# Statistical Models for Assessing the Individuality of Fingerprints

Yongfang Zhu, *Student Member, IEEE*, Sarat C. Dass, and Anil K. Jain, *Fellow, IEEE*

***Abstract*—Following the Daubert ruling in 1993, forensic evidence based on fingerprints was first challenged in the 1999 case of the U.S. versus Byron C. Mitchell 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 and matching error rates are unknown. In order to assess the error rates, we require quantifying the variability of fingerprint features, namely, minutiae in the target population. A family of finite mixture models has been developed in this paper to represent the distribution of minutiae in fingerprint images, including minutiae clustering tendencies and dependencies in different regions of the fingerprint image domain. A mathematical model that computes the probability of a random correspondence (PRC) is derived based on the mixture models. A PRC of $2.25 \times 10^{-6}$ corresponding to 12 minutiae matches was computed for the NIST4 Special Database, when the numbers of query and template minutiae both equal 46. This is also the estimate of the PRC for a target population with a similar composition as that of NIST4.**

***Index Terms*—Fingerprint identification, image processing, pattern classification, pattern clustering methods, pattern matching, statistics.**

## I. Introduction

EXPERT testimony based on fingerprint evidence is delivered in a courtroom by comparing salient features of a latent print lifted from a crime scene with those taken from the defendant. A reasonably high degree of match between the salient features leads the experts to testify irrefutably that the source of the latent print and the defendant are one and the same person. For decades, the testimony of forensic fingerprint experts was almost never excluded from these cases, and on cross-examination, the foundations and basis of this testimony were rarely questioned. Central to establishing an identity based on fingerprint evidence is the assumption of discernible uniqueness; salient features of fingerprints of different individuals are observably different and, therefore, when two prints share many common features, the experts conclude that the sources of the two different prints are one and the same person. The assumption of discernible uniqueness, although lacking sound theoretical and empirical foundations [20], allows forensic experts to offer an unquestionable proof toward the defendant's guilt. To make matters worse, forensic experts are never questioned on the uncertainty associated with their testimonials (that is, how frequently would an observable match between a pair of prints lead to errors in the identification of individuals). Thus, discernible uniqueness precludes the opportunity to establish error rates which should be estimated from collecting population samples, analyzing the inherent feature variability, and reporting the corresponding probability of two different persons sharing a set of common features (known as the probability of random correspondence).

A significant event that questioned this trend occurred in 1993 in the case of Daubert versus Merrell Dow Pharmaceuticals [7] 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 1) the particular technique or methodology has been subject to statistical hypothesis testing, 2) its error rates have been established, 3) standards controlling the technique's operation exist and have been maintained, 4) it has been peer reviewed, and 5) it has a general widespread acceptance [18]. Forensic evidence based on fingerprints was first challenged in the 1999 case of the U.S. versus Byron C. Mitchell [23] under the Daubert ruling, stating that the fundamental premise for asserting the uniqueness of fingerprints had not been objectively tested and its potential matching error rates were unknown. After the U.S. versus Byron C. Mitchell case, fingerprint-based identification has been challenged in more than 20 court cases in the U.S., see, for example, the U.S. versus Llera Plaza [25], [26] in 2002 and U.S. versus Crisp [24] in 2003; and [5] for additional court cases.

The main issue with the admissibility of fingerprint evidence stems from the realization that the individualization of fingerprints has not been subjected to the principles of scientific validation. The uncertainty involved in assessing 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?" As mentioned before, a satisfactory answer to this question requires 1) collecting fingerprint samples from a target population, 2) analyzing the variability of the features from the different fingerprints collected, and 3) defining a notion of similarity between fingerprints and reporting the corresponding probability of two

Manuscript received July 12, 2006; revised March 18, 2007. This work was supported by the National Science Foundation ITR Grant 0312646 and ARO Grant W911NF-06-1-0418. The associate editor coordinating the review of this manuscript and approving if for publication was Dr. Tieniu Tan.

Y. Zhu and S. C. Dass are with the Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824 USA.

A. K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: zhuyongf@msu.edu; sdass@msu.edu; jain@msu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2007.903846

