4.12, i.e., $\lambda(Q,T) \equiv \lambda(C_i, C_j)$. Let $p^*(w\,;\, C_i, C_j)$ denote the Poisson probability

$$p^*(w\,;\, C_i, C_j) = e^{-\lambda(C_i,C_j)} \frac{\lambda(C_i, C_j)^w}{w!}. \tag{5.6}$$

For a fingerprint database consisting of N^* different clusters of distributions, the most representative value for the probability of a random correspondence is reported as the estimate of fingerprint individuality for this database. There are a total of $N^*(N^* - 1)$ possible impostor pairs of fingerprint images (Q,T), where Q and T come from different clusters. Let $T_0 = \{(i,j) : 1 \leq i \leq N^* \, and \, 1 \leq j \leq N^*, i \neq j\}$. The average PRC corresponding to w minutiae matches is given by

$$\overline{\text{PRC}} = \frac{\sum_{(i,j)\in T_0} \pi_i \, \pi_j \, p^*(w\,;\, C_i, C_j)}{\sum\limits_{(i,j)\in T_0} \pi_i \, \pi_j}, \tag{5.7}$$

where $p^*(w\,;\, Q,T)$ is defined in Equation 4.29. Note that $p^*(w\,;\, Q,T)$ is symmetric in Q and T, and thus it is sufficient to consider only the $N^*(N^* - 1)/2$ distinct impostor pairs instead of the total $N^*(N^* - 1)$. Each of the probabilities, $p^*(w\,;\, Q,T)$, is very small, e.g., 10^{-6} or 10^{-7}. Thus, the average PRC in Equation 5.7 is highly affected by the largest of these probabilities, and is, therefore, not reliable as an estimate of typical PRCs arising from the impostor pairs. A better measure would be to consider an average of the trimmed probabilities. Let α denote the percentage of $p^*(w\,;\, Q,T)$ to be trimmed,

and let $p^*(w;\alpha/2)$ and $p^*(w;1-\alpha/2)$, respectively, denote the lower and upper $100\alpha/2$-th percentiles of these probabilities. Define the set of all trimmed $p^*(w\,;\,C_i,C_j)$ probabilities as $\mathcal{T} \equiv \{\,(i,j)\,:\,p^*(w;\alpha/2) \le p^*(w\,;\,C_i,C_j) \le p^*(w;1-\alpha/2)\}$. Then, the α-trimmed mean PRC is

$$\overline{\text{PRC}}_\alpha = \frac{\sum_{(i,j)\in\mathcal{T}} \pi_i\,\pi_j\,p^*(w\,;\,C_i,C_j)}{\sum\limits_{(i,j)\in\mathcal{T}} \pi_i\,\pi_j}. \tag{5.8}$$

The above discussion is general and holds true for any distribution of the query and template minutiae. In particular, when the distribution on the minutiae (both location and direction) are chosen to be uniform as in the model by Pankanti et al. [35], the following expression for $\lambda(Q,T)$ is obtained:

$$\lambda_U(Q,T) = m\,n\,p_L\,p_D, \tag{5.9}$$

where p_L (respectively, p_D) is the probability that S^Q and S^T (respectively, D^Q and D^T) will match. The probability of a location-and-direction match appears as the product p_L and p_D since the minutiae location and direction are distributed independently of each other.

5.4 Summary

A hyper-mixture model was proposed to cluster the mixture models of all the fingers in the sample database into clusters so that fingerprints in the same cluster have similar distributions. The PRCs for the target population can be calculated by the weighted average of PRCs for the clusters from the hyper-mixture models. Study on the clusters in the hyper-mixture model showed that fingerprints of the same class are not uniformly distributed in the clusters of hyper-mixture models and fingerprints of the same class tend to belong to the same cluster.

CHAPTER 6

Assessing Fingerprint Individuality: Compound Stochastic Models

6.1 Motivation

There are two sources of fingerprint variability in matching, namely interclass variability and intraclass variability. Chapter 5 addressed interclass variability, i.e., minutiae variability in different fingers in target population. While still taking minutiae locations and directions as the salient features, this chapter focuses on modeling both intraclass variability and interclass variability. The previous studies discussed in Chapter 2 ([3], [11], [16], [18], [23], [41], [47], [50]) did not model intraclass variability. Though the corrected uniform model by Pankanti et al. [35] estimated parameter l from empirical genuine matching, their model didn't study the intraclass variability intensively. In this chapter, compound stochastic models are developed to account for three sources of minutiae variability, namely, (i) variability in minutiae distributions in different fingers, (ii) variability due to local perturbations arising from non-linear distortion effects in multiple impressions of the same finger, and (iii) variability due to partial prints in multiple acquisitions of the same finger. The three sources of variability mentioned here account for most of the variability in minutiae

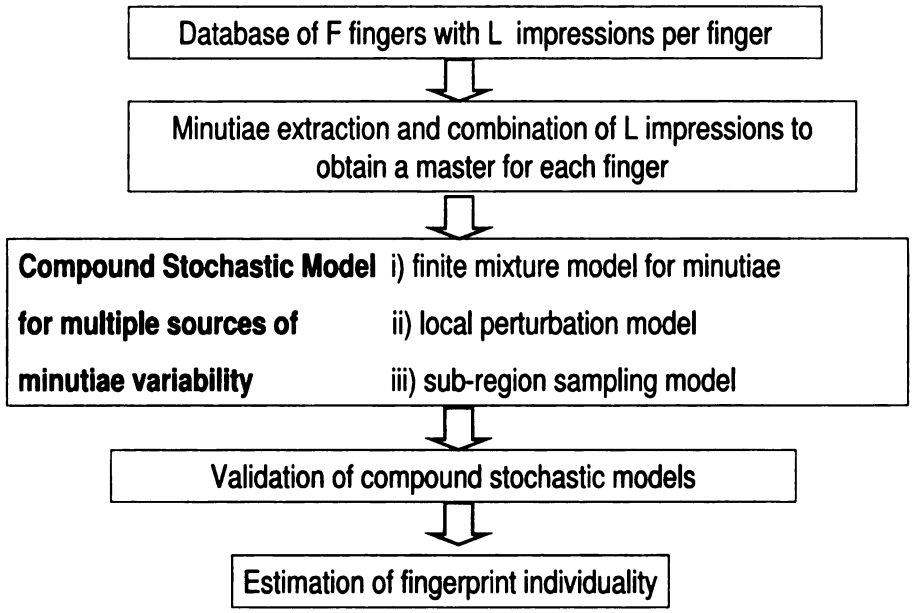


Figure 6.1: Flow chart for constructing compound stochastic model and assessing fingerprint individuality

distributions. Since a compound stochastic model involves both interclass and intraclass variability, it is a more realistic model compared to the hyper-mixture model introduced in Chapter 5 which only addressed the first variability mentioned above. The flow chart in Figure 6.1 gives the steps involved in constructing the compound stochastic model.

6.2 Compound Stochastic Model

6.2.1 Construction of Master Minutiae Set

Suppose a fingerprint database consists of prints of F different fingers with L impressions per finger. Let $\mathcal{F}(f, l)$ denote the l-th impression of the f-th finger. As a first step towards constructing the compound stochastic model for finger f, the minutiae in all the L impressions of a finger, $\mathcal{F}(f, l), l = 1, 2, \ldots, L$, are combined to obtain a "master" set. A

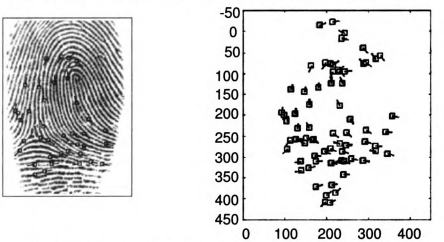
reference impression for each finger f, without loss of generality, $\mathcal{F}(f,1)$, say, the first impression, is chosen as follows: Since the quality and the sensed area corresponding to different impressions of the same finger are typically different, the reference impression is chosen as the one that has relatively good quality computed according to [8] and maximizes, on the average, the number of minutiae matches with all other impressions of that finger.

Once the reference impression is determined, all other impressions are aligned to it via a Procrustes transformation [29] based on the correspondences from the matcher reported in Section 6.4.1. Thus, for each $l = 2,3,\ldots,L$, $\mathcal{F}(f,l)$ is aligned to $\mathcal{F}(f,1)$, using the rigid transformation $T(f,l)$, and correspondences between minutiae in $\mathcal{F}(f,l)$ and $\mathcal{F}(f,1)$ are found. The correspondence between the minutiae sets was achieved by both automatic fingerprint matching and manual verification as described in section 6.4.1. When a minutia in $\mathcal{F}(f,l)$ does not have any corresponding minutiae in $\mathcal{F}(f,1)$, that minutiae is appended to the list of minutiae in $\mathcal{F}(f,1)$. The consolidation of minutiae into the master set in this way eventually results in a total of n consolidated minutiae in $\mathcal{F}(f,1)$ with correspondence sets M_k, $k = 1,2,\ldots,n$. The elements in each M_k are denoted by $\{(S_{kj}, D_{kj}),\ j = 1,2,\ldots,m_k\}$, where S_{kj} and D_{kj} are, respectively, the j-th location and direction of minutiae k. For each set of correspondences, define the mean, or the center, of S_{kj} as $\bar{S}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} S_{kj}$. The mean of D_{kj}, $\bar{D}_k$, is taken to be the phase angle of the complex number $\sum_{j=1}^{m_k} \cos(D_{kj}) + i\sin(D_{kj})$ (see also [28]). The deviations of locations and directions from their respective centers for the k-th minutiae are given by

$$\{(S_{kj} - \bar{S}_k, D_{kj} - \bar{D}_k),\ j = 1,2,\ldots,m_k\}. \tag{6.1}$$

An illustration of the construction of a master minutiae set is presented in Figure 6.2, in which multiple impressions of the same finger (top panel) are aligned to the reference image (bottom left panel) to obtain the master minutiae set with minutiae centers shown (bottom right panel).