1556-6013/$25.00 © 2007 IEEE

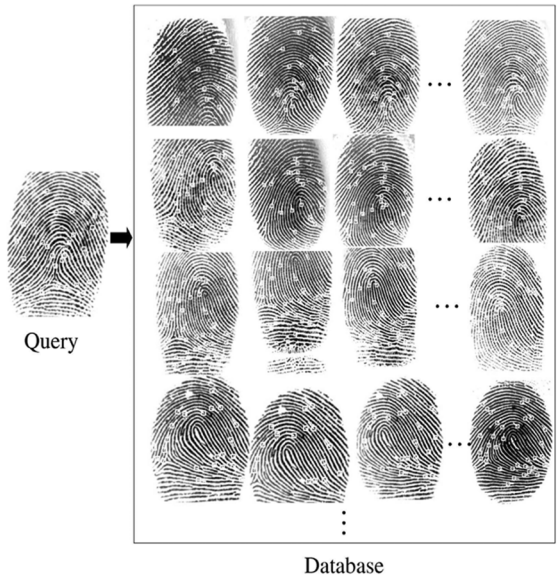



Fig. 1. Intraclass variability in a fingerprint database. Rows correspond to different fingers whereas columns correspond to multiple impressions of the same finger. White boxes in the fingerprint image correspond to the location of fingerprint minutiae.

different individuals sharing a set of common fingerprint features. We address issues 2) and 3) in this paper assuming that a sample of prints is available from a target population and a notion of similarity is given; see also Fig. 1. We do not address the issues and challenges involved in sampling from a target population. Instead, we assume that a database of prints is available and demonstrate how the methodology described in this paper can be used to obtain estimates of fingerprint individuality. If the available database is representative of the target population, then the estimates of fingerprint individuality obtained based on the methodology presented here would generalize the target population. An analysis of variability of fingerprint features requires the development of appropriate statistical models on the space of fingerprint features that are able to represent all aspects of variability observed in these features. Based on these models, the probability of a random correspondence (PRC) (alternatively, the probability that the observed match between features in a pair of prints is purely due to "chance") will be determined.

There have been a few previous studies that addressed the problem of fingerprint individuality using statistical models on fingerprint features. All of these studies utilized minutiae features in fingerprints (both location and direction information) to assess individuality. However, the assumptions made in these studies do not satisfactorily represent the observed variations of the features in actual fingerprint databases. For example, it is known that fingerprint minutiae tend to form clusters [21], [22], but Pankanti *et al.* [18] assumed a uniform distribution on minutiae locations and directions which was then corrected to match empirical results from the databases used in their study. Another assumption made by Pankanti *et al.* is that the minutiae location is distributed independently of the minutiae direction. But, minutiae in different regions of the fingerprint are observed to be associated with different region-specific minutiae directions. Moreover, minutiae points that are spatially close tend to have similar directions with each other. These observations on the distribution of fingerprint minutiae need to be accounted for in eliciting reliable statistical models.

The problem of establishing individuality estimates based on fingerprints is in contrast to DNA typing where the probability of a random correspondence has been studied extensively and quantified (see, for example, [10]). The DNA typing problem (inherently 1-D) is in some sense is simpler to analyze compared to the fingerprint individuality problem (inherently 2-D); also, the act of acquiring fingerprint impressions as well as the condition of the physical finger itself (i.e., cuts and bruises, and distortions) introduces many sources of noise. This paper proposes to determine reliable estimates of the probability of a random correspondence between two fingerprints via appropriate statistical models in a spirit similar to that of DNA typing.

To address the issue of individuality, candidate models have to meet two important requirements: 1) flexibility, that is, the model can represent the observed distributions of the minutiae features in fingerprint images over different databases and 2) associated measures of fingerprint individuality can be easily obtained from these models. In practice, a forensic expert uses many fingerprint features (minutiae location and direction, fingerprint class, inter-ridge distance, etc.) to make the match, but here we only use a subset of these features, namely, the minutiae locations and directions, to keep the modeling problem tractable. We introduce a family of finite mixture models to represent the observed distribution of minutiae locations and directions in fingerprint images. The reliability of the models is assessed using a criteria based on the degree to which the models are able to capture the observed variability in the minutiae locations and directions. We then derive a mathematical model for computing the PRCs based on the elicited mixture models.

The rest of this paper is organized as follows: Section II describes the finite mixture models proposed for the minutiae features (both location and direction). We also develop tests to demonstrate the appropriateness of the mixtures as distributional models for fingerprint minutiae compared to the uniform distribution. Section III develops a new mathematical model for computing the PRC, whereas Section IV describes the experimental results based on the NIST Special Database 4 [17], and FVC2002 [13] database.

## II. Statistical Models on Minutiae Location and Direction

A minutiae is the location of a ridge anomaly in a fingerprint image [14]. Forensic experts and most automatic fingerprint matching systems use minutiae for identification since these features have been shown to be stable and can be reliably extracted from prints. There are many types of ridge anomalies that occur in fingerprint images—examples of these include ridge endings, bifurcations, islands, dots, enclosures, bridges, double bifurcations, trifurcations, and others. However, in this paper, we only consider the two dominant types of minutiae, namely, endings and bifurcations. The main reasons for this are that the occurrence of the other ridge anomalies is relatively rare, and it is easy to consistently detect minutiae endings and bifurcations compared to other minutiae types. Each minutiae is characterized in

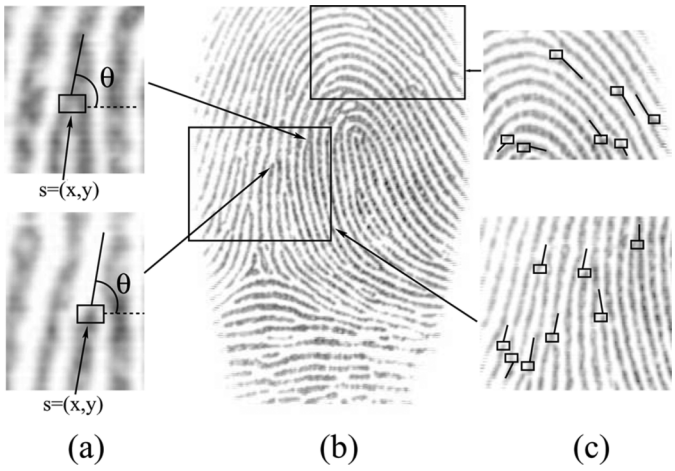



Fig. 2. Minutiae features consisting of the location $s$ and direction $\theta$ for a typical fingerprint image (b): The top (respectively, bottom) panel in (a) shows $s$ and $\theta$ for a ridge bifurcation (respectively, ending). The top (respectively, bottom) panel in (c) shows two subregions in which orientations of minutiae points that are spatially close tend to be very similar.

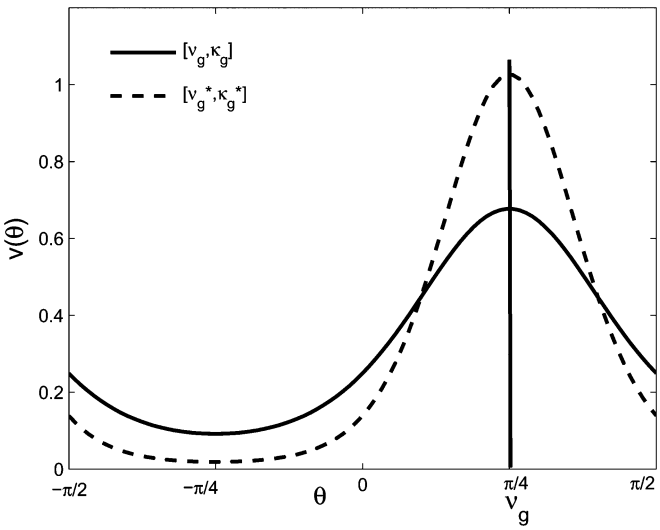



Fig. 3. Probability distribution plots of the Von–Mises distribution with center $\nu_g = \pi/4$, and with two different precisions $\kappa_g$ and $\kappa_g^*$, with $\kappa_g < \kappa_g^*$. The values of $v(\theta)$ at $(\pi/2)$ and $-(\pi/2)$ are equal to each other due to the cyclical nature of the cosine function.

terms of two components: 1) its location (i.e., the spatial coordinates of its position) and 2) its direction (i.e., the angle subtended by the minutiae measured from the horizontal axis). We also do not distinguish between minutiae bifurcation and ending since it is often not easy to distinguish between them by automatic systems. Subsequently, the term "minutiae features" will be used to refer to the location and direction of a minutiae in a fingerprint impression. See Fig. 2 for an example of minutiae features for a fingerprint impression from the FVC2002 DB1 [13] database.

Let $X$ denote a generic random minutiae location and $D$ denote its corresponding direction. Let $S \subseteq \mathbf{R^2}$ denote the subset of the plane representing the fingerprint domain. Then, the set of all possible configurations for $X$ is the $(x, y) \equiv s$ coordinate points in $S$. The minutiae direction $D$ takes values in $[0, 2\pi)$. Denoting the total number of minutiae in a fingerprint image by $k$, we will develop a joint distribution model for the $k$ pairs of minutiae features $(X, D)$: $\{(X_j, D_j),\ j = 1, 2, \ldots, k\}$ that accounts for 1) clustering tendencies (nonuniformity) of minutiae, and 2) dependence between minutiae location and direction ($X_j$ and $D_j$) in different regions of $S$.