(a) Multiple Impressions of the same finger



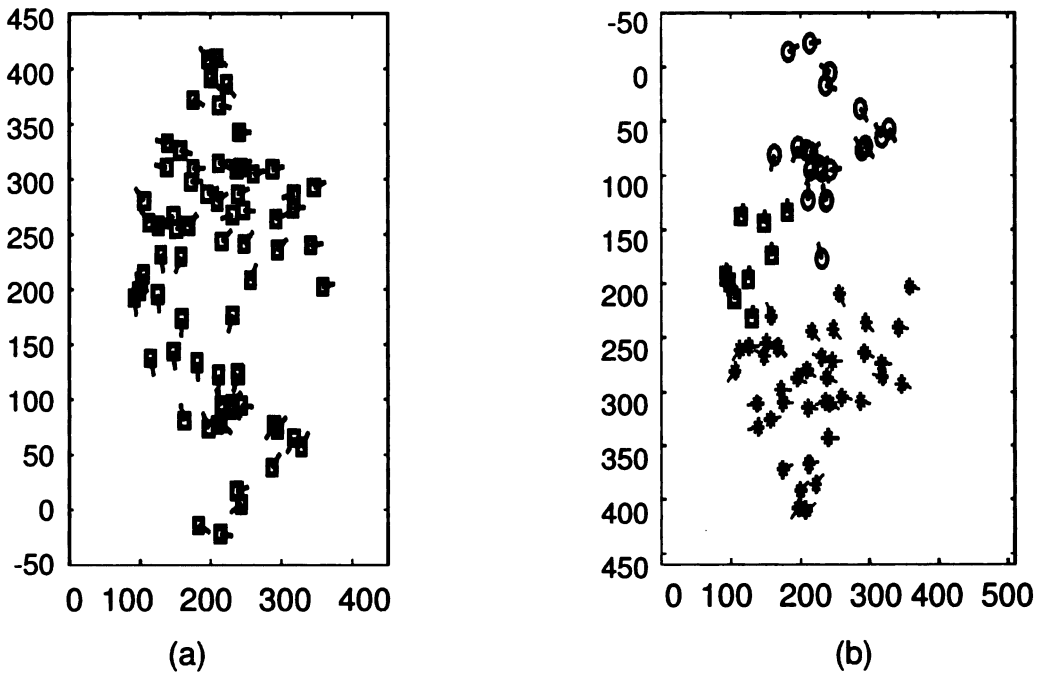
(b) Reference Impression (c) Master

Figure 6.2: Master minutiae set construction. Eight impressions are shown which include the reference impression (b) and the other seven impressions (a). The number of minutiae in each impression in the first row is 29, 30, 27, 32, 32, 38, 24. The number of minutiae in the reference impression in the second row is 38. There are 70 minutiae centers kept in the master minutiae set .

6.2.2 Mixture Model on the Centers: Adaptation of Mixture Models to Compound Stochastic Models

The first two stages of the compound stochastic models consist of developing statistical models on (i) the centers, and (ii) the deviations of the observed minutiae from their respective centers.

Each minutia center in the master, $(\bar{S}_k, \bar{D}_k)$, $k = 1, 2, \ldots, n$, is assumed to be independently distributed according to the mixture density defined in Equation 3.5. The parameters of the mixture model are estimated using the method introduced in Chapter 3. Figure 6.3



(a) Master (b) Components of mixture models

Figure 6.3: The ability of the mixture model to capture clustering characteristics of the master in (a). The eight impressions are shown in Figure 6.2. Three cluster components labeled by circles, squares and asterisk in the mixture model fitted to the minutiae in (a).

shows an example of fitted mixture model to a master minutiae set. Figure 6.3 (a) is the master minutiae set obtained from eight different fingerprint impressions of a finger. Figure 6.3 (b) shows three clusters, labeled with circles, squares and asterisks, obtained by the mixture model fitting.

6.2.3 Local Perturbation Model

The local perturbation model consists of a probability model to capture fingerprint distortions in different finger regions. For the local perturbation model, the domain of the master is first divided into a lattice of b_0 non-overlapping blocks, $\mathcal{B} = \{ B_b, b = 1, 2, \ldots, b_0 \}$. If the mean of the k-th minutiae $\bar{S}_k$ belongs to B_b, then the k-th minutiae is assigned to block

B_b.

In block B_b, the location deviations of all the minutiae that fall in B_b, i.e.,

$$\{S_{kj} - \bar{S}_k : \bar{S}_k \in B_b\},$$

are modeled as a bivariate normal distribution with mean zero and covariance matrix COV_{B_b}. The covariance matrix COV_{B_b} allows for flexible modeling of the dominant directions of distortions via the eigenvalues and eigenvectors of COV_{B_b}. In this chapter, the covariance matrix COV_{B_b} is estimated by

$$\widehat{COV}_{B_b} = \frac{1}{N} \sum_{k:\bar{S}_k \in B_b} \sum_{j=1}^{m_k} (S_{kj} - \bar{S}_k) \cdot (S_{kj} - \bar{S}_k)^T,$$

where $N = \sum_{k:\bar{S}_k \in B_b} m_k$.

The minutiae direction deviations, on the other hand, $\{D_{kj} - \bar{D}_k : \bar{S}_k \in B_b\}$, are modeled as a Von-Mises distribution with mean zero and precision κ_{B_b}. The unknown parameter κ_{B_b} is estimated from the observed deviations in each block B_b (based on the estimation procedure given in [28]).

The local perturbation model assumes that the non-linear distortions of different minutiae within the same block are independent and identically distributed, whereas the distortions in different blocks can be different. Figure 6.4 (a) shows two minutiae, labeled as 1 and 2, in the reference image. The locations and directions of minutiae 1 and 2 in seven other aligned impressions are shown in Figures 6.4 (b) and (c). Note that there are multiple location and direction values for the same minutiae in different impressions of the same finger. The changes in the location and direction values are due to nonlinear distortion introduced during sensing as the three-dimensional finger surface is projected onto a two-dimensional plane. The amount of distortion is different in different regions of the finger. The distortion is usually less in regions closer to the center of a finger, compared to peripheral regions due to nonuniform pressure of the finger against the sensor [7]. In the area closer to the center of the fingerprint image, the pressure is high and the slippage is little and therefore

(a) (b) minutiae 1 (c) minutiae 2

Figure 6.4: Consolidating minutiae: (a) a partial fingerprint image. (b) and (c) show the locations and directions of the two labeled minutiae in (a) from eight aligned impressions.

the distortion is small; However, in the outer region, the pressure is small and slippage can be large which leads to large distortion. Computing the average location ($\bar{S}_k$) and average direction ($\bar{D}_k$) for each minutia smoothes out the noise.

6.2.4 Modeling the Variability of Partial Prints

The third and the final component of the compound stochastic model accounts for minutiae variability due to partial prints in multiple acquisitions of a finger. The partial print region can be determined as the minimal ellipse discussed in Chapter 4.

For the impression $\mathcal{F}(f,l)$, the following parameters uniquely determine the minimal ellipse discussed in section 4.6: the area $[A(f,l)]$, length of major axis $[p_1(f,l)]$, the orientation $[\theta(f,l)]$, and the center $[c(f,l)]$. In our experiments, the ratio of the lengths of the major to minor axes of each ellipse (say, r_0) is fixed. Thus, the effective ellipse parameters reduce to the triplet $E(f,l) \equiv \{A(f,l), \theta(f,l), c(f,l)\}$. Denote the collection of all the ellipse parameters for all the fingerprint impressions in the fingerprint database by $\mathcal{E}$, and let $\mathcal{T} \equiv \{T^{-1}(f,l)\}$, where $T(f,l)$ is a Procrustes transformation [29] used to align $\mathcal{F}(f,l)$ to $\mathcal{F}(f,1)$. With these quantities defined, a conditional minutiae synthesis method can be applied to estimate the fingerprint individuality, which is described in the following section.

6.3 Conditional Minutiae Synthesis

In Chapter 4, an analytical method, namely the Poisson model, based solely on the interclass variability was introduced for the assessment of fingerprint individuality. When additional sources of minutiae variability are considered, larger number of parameters are involved in calculating fingerprint individuality requiring a more elegant analytical formula. An alternative synthesis method, namely the conditional minutiae synthesis method, is developed in this section so that all sources of variability in the stochastic model, i.e., both interclass and intraclass variabilities, are considered in the simulation procedure and the simulated minutiae sets are matched by a matcher to find the PRC. This method replaces the step of fingerprint individuality calculation by matching the simulated minutiae sets.

A minutiae set is synthesized for a finger consisting of a pre-specified number (m) of minutiae. In order to synthesize this minutiae set, the minimal ellipse that encompasses all minutiae needs to be simulated first. The areas of best-fitting ellipses are found, in general, to be strongly correlated with the total number of minutiae, i.e., $m(f,l)$, in a fingerprint impression. As an example, an illustration is given based on FVC 2002 DB1 database. In this database the ellipse area $A(f,l)$ is positively correlated with $m(f,l)$ (see Figure 6.5), while the other variables had no significant correlation. Consequently, a quadratic polynomial in $m(f,l)$, i.e., $Q_0(m(f,l))$, was fitted to the scatter plot of $(m(f,l), A(f,l))$ (Figure 6.5). The residuals from the fit were found to follow a normal distribution with mean μ and standard deviation σ_0, where $\mu = 4.5$ and $\sigma_0 = 1.4 \times 10^4$. As a consequence, the area of a partial print with a fixed number of minutiae can be simulated.

An illustration of the conditional synthesis technique is given based on the FVC 2002 DB1 database. Panels (a) and (b) in Figure 6.6 give an instance of a finger f. Panel (a) shows the constructed master set with m minutiae, and panel (b) shows the corresponding minimal ellipse. The procedure of minutiae synthesis is as follows: (i) A random ellipse is generated whose area is a sample from a normal distribution with mean $Q_0(m)$ and

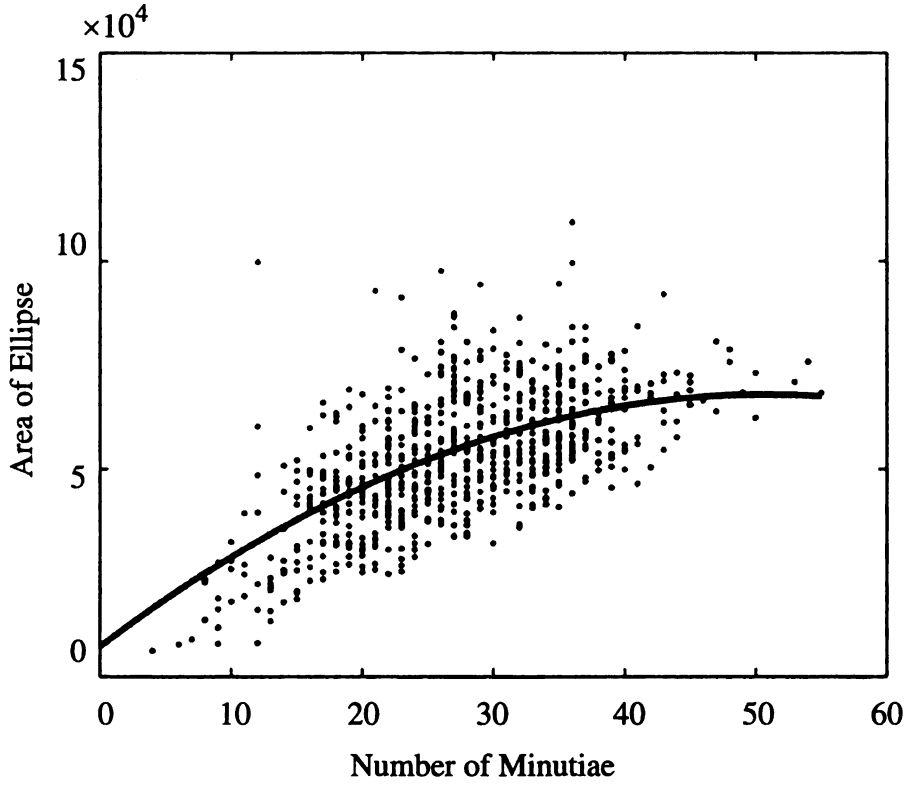


Figure 6.5: Scatter plot of the area of ellipse $[A(f, l)]$ versus the total number of minutiae $[m(f, l)]$, and the fitted quadratic polynomial for the FVC 2002 DB1 [26].