The proposed joint distribution model is based on a mixture consisting of $G$ components or clusters. Let $c_j$ be the cluster label of the $j$th minutiae location and direction $(X_j, D_j)$, $c_j \in \{1, 2, \ldots, G\}$, $j = 1, 2, \ldots, k$. The labels $c_j$ are independently distributed according to a single multinomial with $G$ classes and class probabilities $\tau_1, \tau_2, \ldots, \tau_G$ such that $\tau_j \geq 0$ and $\sum_{j=1}^{G} \tau_j = 1$. Given label $c_j = g$, the minutiae location $X_j$ is distributed according to the density

$$f_g^X(s \mid \mu_g, \Sigma_g) = \phi_2(s \mid \mu_g, \Sigma_g) \tag{1}$$

where $\phi_2$ is the bivariate Gaussian density with mean $\mu_g$ and covariance matrix $\Sigma_g$. Equation (1) states that the minutiae locations arising from the $g$th cluster follow a 2-D Gaussian with mean $\mu_g$ and covariance matrix $\Sigma_g$. The Von–Mises distribution [15] is a typical distribution used to model angular random variables, such as minutiae directions in our case. So we assume the distribution of $j$th minutiae direction, $D_j$ belonging to the $g$th cluster follows the density

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

where $I\{A\}$ is the indicator function of the set $A$ (that is, $I\{A\} = 1$ if $A$ is true, and 0, otherwise), $p_g$ is a real number between 0 and 1, and $v(\theta)$ is the Von–Mises distribution given by

$$v(\theta) \equiv v(\theta \mid \nu_g, \kappa_g) = \frac{2}{I_0(\kappa_g)} \exp\{\kappa_g \cos 2(\theta - \nu_g)\} \tag{3}$$

with $I_0(\kappa_g)$ defined as

$$I_0(\kappa_g) = \int_0^{2\pi} \exp\{\kappa_g \cos(\theta - \nu_g)\} d\theta. \tag{4}$$

In (3), $\nu_g$ and $\kappa_g$ represent the mean angle and the precision (inverse of the variance) of the Von–Mises distribution, respectively. Fig. 3 plots two density functions associated with Von–Mises distributions with common means $\nu_g$ but with two different precisions $\kappa_g < \kappa_g^*$. This figure shows that $\nu_g$ represents the "center" (or modal value) while $\kappa_g$ controls the degree of spread around the center (thus, the density with precision $\kappa_g^*$ has a higher concentration around $\nu_g$). The density $f_g^D$ in (2) can be interpreted in the following way: The ridge flow orientation $\mathcal{O}$ is assumed to follow the Von–Mises distribution (3) with mean $\nu_g$ and precision $\kappa_g$. Subsequently, minutiae arising from the $g$th component have directions that are either $\mathcal{O}$ or $\mathcal{O} + \pi$ with probabilities $p_g$ and $1 - p_g$, respectively.

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

$$f(s, \theta \mid \Theta_G) = \sum_{g=1}^{G} \tau_g f_g^X(s \mid \mu_g, \Sigma_g) \cdot f_g^D(\theta \mid \nu_g, \kappa_g, p_g), \tag{5}$$

where $f_g^X(\cdot)$ and $f_g^D(\cdot)$ are defined as in (1) and (2), respectively. In (5), $\Theta_G$ denotes all of the unknown parameters in the mixture model which includes the total number of mixture components $G$, the mixture probabilities $\tau_g$ $g = 1, 2, \ldots, G$, the component means and covariance matrices of $f_g^X$'s given by $\boldsymbol{\mu}_G \equiv \{\mu_1, \mu_2, \ldots, \mu_G\}$, and $\boldsymbol{\Sigma}_G \equiv \{\Sigma_1, \Sigma_2, \ldots, \Sigma_G\}$; 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 in (5) allows for 1) different clustering tendencies in the minutiae locations and directions via $G$ different clusters, and 2) incorporates dependence between the minutiae location and direction since if $X_j$ is known to come from the $g$th component, then it follows that the direction $D_j$ also comes from the same mixture component.

The mixture density given in (5) is defined on the entire plane $\mathbf{R}^2$ and is not restricted to the fingerprint domain $S$. We correct this by defining the mixture model on the fingerprint area $A \subset S$ as

$$f_A(s, \theta \mid \Theta_G) = \frac{f(s, \theta \mid \Theta_G)}{\int_{s \in A} \int_{\theta=0}^{2\pi} f(s, \theta \mid \Theta_G) d\theta ds}. \tag{6}$$

If most of the fingerprint area $A$ encompasses the entire rectangular sensing area, $S$ (i.e., $A \approx S$)

$$f_A(s, \theta \mid \Theta_G) \approx f(s, \theta \mid \Theta_G) \tag{7}$$

since then the denominator in (6)

$$\int_{s \in A} \int_{\theta=0}^{2\pi} f(s, \theta \mid \Theta_G) d\theta ds \approx 1. \tag{8}$$

To estimate the unknown parameters in the model, we develop an algorithm based on hierarchical agglomeration and the electromagnetic (EM) algorithm [8], [16] for unrestricted multivariate mixture models. The optimal number of components $G^*$ is selected using the Bayes information criteria (BIC). The BIC has been widely used in various model selection problems and has the property that it selects a model that is most parsimonious (the one having the least number of model parameters). Details of the EM algorithm and the BIC used here are given in Appendix A of the technical report [27].

Fig. 4 illustrates the fit of the mixture model to two fingerprint images from the NIST 4 database. 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 minutiae feature in (a) and (b). The cluster label of $(X_j, D_j)$ is estimated according to [27, (46) in Appendix A] after the EM algorithm has converged. Panels (e) and (f) plot the minutiae features in the 3-D $(X, D)$ 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. In panels (c-f), minutiae from the same cluster are labeled with the same color, shape, and number.

Another way to show the effectiveness of the fit of the models to the observed data is to simulate a realization from the fitted models. Fig. 5(a) and (b) shows two fingerprints whose minutiae features were fitted with the mixture distribution in (6). Fig. 5(e)–(f) shows a simulated realization of minutiae when each $X$ and $D$ is assumed to be uniformly distributed independently of each other. Note that there is good agreement, in the distributional sense, between the observed [Fig. 5(a) and (b)] and simulated minutiae locations and directions from the proposed models [Fig. 5(c) and (d)] but no such agreement exists for the uniform model [Fig. 5(e) and (f)].

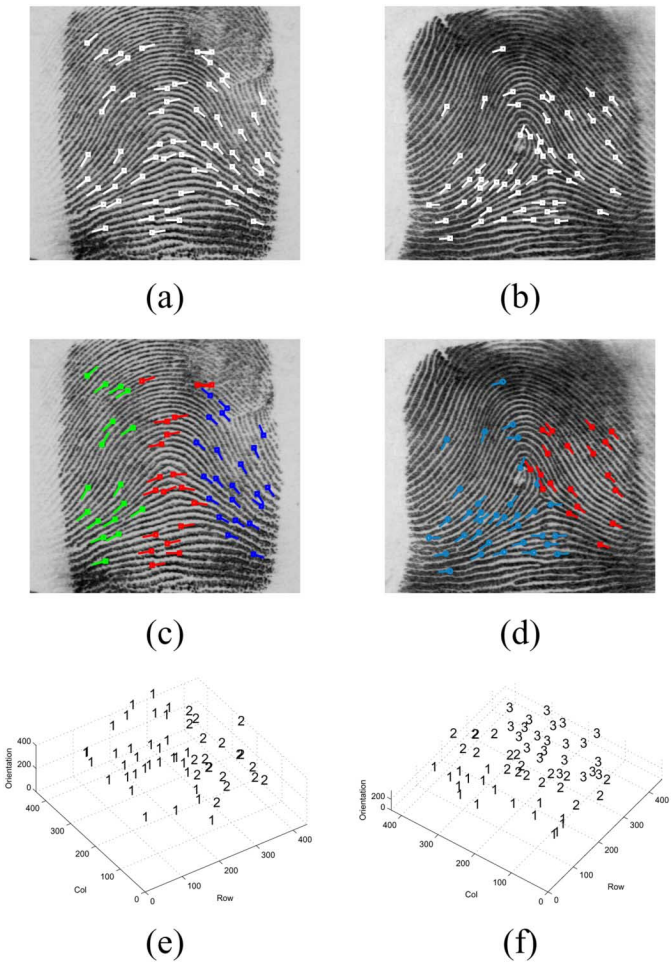



Fig. 4. Assessing the fit of the mixture models to minutiae location and direction: Observed minutiae locations (white boxes) and directions (white lines) are shown in panels (a) and (b) for two different fingerprints from the NIST Special Database 4. Panels (c) and (d), respectively, show the clusters in different colors (light blue, red, and deep blue). The clusters (and associated cluster labels) in 3-D space are shown in panels (e) and (f) with $x$, $y$, $z$ as the row, column, and the orientation of the minutiae in (a) and (b).

## III. Probability of Random Correspondence (PRC)

The probability of a random correspondence (PRC) is the chance that an arbitrary impostor fingerprint from a target population will share a sufficiently large number of minutiae with the query. Small (respectively, large) values of the PRC imply that it is unlikely (respectively, likely) that minutiae in a fingerprint of an individual other than the query source will match those of the query. Let $Q$ (respectively, $T$) denote the query (respectively, file or template) fingerprint image from the individual $I_Q$ (respectively, $I_T \neq I_Q$). To compute the PRC, we first define a minutiae match between $Q$ and $T$. A pair of minutiae features in $Q$ and $T$, $(X^Q, D^Q)$ and $(X^T, D^T)$ respectively, is said to match if for fixed positive numbers $r_0$ and $d_0$

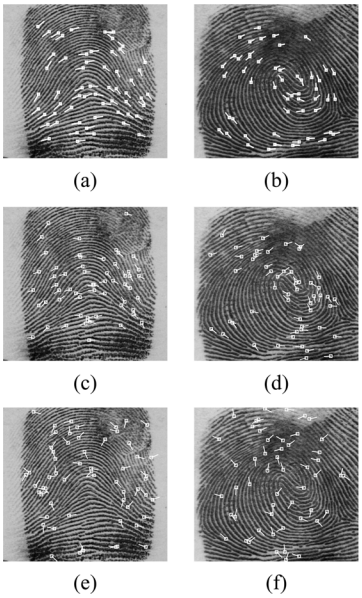



Fig. 5. All $(X, D)$ realizations from the proposed model [(c) and (d)], and from the uniform distribution [(e) and (f)] for two different images [(a) and (b)]. The true minutiae locations and directions are marked in (a) and (b).