standard deviation σ_0, and the orientation and center of the ellipse are randomly selected from the second and third components of $\mathcal{E}$. (ii) a minutia center is then generated from the mixture model of master f. (iii) A deviation is generated according to the local perturbation model, and compounded to the generated minutiae center from step (ii). This synthesized minutiae is retained if its location lies within the ellipse in step (i) and rejected otherwise. Steps (ii) and (iii) are repeated until m synthesized minutiae fall inside the ellipse. In Figure 6.6, panel (c) shows the m synthesized minutiae centers from the mixture model, whereas panel (d) shows the synthesized minutiae after compounding with the local perturbation model. (iv) Finally, the m minutiae are transformed by a random rigid transformation from $\mathcal{T}$ to form the synthesized impression (Figure 6.6 (e)). During this synthesis procedure, the inter-minutiae distances in an impression are controlled so that they are no smaller than

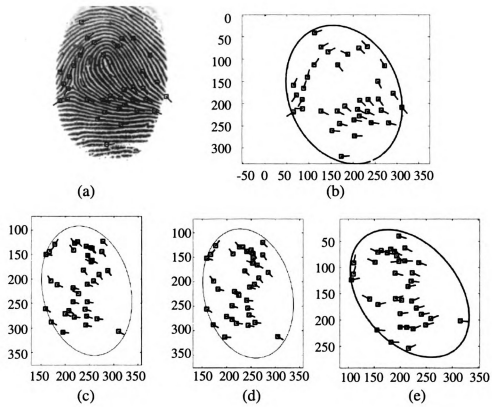


Figure 6.6: Simulating $m_0 = 36$ minutiae for FVC 2002 DB1: (a) Finger impression, (b) Minutiae and minimal ellipse for the impression, (c) Random ellipse and synthesized minutiae centers from the mixture model, (d) Synthesized minutiae after compounding with local perturbations, and (e) Synthesized impression after rigid transformation.

the minimal inter-minutiae distance estimated from the empirical databases. Note from Figures 6.6 (c-e) that the synthesized fingerprint has similar clustering characteristics and dependence structure as the original fingerprint (Figure 6.6 (a)). The fingerprint synthesis procedure described above is able to simulate any number of synthetic impressions of a finger (as a minutia set) with a pre-determined number of minutiae (m) while preserving the clustering and dependence characteristics of the minutiae.

Given a query fingerprint Q with n minutiae and a template T with m minutiae, the probability that Q and T share exactly w minutiae is needed to estimate fingerprint individuality.

This probability is given by the expression

$$p(w|m,\, n) = P\{S(Q,T) = w \,|\, \#Q = n, \#T = m,\, I_Q \neq I_T\}, \tag{6.2}$$

where $S(Q,T)$ is the number of minutiae matches between Q and T as determined by a matcher.

The conditional synthesis technique, described earlier in this section, is applied to simulate fingerprint impressions so that each Q (respectively, T) has exactly n (respectively, m) minutiae. Without loss of generality, we assume $m = n$. Corresponding to each finger f in the database, multiple synthetic impressions are generated based on the conditional synthesis technique. The resulting synthetic database is denoted as follows,

$$\{\mathcal{F}^*(f,l),\, l = 1,\, 2,\, ...,\, H,\, f = 1,\, 2,\, ...,\, F\},$$

where $\mathcal{F}^*(f,l)$ is the l-th synthetic impression from finger f. In order to obtain the distribution of the number of impostor matches for the synthetic database, the fingerprint matcher reported in [40] is applied to each pair of impostor fingerprints. A description of the matcher is described in section 6.4. To compute the probability of w matches, namely $p(w|m,\, n)$, the number of impostor pairs that resulted in w minutiae matches is counted, and this number is then divided by the total number of impostor pairs. Thus $p(w|m,\, n)$ is given by the following equation

$$p(w|m,\, n) = \frac{\sum_{l=1}^{H}\sum_{l'=1}^{H}\sum_{f=1}^{F}\sum_{\substack{f'=1 \\ f\neq f'}}^{F} I_w\{(f,l),(f',l')\}}{F(F-1)H^2}, \tag{6.3}$$

where $I_w\{(f,l),(f',l')\}$ is 1 if $S(\mathcal{F}^*(f,l),\mathcal{F}^*(f',l'))$ equals w, and 0 otherwise. Note that $p(w|m,\, n)$ provides an estimate of the probability in Equation 6.2 based on the synthetic database.

Although the compound stochastic models incorporate more sources of minutiae variability compared to the hyper-mixture model, they have limited capability to estimate fin-

gerprint individuality. While Equation 6.3 gives reliable estimates of fingerprint individuality for small and moderate values of w, the estimate obtained for large w is not reliable. In the case of large w, the true value of Equation 6.2 is extremely small. As a consequence, Equation 6.3 gives zero as the estimate of Equation 6.2, due to limitations of numerical precision.

6.4 Description of Matchers

6.4.1 Matcher for Master Construction

During master construction, in order to find correspondence between multiple impressions of the same finger, an adaptive elastic string matcher, developed by Jain et al. [1], was applied. One can reference the original article for details of this matcher. After applying the matcher, the reported correspondence was manually checked to remove false matched minutiae pairs and to add true matched minutiae pairs that were not detected by the matcher.

6.4.2 Matcher for Synthesized Minutiae Matching

In order to match the minutiae sets synthesized from the compound stochastic models, a matcher developed by Ross et al. [40] was used. Unlike the matcher for construction above which utilizes the fingerprint ridges besides minutiae, this matcher utilizes only minutiae information. The matcher was implemented as follows. First, two minutiae, one query minutia and one template minutia were selected to form a reference minutiae pair. form a reference minutiae pair. Then the two minutiae sets were aligned by translating the query minutiae set, so that the reference minutiae pair had identical locations. Next, the query minutiae set was rotated about its reference minutiae, which maximized the number of minutiae that were paired (i.e., fell within a tolerance window) with the template minutiae set. The above procedure was repeated till all possible reference minutiae pairs were

considered, and the maximum number of matched minutiae pairs was reported.

6.5 Summary

A family of compound stochastic models was developed to account for both interclass variability and intraclass variability of fingerprints. Based on the models, a conditional minutiae synthesis method was introduced to simulate minutiae sets, which were then compared between simulations of different models (i.e., fingers) with a matcher. Based on the obtained probability distribution of the number of matched minutiae pairs, fingerprint individuality was then reported as the probability that the matched minutiae pairs exceeds a given threshold.

CHAPTER 7

Experimental Results

7.1 Fingerprint Databases

The methodology for assessing the individuality of fingerprints is validated on three target populations, namely, the NIST 2000 SD 4 [30] (denoted as "NIST" in this chapter), FVC 2002 DB1 (denoted as "DB1" in this chapter) and FVC 2002 DB2 [26] (denoted as "DB2" in this chapter) fingerprint databases. All the three databases are publicly available. The NIST database contains 2,000 8-bit gray scale fingerprint image pairs of size 512-by-512 pixels. Because of the relative large size of the images in the NIST, the first image of each pair was used for statistical modeling. Minutiae could not be automatically extracted from two images of the NIST due to poor quality. Thus, the total number of fingerprints used in the experiments for NIST is $F = 1,998$. For the FVC 2002, the DB1 impressions (image size $= 388 \times 374$) are acquired using the "TouchView II" optical sensor by Identix, while the DB2 impressions (image size $= 296 \times 560$) are acquired using the "FX2000" optical sensor by Biometrika. Both DB1 and DB2 databases consist of fingerprints of 100 different fingers with 8 impressions per finger. Because of the small size of the DB1 and DB2 databases, the minutiae consolidation procedure was adopted to obtain a master minutiae set for each finger. The mixture models were subsequently fitted to each master. Figure 7.1 shows two

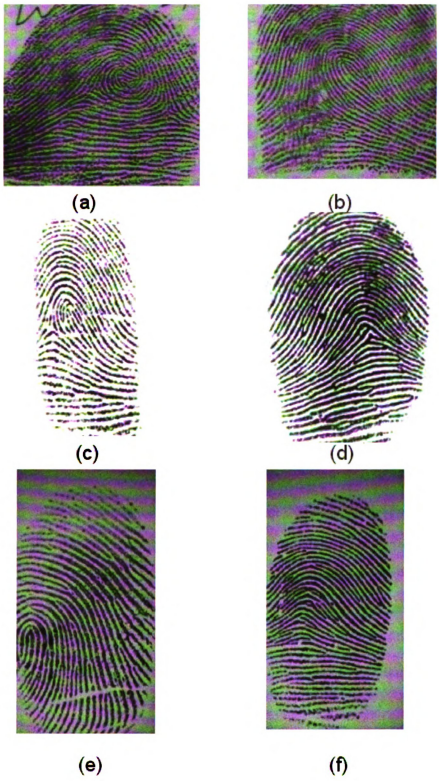


Figure 7.1: Examples of fingerprint images from different databases. Images (a-b) are from NIST database [30]; Images (c-d) are from DB1 and images (e-f) are from DB2 [26] .

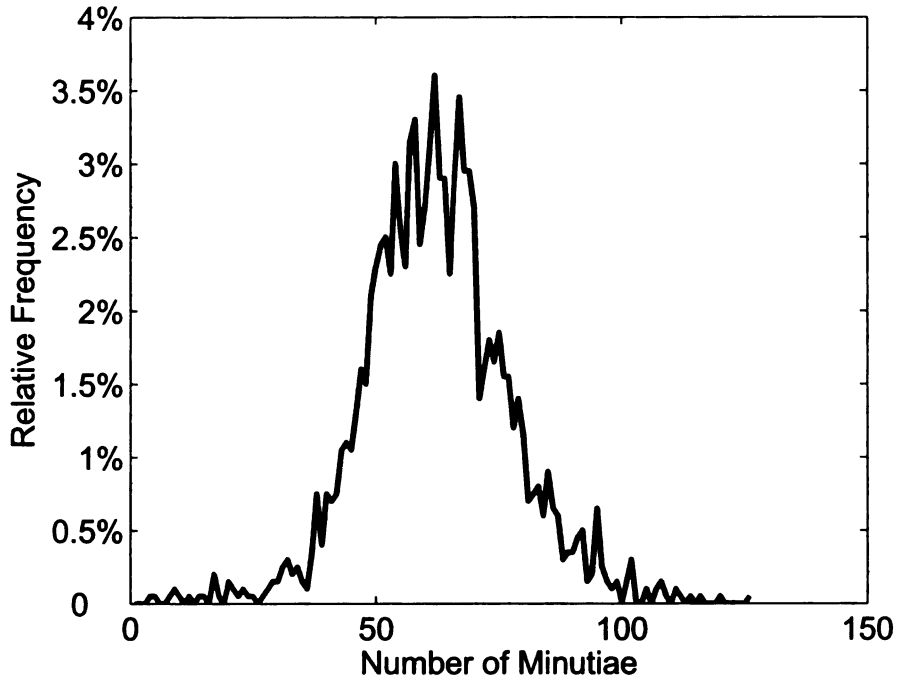
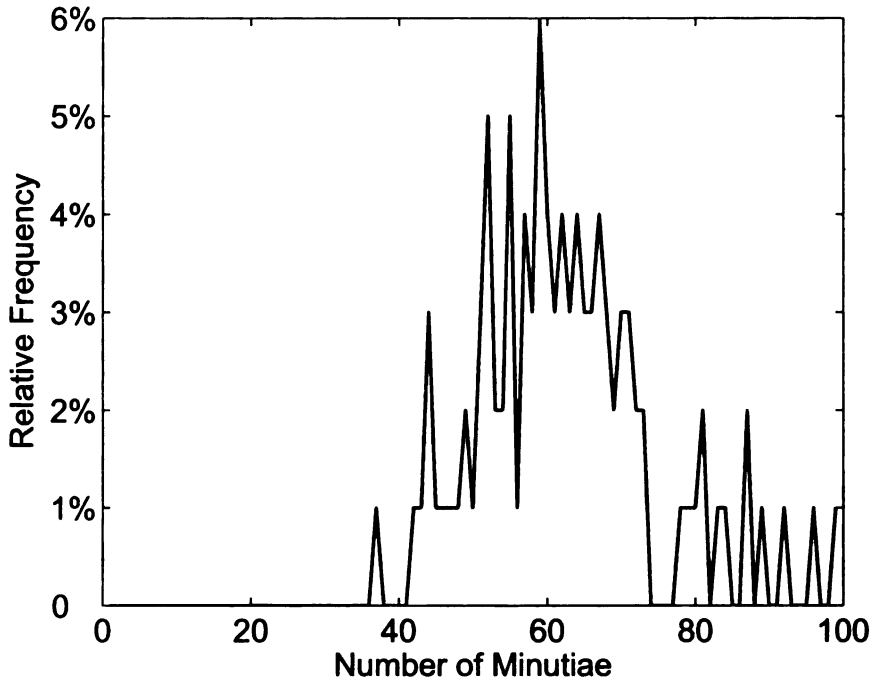


Figure 7.2: Empirical distribution of the number of minutiae (m,n) in the NIST database. The average number of minutiae is 62.

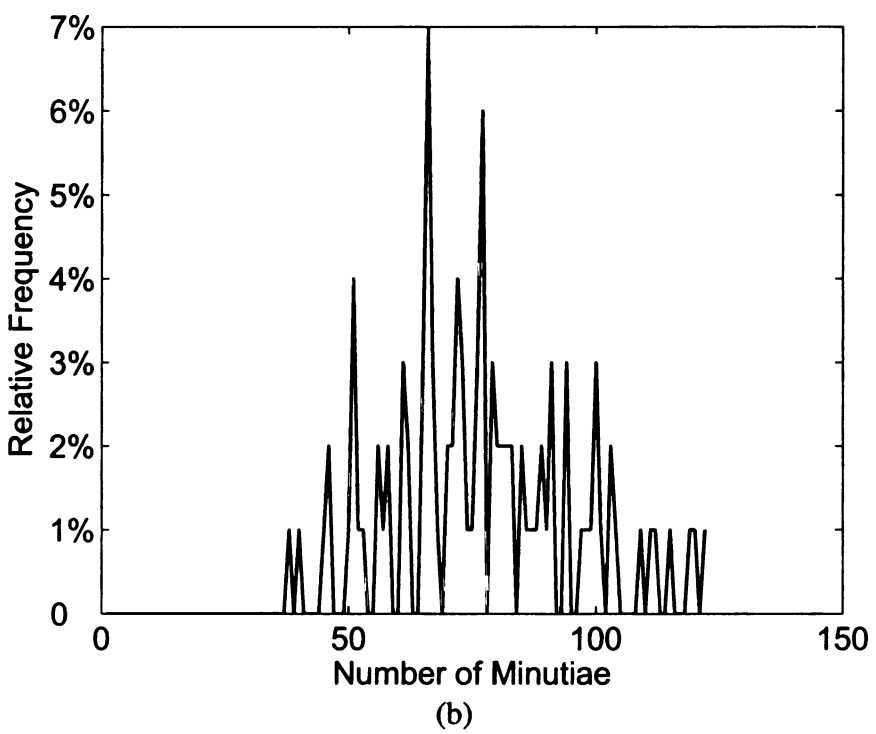
fingerprint images from each of the three databases.

The distribution of the number of minutiae (m, n) for the images in NIST is shown in Figure 7.2 and those of the master minutiae sets from DB1 and DB2 are given in Figure 7.3, panels (a) and (b), respectively (The distribution of m and the distribution of n are identical, and hence only one histogram is shown). The average number of minutiae for the images in NIST and the master minutiae sets in DB1 and DB2 databases are approximately 62 , 63 and 77, respectively.

Based on the three databases, experiments were performed to validate the fingerprint individuality models developed in this thesis.



(a)



(b)

Figure 7.3: Empirical distributions of the number of minutiae (m,n) in the master prints constructed from (a) DB1 database, and (b) DB2 database. Average number of minutiae in the master minutiae set for the two databases are 63 and 77, respectively.

7.2 Fitting the Mixture Models

The best fitting mixture model (see Equation 3.5) was found for minutiae in each fingerprint and the goodness-of-fit test was applied to each database. The results for the goodness of fit for the mixture model as well as those for the uniform model (see Equation 3.22) are reported in Tables (7.1- 7.3) with $W = 2, 4, 10, 20, 50$ and $V = 4, 6, 9$. For all the three databases, the numbers of fingerprint images with p-values above (corresponding to acceptance of H_0) and below the threshold 0.05 (corresponding to rejection of H_0) were computed. The results show that the mixture model is generally a better fit to the observed minutiae distribution compared to the uniform. For example, when $W = 10$ and $V = 4$, the mixture is a good fit to $1,948$ out of $1,998$ images from the NIST (corresponding to p-values above 0.05) based on the Freeman-Tukey test. For the Chi-square test, this number is $1,945$. In comparison, the uniform model is a good fit to only 360 and 352 images based on the Freeman-Tukey and Chi-square tests, respectively.

7.3 Fitting the Hyper-mixture Models

Hyper-mixture model assumes that there are N^* clusters in the target population, which can be estimated based on the gap statistic G_N. The gap statistic G_N as a function of the number of clusters for the NIST was shown in Figure 5.1, and the plots for the DB1 and DB2 databases are shown in Figure 7.4. Based on these figures, N^* for the NIST, DB1 and DB2 databases are estimated as 9, 12 and 33, respectively.

7.3.1 A Check to See if the Clusters of Mixture are Meaningful

Replicability is an important feature for a meaningful cluster analysis. Ideally, the cluster analysis is meaningful when re-performing it on a new sample produces similar results as the original clustering. In reality, if the variation in the clusters is within a reasonable limit, the cluster analysis can be considered to be reliable. To evaluate reliability of the hyper-

NIST					
Chi-square Test ($\alpha = 0.05$)			Freeman-Tukey Test ($\alpha = 0.05$)		
(W,V)	Mixture Accepted	Uniform Accepted	Mixture Accepted	Uniform Accepted	Average# Blocks
(2,4)	1,567	708	1,542	714	6.2
(4,4)	1,914	153	1,909	179	7.2
(10,4)	1,945	352	1,948	360	8.2
(20,4)	1,935	390	1,938	397	8.3
(50,4)	1,940	405	1,938	395	8.3
(2,6)	987	425	911	416	7.0
(4,6)	1,880	135	1,877	148	7.4
(10,6)	1,942	331	1,939	327	8.2
(20,6)	1,937	396	1,937	395	8.3
(50,6)	1,939	403	1,936	392	8.3
(2,9)	981	462	933	456	7.4
(4,9)	1,882	137	1,865	150	7.5
(10,9)	1,944	325	1,945	330	8.2
(20,9)	1,938	393	1,942	396	8.3
(50,9)	1,939	407	1,937	392	8.3

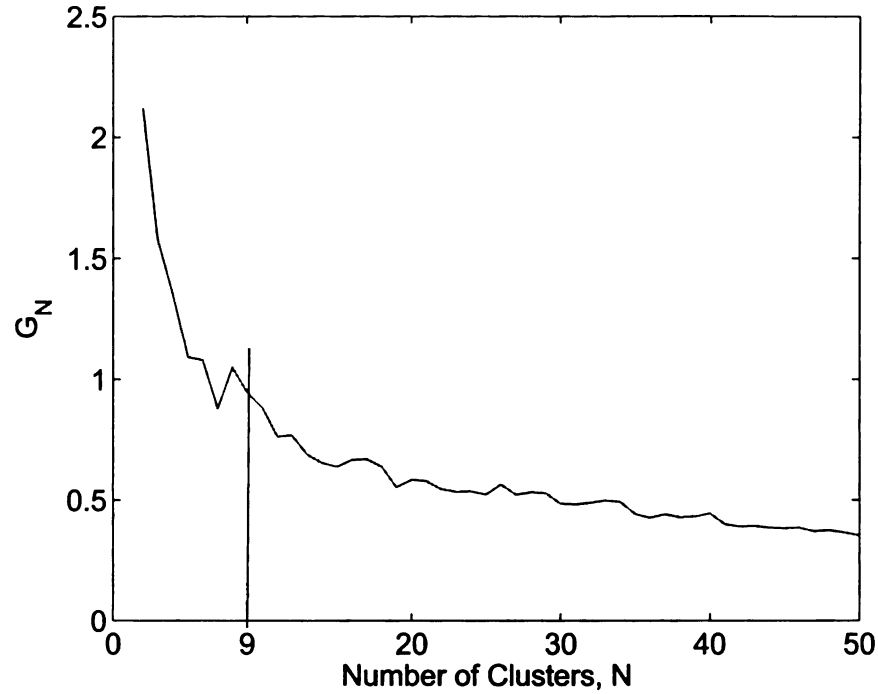
Table 7.1: Results of the Freeman-Tukey and Chi-square tests for testing the goodness-of-fit of the mixture and uniform models on NIST. (W, V) means the whole minutiae location space S is partitioned into W equal-size rows by W equal-size columns and the minutiae direction space D is partitioned into V equal-size blocks initially, prior to merging blocks of insufficient minutiae with their neighboring blocks. Entries correspond to the number of fingerprints in each database with p-values above 0.05. The total number of mixture models that are tested in NIST is 1,997 because three out of the 2,000 fingerprints don't have enough minutiae (at least 5) for the tests.