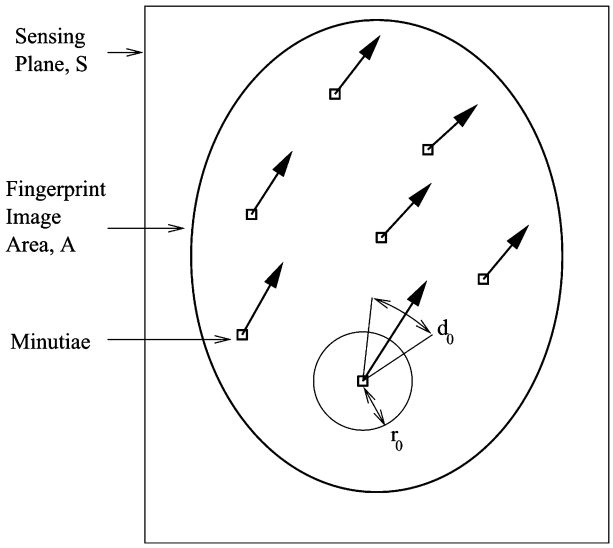



Fig. 6. Identifying the matching region for a query minutiae.

$$|X^Q - X^T|_s \leq r_0 \text{ and } |D^Q - D^T|_a \leq d_0 \tag{9}$$

where

$$|X^Q - X^T|_s \equiv \sqrt{(x^Q - x^T)^2 + (y^Q - y^T)} \tag{10}$$

is the (spatial) Euclidean distance between the minutiae locations $X^Q = (x^Q, y^Q)$ and $X^T = (x^T, y^T)$, and

$$|D^Q - D^T|_a \equiv \min(|D^Q - D^T|, 2\pi - |D^Q - D^T|) \tag{11}$$

is the angular distance between the minutiae directions $D^Q$ and $D^T$; see Fig. 6. The definition of a match in (9) is the same as that used by Pankanti *et al.* [18] and depends on two parameters $r_0$ and $d_0$. Large (respectively, small) values of the pair $(r_0, d_0)$ will lead to spurious (respectively, 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 presented in Section III-C and postponed for now.

Suppose the following quantities are available:

1) $m$: number of minutiae in query fingerprint $Q$ with minutiae denoted by $\left(X_i^Q, D_i^Q\right), i = 1, 2, \ldots, m$;
2) $n$: number of minutiae in file fingerprint $T$ with minutiae denoted by $\left(X_i^T, D_i^T\right), i = 1, 2, \ldots, n$;
3) $w$: number of matching minutiae between $Q$ and $T$.

We assume that the query and file minutiae are distributed independently according to the mixture densities

$$f_Q(X^Q, D^Q) = f\left(X^Q, D^Q \mid \Theta_G^Q\right) \tag{12}$$

and

$$f_T(X^T, D^T) = f\left(X^T, D^T \mid \Theta_G^T\right) \tag{13}$$

respectively. Then, the PRC is the probability of obtaining exactly $w$ matches between $Q$ and $T$ when $I_Q \neq I_T$. In order to compute the probability of obtaining $w$ matches, we impose the condition that the minutiae sets of $Q$ and $T$ cannot be too close to each other; this is a reasonable assumption to make since minutiae can only occur on ridges and, therefore, should be at least one inter-ridge distance away from one another (see also Pankanti *et al.* [18] for a similar condition). In [27, App. B], we show that the probability of obtaining exactly $w$ matches, given that there are $m$ and $n$ minutiae in $Q$ and $T$, respectively, is given by the expression

$$p^*(w; Q, T) = \frac{e^{-\lambda(Q,T)}\lambda(Q,T)^w}{w!} \tag{14}$$

for large $m$ and $n$; (14) corresponds to the Poisson probability mass function with mean $\lambda(Q, T)$ given by

$$\lambda(Q, T) = mnp(Q, T) \tag{15}$$

where

$$p(Q, T) = P\left(|X^Q - X^T|_s \leq r_0 \text{ and } |D^Q - D^T|_a \leq d_0\right) \tag{16}$$