DB1					
Chi-square Test ($\alpha = 0.05$)			Freeman-Tukey Test ($\alpha = 0.05$)		
(W,V)	Mixture Accepted	Uniform Accepted	Mixture Accepted	Uniform Accepted	Average# Blocks
(2,4)	57	14	55	9	4.7
(4,4)	60	2	57	2	4.8
(10,4)	90	0	89	0	4.4
(20,4)	92	0	90	0	4.8
(50,4)	93	0	92	0	4.9
(2,6)	38	10	36	7	5.1
(4,6)	48	1	45	0	5.0
(10,6)	87	0	85	0	4.6
(20,6)	91	0	89	0	4.8
(50,6)	92	0	91	0	4.8
(2,9)	41	14	41	12	5.4
(4,9)	46	0	47	0	5.4
(10,9)	86	0	87	0	4.6
(20,9)	91	0	89	0	4.8
(50,9)	93	0	92	0	4.9

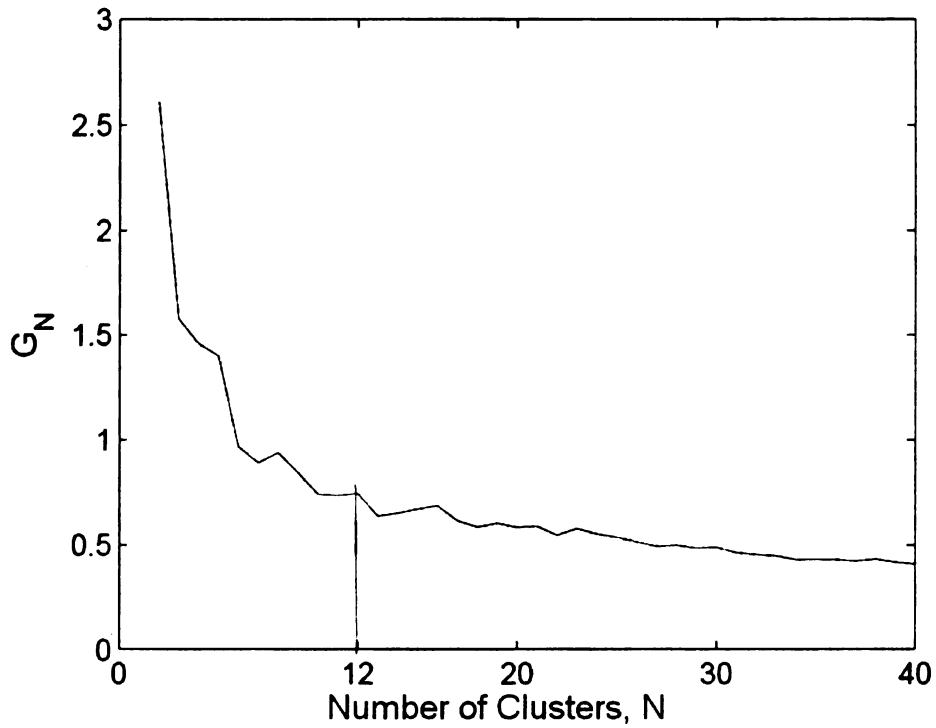
Table 7.2: Results of the Freeman-Tukey and Chi-square tests for testing the goodness-of-fit of the mixture and uniform models on DB1. (W, V) means the whole minutiae location space S is partitioned into W equal-size rows by W equal-size columns and the minutiae direction space D is partitioned into V equal-size blocks initially, prior to merging blocks of insufficient minutiae with their neighboring blocks. Entries correspond to the number of fingerprints in each database with p-values above 0.05. The total number of mixture models that are tested in this database is 100 since all master minutiae sets have total number of minutiae more than 5.

DB2					
Chi-square Test ($\alpha = 0.05$)			Freeman-Tukey Test ($\alpha = 0.05$)		
(W,V)	Mixture Accepted	Uniform Accepted	Mixture Accepted	Uniform Accepted	Average# Blocks
(2,4)	47	4	44	8	6.9
(4,4)	67	2	64	5	7.4
(10,4)	92	1	94	2	8.8
(20,4)	94	5	94	5	10.2
(50,4)	95	3	93	4	10.5
(2,6)	37	6	35	6	7.5
(4,6)	57	2	52	2	8.0
(10,6)	95	2	94	2	9.0
(20,6)	96	5	95	5	10.2
(50,6)	95	4	93	4	10.5
(2,9)	27	4	28	6	8.0
(4,9)	37	1	38	1	8.5
(10,9)	94	2	93	3	9.1
(20,9)	95	4	94	5	10.2
(50,9)	95	4	93	4	10.5

Table 7.3: Results of the Freeman-Tukey and Chi-square tests for testing the goodness of fit of the mixture and uniform models on DB2 database. (W, V) means the whole minutiae location space S is partitioned into W equal-size rows by W equal-size columns and the minutiae direction space D is partitioned into V equal-size blocks initially, prior to merging blocks of insufficient minutiae with their neighboring blocks. Entries correspond to the number of fingerprints in each database with p-values above 0.05. The total number of mixture models that are tested in this database is 100 since all master minutiae sets have total number of minutiae more than 5.



(a) DB1



(b) DB2

Figure 7.4: Estimating the number of clusters for FVC database.The estimated number of clusters for DB1 and DB2 are 9 and 12, respectively. The horizontal axis shows the number of clusters N and the vertical axis is the value of gap statistic at N.

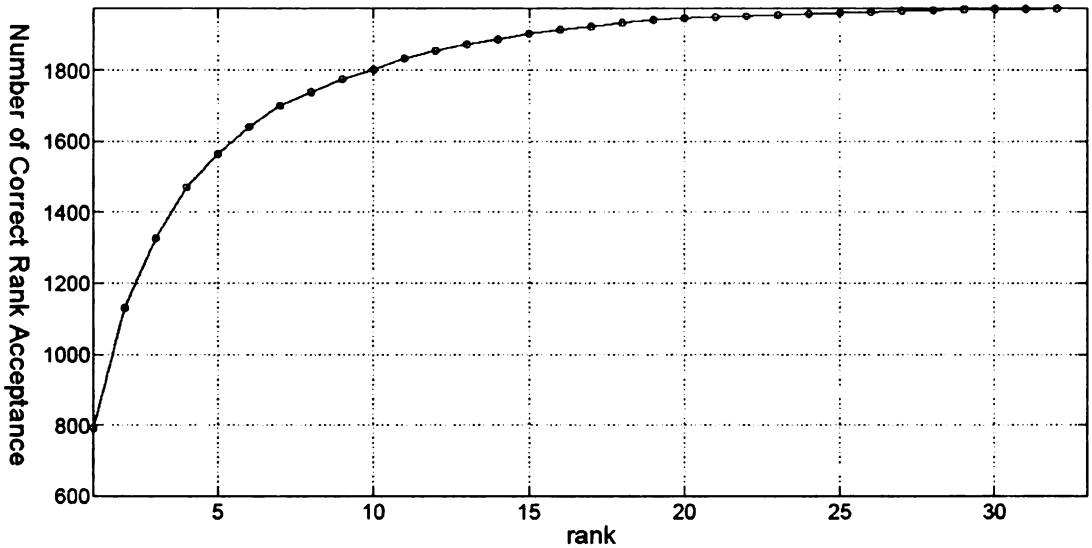


Figure 7.5: Analysis of Hyper-mixture model using the NIST. The figure shows the cumulative distribution of the ranks for the fingerprints in the validation set. The horizontal axis shows the rank, and the vertical axis shows the number of fingerprints that have a rank less than or equal to a given rank.

mixture model, a check [4] was implemented on the NIST database. In this database, there are two impressions for each finger. Thus for the total of 2,000 fingers, there are 2,000 pairs of images. Most of the pairs (1,975 out of the 2,000 pairs) share at least 3 minutiae features (the fitting process of the mixture models requires at least 3 minutiae features), but 25 pairs do not. The check was performed only for the minutiae shared by each of the 1,975 pairs.

First, the two databases were partitioned into two sets, namely a development set and a validation set. Each set has one and only one impression for each finger. When applying the hyper-mixture model on the development set, the fingerprints were grouped into 33 clusters. For each fingerprint in the validation set, a Hellinger distance between the fingerprint and each of the 33 clusters was calculated. A fingerprint is similar to a cluster if the Hellinger distance between them is small. Ideally, the Hellinger distance between a fingerprint and a cluster containing its counterpart (i.e., the fingerprint of the same finger) should be the smallest, which is given a rank one. However, in reality, it is not always true. Therefore, a rank k can be given to each fingerprint, which describes that the Hellinger dis-

tance between the fingerprint and the cluster containing its counterpart is the kth smallest among the Hellinger distances for the 33 clusters. The rank k characterizes quantitatively the reliability of the clustering method, i.e., the smaller the k, the more reliable the clustering method is. Figure 7.5 shows the cumulative distribution of the ranks for the fingerprints in the validation set. The horizontal axis shows the rank, and the vertical axis shows the number of fingerprints that have a rank less than or equal to a given rank. As the curve shows, 1,564 out of the 1,975 mixture models in the validation set were within rank 5, indicating good reliability of the clustering procedure, that is, the clustering is capturing some aspects of inter-class variability in the population.

7.3.2 Evaluation of Hyper-mixture Models on Assessment of Fingerprint Individuality

In order to improve computing efficiency, hyper-mixture model was introduced to replace the individual mixture models with an average density of similar mixture models. Obviously, this modification should not significantly change the individuality estimates if the clusters represent original characteristics of the population. Therefore, it is necessary to compare the hyper-mixture model with the individual mixture models. Experiments were carried out on the three databases. The PRCs estimated by the hyper-mixture model (Equation 5.8) are shown in the first row in each block of Table 7.4. The PRCs estimated by the individual mixture models, on the other hand, are calculated according to Equation 7.1, and are shown in the middle row in each block of Table 7.4.

$$\overline{\mathrm{PRC}}_{Mixture} = \frac{\sum_{\{(i,j)\in\{1,2,\ldots,L,\},i\neq j\}} p(w\,;\,i,j)}{L\times(L-1)}, \tag{7.1}$$

where

$$p(w\,;\,i,j) = e^{-\lambda(i,j)}\frac{\lambda(i,j)^w}{w!}, \tag{7.2}$$

and $\lambda(i,j)$ is estimated by Equation 4.13 (i.e., Poisson model) for fingers i and j.

Database	(m, n, w)	Method	PRC	PRC Ratio
NIST	$(62, 62, 12)$	Individual Mixture Model	6.1×10^{-3}	1
NIST	$(62, 62, 12)$	Hyper-mixture Model	4.1×10^{-4}	1/14
NIST	$(62, 62, 12)$	Random Grouping	1.0×10^{-5}	1/554
DB1	$(63, 63, 12)$	Individual Mixture Model	2.7×10^{-2}	1
DB1	$(63, 63, 12)$	Hyper-mixture Model	5.9×10^{-3}	1/5
DB1	$(63, 63, 12)$	Random Grouping	5.5×10^{-5}	1/491
DB2	$(77, 77, 12)$	Individual Mixture Model	3.2×10^{-2}	1
DB2	$(77, 77, 12)$	Hyper-mixture Model	8.4×10^{-3}	1/4
DB2	$(77, 77, 12)$	Random Grouping	1.6×10^{-3}	1/20

Table 7.4: Comparison of PRCs estimated from three different methods: (I) Individual mixture models, (II) Hyper-mixture models, and (III) Random grouping. The right-most column shows the relative ratios of the PRCs.

Comparing the top two rows in each block of Table 7.4, it appears that the PRCs from the individual mixture model and those from the hyper-mixture model are similar. However, without a reference, it is difficult to see how similar they are. That is why a third row has been included in each block of Table 7.4, which is the PRCs from random grouping, i.e., while keeping the same number of clusters and the same number of fingerprints in each cluster as in the hyper-mixture model, the members in each cluster are randomly selected from the entire database with no replacement. In each block of Table 7.4, as the right-most column shows, the PRCs from random grouping (the bottom row) is always at least 20 times smaller than those from the individual mixture model (the top row); whereas the PRCs from the hyper-mixture model (the middle row) is at most 14 times smaller than those from the individual mixture model. Therefore, compared with the PRCs from random grouping, those from the hyper-mixture model are more similar to those from the individual mixture models, which gives support to the clustering based on hyper-mixture models.

Database	(m, n, w)	N^*	Mean Fingerprint area	Mean λ Hyper-Mixture Model	$\overline{\text{PRC}}_\alpha$
NIST	$(62, 62, 12)$	33	2.5×10^5	2.5	4.1×10^{-4}
DB1	$(63, 63, 12)$	9	1.2×10^5	5.1	5.9×10^{-3}
DB2	$(77, 77, 12)$	12	1.8×10^5	5.14	8.4×10^{-3}

Table 7.5: The number of clusters, N^*, as well as mean λ and $\overline{\text{PRC}}_\alpha$ based on the hyper-mixture models for the three databases.

7.4 Assessment of Fingerprint Individuality with the Hyper-mixture Models

For the three databases, the agglomerative clustering procedure in Chapter 5 was carried out for the fitted mixture models to estimate the number of clusters, i.e., N^*. The results are shown in Table 7.5, which also gives the following quantities for each database: the numbers of minutiae in master minutiae sets (namely m and n), the fingerprint area, and the parameter λ for the mixture model representing the expected mean number of impostor matches from the mixture models. The last column in Table 7.5 gives the mean PRC, $\overline{\text{PRC}}_\alpha$, corresponding to $w = 12$ based on the hyper-mixture model (i.e., obtaining 12 or more matches). The parameter α was chosen to be 0.05 to correspond to the 5% trimmed mean of the probabilities. Note that while the mean values of m and n for the NIST and DB1 databases are similar, the mean of λ for DB1 is much larger than that for NIST database, resulting in a much larger mean PRC for DB1 compared to that for NIST database. Comparing DB1 and DB2, the mean λ remains the same but the mean value of minutiae in DB2 is much larger than that in DB1 (77 vs. 63). A larger number of total minutiae implies a greater chance of obtaining a random match and hence a larger value for the PRC.

A comparison of $\overline{\text{PRC}}_\alpha$ ($\alpha = 0.05$) was carried out for two different choices of λ for the Poisson model: (i) the λ that was derived from the cluster of the mixture models (see

Database	(m, n, w)	Hyper-Mixture Model	Uniform	Empirical
NIST	$(62, 62, 12)$	4.1×10^{-4}	2.9×10^{-7}	3.4×10^{-3}
DB1	$(63, 63, 12)$	5.9×10^{-3}	1.0×10^{-4}	1.4×10^{-2}
DB2	$(77, 77, 12)$	8.4×10^{-3}	8.4×10^{-5}	1.9×10^{-2}

Table 7.6: A comparison of the $\overline{\text{PRC}}_\alpha$ obtained from the mixture and uniform models based on mean m, n with empirical values.

Database	(m,n,w)	Hyper-Mixture Model Mean λ	Uniform Model Mean λ	Empirical
NIST	$(62, 62, 12)$	2.5	1.5	7.1
DB1	$(63, 63, 12)$	5.1	3.0	8.0
DB2	$(77, 77, 12)$	5.1	3.0	8.6

Table 7.7: A comparison of the mean number of matches obtained from the mixture and uniform models and empirical matches.

Equations 4.29, 5.6 and 5.8), and (ii) the λ that was derived from the uniform model (see Equations 5.9 and 5.8). The values of m and n are taken to be the mean in each database. The $\overline{\text{PRC}}_\alpha$'s obtained from the mixture model are reported in Table 7.5. Table 7.6 gives the $\overline{\text{PRC}}_\alpha$ from the mixture and uniform models corresponding to $w = 12$ from the NIST and FVC 2002 based on the fingerprint area. Note that the fingerprint individuality estimates using the mixture models are at least one order of magnitude higher compared to the uniform model. It is because when minutiae from the query and template have similar clustering tendencies, a larger number of random matches will arise compared to the uniform model. The empirical PRCs for $w = 12$ in each database is the proportion of impostor pairs with 12 or more matches among all pairs that have m and n values within ± 5 of the mean. Using the matcher reported in [40], the n query minutiae (S_i^Q, D_i^Q), $i = 1, 2, \ldots, n$, are optimally aligned with the m template minutiae (S_j^T, D_j^T), $j = 1, 2, \ldots, m$, to obtain the best number of matches between each impostor pair. The mean number of impostor minutiae matches for each database is reported in Table 7.7. Note that the empirical number of matches and the PRCs are closer to the values derived from the mixture models compared

Database	(m, n, w)	N^*	Mean λ	Hyper-Mixture	Uniform
			Hyper-Mixture Model		
NIST	$(46, 46, 12)$	33	1.9	2.3×10^{-6}	5.0×10^{-10}
DB1	$(46, 46, 12)$	9	2.7	5.6×10^{-5}	2.8×10^{-7}
DB2	$(46, 46, 12)$	12	1.8	4.1×10^{-6}	3.2×10^{-9}

Table 7.8: A comparison between $\overline{\text{PRC}}_\alpha$ obtained from the mixture and uniform models for $m = n = 46$ and $w = 12$.

Database	(m,n)	Mean Overlapping Area (pixel^2)	M
NIST	(52,52)	112,840	413
DB1	(51,51)	71,000	259
DB2	(63,63)	110,470	405

Table 7.9: Table giving the mean m and n in the overlapping area, the mean overlapping area and the value of M for each database.

to those from the uniform model, suggesting the appropriateness of the mixture models in representing the distribution of minutiae.

Since the mathematical model for the PRC was developed for any combination of m, n and w, the trimmed mean PRC value corresponding to $m = n = 46$ and $w = 12$ can be found for the three databases as an example. These PRCs are given in Table 7.8 for the mixture and uniform distributions. Note, again, that the PRCs derived from the mixture model are orders of magnitude higher compared to those from the uniform model.

In the following paragraphs, the results obtained from the proposed methodology in this thesis are compared with those of Pankanti et al. [35], introduced in Section 2.3. There are two main differences between the experiments presented in this section and the ones discussed in the previous paragraphs (i.e., Tables 7.6 and 7.8). First, the "corrected" uniform model of Pankanti et al. [35], instead of the fully uniform model, is considered (the "corrected uniform model" was discussed in section 2.3.3). Second, the overlapping area between the query and the template, instead of the whole fingerprint area, is considered. In

Database	(m, n, w)	Empirical	Mixture Model	Pankanti
NIST	$(52, 52, 12)$	7.1	3.1	1.2
DB1	$(51, 51, 12)$	8.0	4.9	2.4
DB2	$(63, 63, 12)$	8.6	5.9	2.5

Table 7.10: A comparison between the mean λ obtained from the mixture and uniform models and the mean number of matched minutiae from the empirical matches in the overlapping area.

Database	(m,n,w)	Empirical	Mixture Model	Pankanti
NIST	(52,52,12)	3.9×10^{-3}	4.4×10^{-3}	4.3×10^{-8}
DB1	(51,51,12)	2.9×10^{-2}	1.1×10^{-2}	4.1×10^{-6}
DB2	(63,63,12)	6.5×10^{-2}	1.1×10^{-2}	4.3×10^{-6}

Table 7.11: A comparison between fingerprint individuality estimates using the (a) Poisson and mixture models, and (b) the corrected uniform model of Pankanti et al. [35].

other words, the overlapping area model was utilized to estimate fingerprint individuality. Since mixture models were used in the overlapping area model instead of hyper-mixture models, the comparison is between mixture models and the "corrected uniform model".

In order to compare the fingerprint individuality estimates using the mixture model and the model by Pankanti et al. [35], we first need to find the overlapping area between the query and template. This is done as follows. The query and template fingerprints in the NIST and FVC databases are first aligned using a Procrustes transformation [29] based on the minutiae correspondence obtained from the matcher described in section 6.4.2. Then, bounding boxes encompassing all minutiae points in the query and template fingerprints are determined. The overlapping area between the two bounding boxes is taken to be the overlapping area between the query and template fingerprints. Thus the fingerprint individuality estimates presented here are dependent on the matcher. In order to compute the Poisson probabilities, overlapping area model is used. Meanwhile, the fingerprint individuality estimates based on the corrected uniform model is also obtained. Table 7.10 gives the mean λ of the Poisson model in the overlapping area for the NIST and FVC databases.

The mean λ's (i.e., the theoretical mean numbers of matches) obtained from the hyper-mixture density model are closer to those from the empirical results, compared to those from the corrected uniform model, which illustrates superiority of the mixture model over the corrected uniform model. Table 7.11 shows the PRCs corresponding to the mean m and the mean n, compared with the empirical PRCs. The empirical PRC is computed as the proportion of impostor pairs with 12 or more matches among all pairs with m and n values within ± 5 of the mean in the overlapping area. Note that as m or n or both increase, the values of PRC for both models become large because it becomes much easier to obtain spurious matches for larger m and n values. More important, however, is the fact that the Poisson probabilities based on the mixture models are, again, orders of magnitude larger than those from the corrected uniform model. Also the PRCs corresponding to the hyper-mixture model are closer to the empirical counterparts, compared to those corresponding to the corrected uniform model, confirming again the reliability of the mixture models.