denotes the probability of a match when $(X^Q, D^Q)$ and $(X^T, D^T)$ are random minutiae from (12) and (13), respectively. The mean parameter $\lambda(Q, T)$ can be interpreted as the expected number of matches from the total number of $mn$ possible pairings between $m$ minutiae in $Q$ and $n$ minutiae points in $T$ with the probability of each match being $p(Q, T)$. As a rule of thumb, taking $m, n \geq 40$ results in the Poisson being a reasonably good approximation of the exact matching probability (see [27] for details). The Poisson distribution in (14) is obtained using arguments similar to when a binomial distribution with a large number of trials and small probability of "success" can be approximated by a Poisson distribution, provided the expected number of "successes" is moderate. For this reason, the Poisson approximation is also called the law of rare events. In our case, if we define "success" to be a minutiae

match, then 1) the number of trials $mn$ is large and 2) the probability of a success $p(Q,T)$ is small, and 3) the number of impostor matches between $Q$ and $T$ is moderate (not exceeding 10 in the databases we worked with), thus, justifying the use of the Poisson law.

The previous discussion is general and holds true for any distribution for the query and file minutiae. In particular, when the distributions on the minutiae (both location and direction) are chosen to be uniform, we obtain the following expression for $\lambda(Q,T)$:

$$\lambda_U(Q,T) = mnp_L p_D \tag{17}$$

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

For a fingerprint database consisting of $F$, different fingers with a single impression per finger, we wish to find the most representative value for the probability of a random correspondence—PRC–for this database. There are a total of $F(F-1)/2$ pairs of impostor fingerprint images $(Q,T)$ from the entire database. The average PRC corresponding to $w$ minutiae matches is given by

$$\overline{\mathrm{PRC}} = \frac{2}{F(F-1)} \sum_{(Q,T)\mathrm{impostorpairs}} p^*(w;Q,T) \tag{18}$$

where $p^*(w;Q,T)$ is as defined in (14); note that $p^*(w;Q,T)$ is symmetric in $Q$ and $T$ and, thus, it is sufficient to consider only the $F(F-1)/2$ distinct impostor pairs instead of the total $F(F-1)$. Each probability $p^*(w;Q,T)$ is a very small number, such as $10^{-6}$ or $10^{-7}$. Thus, the average PRC in (18) 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 $\alpha$ 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. The $\alpha$-trimmed mean is given by

$$\overline{\mathrm{PRC}}_\alpha = \frac{2}{F(F-1)(1-\alpha)} \sum_{(Q,T)\mathrm{impostor}} p^*_\alpha(w;Q,T) \tag{19}$$

where as shown in (20), at the bottom of the page.

### A. Incorporating Multiple Impressions Per Finger

We would like the model to be as accurate as possible. Since each fingerprint often only gives a partial impression of a finger, combining multiple impressions of a finger results in better model fits. To utilize multiple impressions of a finger (such as from databases in the fingerprint verification competitions (FVCs) [12], [13], we combine minutiae from different impressions into a single "master" on which the mixture model is fit. The minutiae consolidation procedure we follow is described in detail in [27] and [28]. An illustration of the consolidation procedure is shown in Fig. 7 where multiple impressions of the same finger (a) are aligned to the reference image (b) to obtain the master fingerprint (c). The process of minutiae consolidation has two advantages: 1) a more reliable fit of the mixture model is obtained and 2) the assumption of large $m$ and $n$ required for computing the individuality estimates is satisfied. PRCs for $w$ matches are then obtained using (14) for the $F(F-1)/2$ impostor master pairs. The consolidation process involves averaging the location and direction of the same minutiae obtained from the multiple impressions. This helps smooth out any nonlinear distortion effects that can affect the estimate of fingerprint individuality. In this paper, we do not model the variability in the partial prints corresponding to each finger as was done in [28].

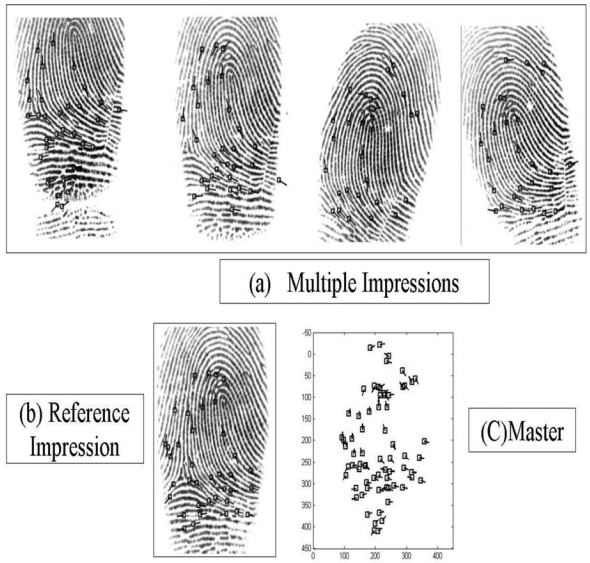



Fig. 7. Master fingerprint construction. (a) Four different impressions of a finger. (b) Reference impression. (c) Master.