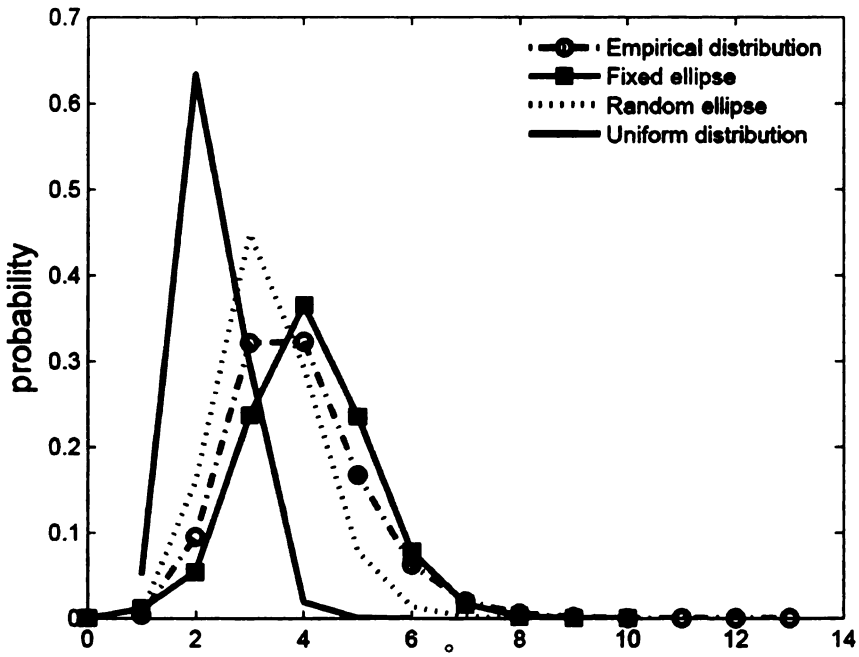
7.5 Estimation of Fingerprint Individuality with the Compound Stochastic Model

Instead of all the three databases as in other experiments, only two databases, namely DB1 and DB2, were used to demonstrate effectiveness of the proposed compound stochastic models. There are two reasons not to use NIST. First, in the NIST database, there are only two impressions for each finger; whereas in DB1 and DB2, each finger has eight impressions. Therefore the NIST might not have sufficient data to model the local perturbation for each finger. Second, the ink fingerprints in the NIST database have very large area and cover most of the fingertips, which makes it difficult to show the effectiveness of the partial print model.

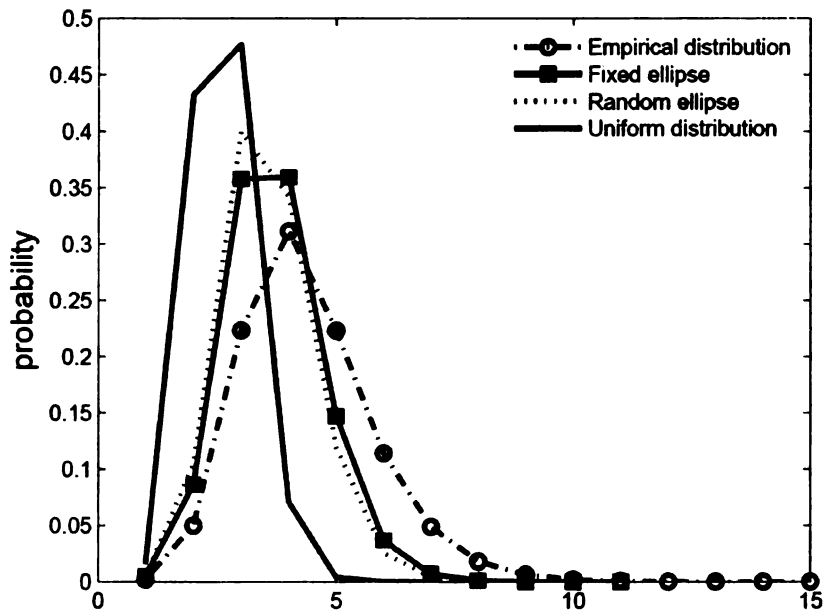
To validate the models, a synthetic database consisting of F fingers with L impressions per finger was generated. For finger f, a total of n (the actual number of consolidated

minutiae) minutiae were synthesized from the fitted mixture model (for minutiae centers) and the local perturbation model (for deviations from the minutiae centers). The parameterized ellipse for the l-th impression was then used to select a subset of the synthetic minutiae set. Subsequently, the rigid transformation $T(f, l)$ was used to obtain a synthetic impression. Since the ellipses used in this synthesis are the ellipses from the original fingerprint impressions, this simulation is called the fixed ellipse simulation. The distribution of the number of impostor minutiae matches for this synthetic database is obtained using the matcher described in [40]. This distribution is represented by the solid line with squares (□) labeled as "fixed-ellipse" in Figures 7.6 (a-b). Another synthetic database of F fingers with L impressions per finger was constructed using the conditional minutiae synthesis technique so that the number of minutiae for the l-th impression of finger f equals to the observed number of minutiae in the l-th impression of finger f, namely $m(f, l)$. When simulating the best fitting ellipse, the ratio of lengths of major to minor axes for the DB1 and DB2 were taken to be the mean values, namely 1.48 and 1.90, respectively. Since this synthesis simulates random ellipses from the partial-print model, it is called the random ellipse synthesis. The corresponding distribution of the number of impostor matches is represented in Figures 7.6 (a-b) by dashed lines labeled as "random ellipse". The distributions based on the real fingerprint impressions and based on the uniform distribution (i.e., with no clustering tendency) for the minutiae centers and deviations were also obtained (denoted by the dot-dashed lines with circles and by the solid lines, respectively). Note the close agreement between the impostor distributions of the synthesized and empirical databases, demonstrating the adequateness of the proposed compound stochastic models in representing the distribution of minutiae in the databases.

Fingerprint individuality estimates computed using Equation 6.3 (using 20 synthetic impressions per finger, i.e., $H = 20$) are given in Table 7.12 for DB1 and DB2 databases. For example, given a query and a template impression with 36 minutiae, the estimated probability of getting more than or equal to 12 matches is 7.2×10^{-7} for DB1. When



(a) Number of Minutiae Matches for FVC2002 DB1



(b) Number of Minutiae Matches for FVC2002 DB2

Figure 7.6: A comparison between the synthetic and empirical impostor distributions for the number of minutiae matches.

(m, n, w)	DB1	DB2
$(26, 26, 12)$	0*	0*
$(36, 36, 12)$	7.2×10^{-7}	0*
$(46, 46, 12)$	2.6×10^{-5}	7.9×10^{-7}

Table 7.12: The estimation of fingerprint individuality from the compound stochastic model when $w = 12$. Notice that when $m = n = 26$, none of two synthesized minutiae sets share 12 or more matched minutiae. Hence the probability cannot be estimated accurately for the case of $m = n = 26$.

$m = n = 26$, none of two synthesized minutiae sets share 12 or more matched minutiae, and thus the PRC is reported as zero. Since the fingerprint individuality (i.e., the PRC) is very small, to estimate such small probability, the sample size of the synthesized database needs to be large; otherwise a trivial result (i.e., zero) as for $m = n = 26$ will occur. The PRCs from the compound stochastic models are slightly smaller than those from the hyper-mixture models as shown in Table 7.8. For example, when $m = n = 46$, estimated probability of obtaining more than or equal to 12 matches for DB2 of hyper-mixture model and compound stochastic model are 4.1×10^{-6} and 7.9×10^{-7}, which can be explained by the multiple sources of minutiae variability from the compound stochastic model compared to single source of minutiae variability from the hyper-mixture model.

Table 7.13 gives fingerprint individuality estimates from the compound stochastic model for the "12-point match criteria" (see [2] and [35]) based on DB1 database. For comparison purposes, the fingerprint individuality estimates of the corrected uniform model by Pankanti et. al [35] are also given. Recall that the PRCs from the corrected uniform model are computed based on the number of minutiae in the query and template that occur in the overlapping area. The parameters n and m from the mixture models represent the total number of minutiae in a query and template, respectively. In order to make valid comparisons, the mean number of minutiae occurring in the overlapping area is found. When $n = m = 36$, this mean number is approximately 25. Consequently, the estimate 1.0×10^{-10} in Table 7.13 was calculated using Equation 2.13 from the corrected uniform

(m, n, w)	Compound Stochastic model	Corrected Uniform Model
$(36, 36, 12)$	7.2×10^{-7}	1.0×10^{-10}
$(46, 46, 12)$	2.6×10^{-5}	3.9×10^{-8}

Table 7.13: A comparison of fingerprint individuality estimates. n and m are the total number of minutiae in the query and the template, respectively. w is the number of matches between the query and template fingerprints.

model based on the combination $(25, 25, 12)$. Note that the estimates of the compound stochastic models are orders of magnitude higher compared to those of the corrected uniform model. This is due to the fact that the compound stochastic model accounts for the clustering tendency of minutiae via mixture model whereas the corrected uniform model does not, which indicates that the compound stochastic models gives a more realistic estimate of fingerprint individuality compared to the corrected uniform model. In conclusion, the experiment results indicates that the compound stochastic models consider more sources of minutiae variability but are not able to estimate the PRC at the tails of the distribution. However, if only the variability of minutiae in different fingerprints are considered, the hyper-mixture models are more computational efficient (no minutiae synthesis and matching is required) in fingerprint individuality assessment via Poisson approximation. The mixture models are also better models for representing the distributions of minutiae compared to the uniform or "corrected" uniform models.

7.6 Summary

To validate and compare models on fingerprint individuality, the models were implemented on three different databases, namely NIST 2000 SD 4, FVC 2002 DB1 and DB2. For all the databases, goodness-of-fit tests on the mixture models showed better performance than the uniform model. For computational efficiency and without jeopardizing the inherent interclass variability, the hyper-mixture model was used on the entire database/population of fingerprints. The compound stochastic model is a further development, accounting in-

traclass variability of fingerprints. In order to compare between the compound stochastic model and the uniform model, synthetic minutiae sets were generated based on each of the models. Visually comparing the matching distributions of the synthetic minutiae sets with empirical distribution (Figure 7.13) shows that the results from the compound stochastic model is more similar than those from the uniform model to the empirical distribution, favoring the compound stochastic model. Finally, the PRCs from the hyper-mixture model, those from the compound stochastic model (based on the synthetic minutiae sets), those from the uniform model, and those from the corrected uniform model were compared with the empirical PRCs. The results show that the PRCs from the hyper-mixture model and the compound stochastic model are orders of magnitude higher than those from the uniform model and the corrected uniform model. Since the proposed models better represent minutiae distributions, we believe that the PRCs reported in this thesis are better estimates of fingerprint individuality compared to previous works.

CHAPTER 8

Conclusions and Future Directions

The task in fingerprint individuality is to develop statistical measures that characterize the extent of uniqueness of a fingerprint. The measure can be taken as the probability of finding another fingerprint that is sufficiently similar to a given query fingerprint in a target population. A satisfactory estimate of fingerprint individuality will make it possible for forensic experts to determine the admissibility of fingerprints as evidence in courts of law where fingerprint-based evidence is increasingly being challenged. The main issue in the assessment of fingerprint individuality is to satisfactorily model two sources of fingerprint variability, namely, the intraclass variability and interclass variability of fingerprint features. This thesis developed a mixture model for the interclass variability, and a compound stochastic model for minutiae intraclass variability. Publications for this research are [12], [53], and [52].

As models for minutiae interclass variability, the mixture models provide a flexible way to represent a variety of observed minutiae distributions in different fingers. Goodness-of-fit tests showed that the mixture model better represents the characteristics of minutiae features observed in fingerprint images compared to the uniform model. For example, for FVC 2002 DB2 with $W = 10$ and $V = 4$, the rejection rate for the mixture model and the uniform models are $1 - 94/100 = 6\%$ and $1 - 2/100 = 98\%$, respectively, based on

Freeman-Tukey test at the 0.05-level.

Although the proposed mixture model is successful in capturing interclass variability, it does not address intraclass variability. Therefore a compound stochastic model was developed to model intraclass variability. More specifically, the model quantifies three main sources of intraclass variability, namely, the nonlinear deformation, local perturbation, and partial fingerprint. A synthesis technique was then used to validate the compound stochastic model and to estimate fingerprint individuality. Experimental results showed that the impostor matching distributions of synthesized databases based on the compound stochastic model were closer to the corresponding empirical matching distributions compared to the distributions based on the uniform model. This observation indicates the superiority of the compound stochastic model over the uniform model.

To estimate individuality of a target population using the above models of minutiae variability, there are two different approaches, namely, synthesis and analytical approaches. In the synthesis approach, minutiae sets are synthesized by the models and a matcher is applied to the synthesized minutiae sets to obtain impostor matching distribution. Fingerprint individuality is then estimated based on the observed number of matches in the synthetically generated database. On the other hand, in the analytical approach, fingerprint individuality is estimated based on a mathematical formula from the Poisson model. This approach is implemented when minutiae density is calculated based on the hyper-mixture model, which only considers minutiae interclass variability. The Poisson model enables analytical estimation of fingerprint individuality. However, the implementation of the Poisson model becomes infeasible for the compound stochastic model where additional sources of variability are considered. As an alternative method, the synthesis approach is good at incorporating multiple sources of minutiae variability, which produces more realistic estimate. Therefore, the synthesis approach is used for the compound stochastic model, which incorporates both interclass variability and intraclass variability. Although the synthesis approach has many advantages, its matching procedure to obtain the impostor matching

distribution, nevertheless, is time-consuming and doesn't produce reliable estimates of fingerprint individuality, especially for very small values of PRC. By contrast, the analytical method has the advantage of high efficiency, but has the disadvantage of not accounting for intraclass variability.

The PRCs obtained from the proposed models were reported and compared with those of the "corrected" uniform model of Pankanti et al. [35] as well as with empirical results which is matcher dependent. It was found that the estimation based on the proposed approach in this thesis is closer to the empirical results compared with those from the "corrected" uniform model. Also, the PRCs from the proposed models are orders of magnitude larger than those from the corrected uniform model which can be explained by the similar clustering tendencies of minutiae from different fingers. Since the proposed models better represent minutiae distributions, we believe that the PRCs reported in this thesis are better estimates of fingerprint individuality compared to previous works.

There are different ways to improve the model presented in this thesis. First of all, the mixture model can be improved. Instead of using a Gaussian mixture model, a t-mixture model, with heavier tails can be applied. As many distributions for angular data can be used to model minutiae directions, such as wrapped Cauchy distribution, wrapped normal distribution, a detailed study on these distributions is needed to choose an optimal distribution model for minutiae directions. Secondly, a non-homogenous Poisson process model incorporating all aspects of minutiae variability (from sources such as superposition of ghost points, thinning, censoring, and uncertainty in the correspondence function) can be developed to incorporate intraclass variability and to develop analytical models. Another direction is to explicitly model spatial dependence of neighboring minutiae. A model for spatial correlation of minutiae distribution will shed more light on quantifying the tendency of minutiae to cluster spatially.

BIBLIOGRAPHY

[1] S. Pankanti A. K. Jain, L. Hong and R. Bolle. An identity authentication system using fingerprints. *Proc. IEEE*, 85(9):103–110, 1997.

[2] J. Buscaglia B. Budowle and R. C. Perlman. Review of the scientific basis for friction ridge comparisons as a means of identification: Committee findings and recommendations. *Forensic Science Communications,1(2) Online at:http://www.fbi.gov/hq/lab/fsc/backissu/jan2006/research/2006 01 research02.htm*, 2006.

[3] V. Balthazard. De lidentification par les empreintes ditalis. *Comptes Rendus, des Academies des Sciences*, 1862(152), 1911.

[4] Claudio Barbarnelli. Evaluating cluster analysis solutions: An application to the italian neo personality inventory. *European Journal of Personality*, 16:s43–s55, 2002.

[5] A. M. Bazen and S. H. Gerez. Systematic methods for the computation of the directional fields and singular points of fingerprints. *IEEE Trans. PAMI*, 24(7):905–919, 2002.

[6] John Berry and David A. Stoney. The history and development of fingerprinting. In Henry C. Lee and R.E. Gaensslen, editors, *Advances in Fingerprint Technology*, pages 1–40. CRC Press, Florida, 2nd edition, 2001.

[7] R. Cappelli, D. Maio, and D. Maltoni. Modelling plastic distortion in fingerprint images. *Proceedings of the Second International Conference on Advances in Pattern Recognition (ICAPR 2001)*, pages 369–376, 2001.

[8] Y. Chen, S. Dass, and A. K. Jain. Fingerprint quality indices for predicting authentication performance. *Proc. of Audio- and Video-based Biometric Person Authentication (AVBPA)*, pages 160–170, 2005.

[9] Simon Cole. "Is Fingerprint Identification Valid? Rhetorics of Reliability in Fingerprint Proponents Discourse". *Law & Policy*, 28(1):109–135, January, 2006.

[10] Suzanne Collins. Judge throws out fingerprint evidence in murder. *http://wjz.com/topstories/shooting.evidence.judge.2.431460.html*, 2007.

[11] H. Cummins and C. Midlo. *Fingerprints, Palms and Soles: An Introduction to Dermatoglyphics*. Dover Publications, Inc., New York, 1961.

[12] S.C. Dass, Y. Zhu, and A.K. Jain. Statistical models for assessing the individuality of fingerprints. *Fourth IEEE workshop on Automatic Identification Advanced Technologies*, 2:1–7, 2005.

[13] Daubert v. Merrel Dow Pharmaceuticals Inc, 509 U.S. 579, 113 S. Ct. 2786, 125 L.Ed.2d 469 (1993).

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38, 1977.

[15] LLC Elephant Engineering, a division of Syntronics. Latentmaster 2004. *http://www.latentmaster.com/*.

[16] F. Galton. *Finger Prints*. London: McMillan, 1892.

[17] Radim Halir and Jan Flusser. Numerically stable direct least square fitting of ellipses. *Proc. of the 6th International Conference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media*, pages 125–132, 1998.

[18] E. R. Henry. *Classification and Uses of Fingerprints*. London: Routledge, 1900.

[19] Geoffrey W. Hill. Evaluation and inversion of the ratios of modified Bessel functions, $I_1(x)/I_0(x)$ and $I_1(x)/I_{.5}(x)$. *ACM Transactions on Mathematica Software*, 7(2):199–208, 1981.

[20] Lin Hong, Yifei Wan, and Anil K. Jain. Fingerprint image enhancement: Algorithms and performance evaluation. *IEEE Transactions on PAMI*, 20(8):777–789, Aug 1998.

[21] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Inc. Upper Saddle River, NJ, 1988.

[22] Joe H. Ward Jr. Hierarchical grouping to optimize an objective functions. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[23] C. Kingston. Probabilistic analysis of partial fingerprint patterns. *Ph.D. Thesis, University of California, Berkeley*, 1964.

[24] H. Bayo Lawal. Comparison of the X2, Y2, freeman-tukey and williams's improved G2 test statistics in small samples of one-way multinomials. *Biometrika*, 71(2):415–418, 1984.

[25] L. LeCam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[26] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and Anil K. Jain. FVC2002: Fingerprint verification competition. In *Proceedings of the International Conference on Pattern Recognition*, pages 744–747, 2002. Online: http://bias.csr.unibo.it/fvc2002/databases.asp.

[27] Davide Maltoni, Dario Maio, Anil K. Jain, and Salil Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag, 2003.

[28] K. V. Mardia. *Statistics of Directional Data*. Academic Press, 1972.

[29] K. V. Mardia and I. L. Dryden. The complex Watson distribution and shape analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(4):913–926, 1999.

[30] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation - an overview. *Digital Signal Processing*, 10:1–18, 2000.

[31] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

[32] O. Nakamura, K. Goto, and T. Minami. Fingerprint classification by directional distribution patterns. *Systems, Computers, Controls*, 13(5):81–89, 1982.

[33] NIST: 8-bit gray scale images of fingerprint image groups (FIGS). Online: http://www.nist.gov/srd/nistsd4.htm.

[34] L. O'Gorman and J. V. Nickerson. An approach to fingerprint filter design,. *Pattern Recognition*, pages 362–385, 1987.

[35] Sharath Pankanti, Salil Prabhakar, and Anil K. Jain. On the individuality of fingerprints. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(8):1010–1025, 2002.

[36] FBI National Press. Statement on Brandon Mayfield case. *FBI National Press:http://www.fbi.gov/pressrel/pressrel04/mayfield052404.htm*, 2004.

[37] A. R. Rao and R. C. Jain. Computerized flow field analysis: Oriented texture fields. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 14(7):693–709, 1992.

[38] N. Ratha and R. Bolle. Fingerprint image quality estimation. *BM Computer Science Research Report RC 21622*, 1999.

[39] Ralph Riddiough and John McColl. *Maths in Action - Advanced Higher Statistics 1: Advanced Higher Statistics 1*. Nelson, 2001.

[40] A. Ross, S. Dass, and A. K. Jain. A deformable model for fingerprint matching. *Pattern Recognition*, 38(1):95–103, 2005.

[41] T. Roxburgh. Work on evidential value of fingerprints. *Sankhya: Indian Journal of Statistics*, 1(50):189–214, 1933.

[42] S. Pankanti S. Prabhakar and A. K. Jain. Biometric recognition: Security and privacy concerns. *IEEE Security and Privacy Magazine*, 1(2):33–42, 2003.

[43] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[44] S. C. Scolve. The occurence of fingerprint characteristics as a two dimensional process. *Journal of the American Statistical Association*, 74(367):588–595, 1979.

[45] George W. Snedecor and William G. Cochran. *Statistical Methods, Eighth Edition.* Iowa UNiversity Press, 1989.

[46] V. S. Srinivasan and N. N. Murthy. Detection of singular points in fingerprint images. *Pattern Recognition*, 25(2):139–153, 1992.

[47] D. A. Stoney and J. I. Thornton. A critical analysis of quantitative fingerprint individuality models. *Journal of Forensic Sciences*, 31(4):1187–1216, 1986.

[48] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[49] U. S. v. Byron Mitchell. Criminal Action No. 96-407, U. S. District Court for the Eastern District of Pennsylvania, 1999.

[50] B. Wentworth and H. H. Wilder. *Personal Identification,*. R. B. Badger, Boston,, 1918.

[51] C. L. Wilson, G. T. Candela, and C. I. Watson. Neural network fingerprint classifiaction. *J. Artificial Neural Networks*, 2:203–228, 1994.

[52] Y. Zhu, S. C. Dass, and A.K. Jain. Statistical models for assessing the individuality of fingerprints. *IEEE Trans. on Information Forensics and Security, 2007*, 2:391–401.

[53] Y. Zhu, S.C. Dass, and A.K. Jain. Statistical models for fingerprint individuality. *Proceedings of the International Conference on Pattern Recognition(ICPR)*, 3:532–535, 2006.

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03062 4617