### B. Identifying Clusters of Fitted Mixture Models

In order to compute the probability of random correspondence based on the mixture models, our methodology involves fitting a separate mixture model to each fingerprint impression/master from a target population. An important difference between the proposed methodology and previous work is that we fit mixture models to each finger/master, whereas previous studies assumed a common distribution for all fingers/impressions. Assume a common minutiae distribution for all

$$p^*_\alpha(w;Q,T) = \begin{cases} p^*(w;Q,T), & \text{if } p^*\left(w;\frac{\alpha}{2}\right) \le p^*(w;Q,T) \le p^*\left(w;\frac{1-\alpha}{2}\right), \text{and} \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

fingerprint impressions has a serious drawback, namely, that the true distribution of minutiae may not be modeled well. For example, it is well known that the five major fingerprint classes in the Henry system of classification (i.e., right loop, left loop, whorl, arch, and tented arch) have different class-specific minutiae distributions. Thus, using one common minutiae distribution may smooth out important clusters in the different fingerprint classes. Moreover, PRCs depend heavily on the composition of each target population. Consider the following example: The proportion of right loop, left loop, whorl, arch, and tented arch classes of fingerprints is 31.7%, 33.8%, 27.9%, 3.7%, and 2.9%, respectively, in a sample of the British population as reported in [6]. Thus, PRCs computed for fingerprints from this population will be largely influenced by the mixture models fitted to the right loop, left loop, and whorl classes compared to arch and tented arch. More important is the fact that the PRCs will change if the class proportions change (for example, if the target population has an equal number of fingerprints in each class, or with class proportions different from the ones given before). By fitting separate mixture models to each finger, we ensure that the composition of a target population is correctly represented.

To formally obtain the composition of a target population, we adopt an agglomerative hierarchical clustering procedure [9] on the space of all fitted mixture models to cluster $F$ mixture densities. The dissimilarity measure between the estimated mixture densities $f$ and $g$ is taken to be the Hellinger distance [11]

$$H(f,g) = \int_{x\in S}\int_{\theta\in[0,2\pi)} (\sqrt{f(x,\theta)} - \sqrt{g(x,\theta)})^2 dx d\theta. \quad (21)$$

The Hellinger distance $H$ is a number bounded between 0 and 2, with $H = 0$ (respectively, $H = 2$) if and only if $f = g$ (respectively, $f$ and $g$ have disjoint support). For a database with $F$ fingers, we obtain a total of $F(F-1)/2$ Hellinger distances corresponding to the $F(F-1)/2$ mixture density pairs. The resulting dendrogram can be cut to form $N$ clusters of mixture densities $C_1, C_2, \ldots, C_N$ say, based on a threshold $\mathcal{T}$. Note that $N = 1$ when $\mathcal{T} = 2$ and as $\mathcal{T}$ decreases to 0, $N$ increases to $F(F-1)/2$. When the number of clusters is $N$, we define the within cluster dissimilarity as

$$W_N = \sum_{i=1}^{N} \frac{1}{2|C_i|} \mathcal{D}(C_i) \quad (22)$$

where

$$\mathcal{D}(C_i) = \sum_{f,g\in C_i} H(f,g) \quad (23)$$

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, $W_N$ decreases to 0. To choose the optimal number of clusters, we use the "elbow criteria": Let $G_N = |W_N - W_{N-1}|$ denote the absolute difference between the within cluster dissimilarities $W_{N-1}$ and $W_N$. We select $N^*$ as the number of clusters if the values of $G_N$ for $N > N^*$ are insignificant (close to 0) compared to the value of $G_{N^*}$. The criteria is named after the "elbow" that is created at $N = N^*$ in the plot of $W_N$ versus $N$.

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

$$\bar{f}_i(x,\theta) = \frac{1}{|C_i|} \sum_{f\in C_i} f(x,\theta) \quad (24)$$

$i = 1, 2, \ldots, N^*$. The mean parameter $\lambda(Q,T)$ in (15) depends on $Q$ and $T$ via the mean mixture densities of the clusters from which $Q$ and $T$ are taken. If $Q$ and $T$, respectively, belong to clusters $C_i$ and $C_j$, we have $\lambda(Q,T) \equiv \lambda(C_i, C_j)$ with the mean mixture densities of $C_i$ and $C_j$ used in place of the original mixture densities in (16). Let $p^*(w; C_i, C_j)$ denote the Poisson probability

$$p^*(w; C_i, C_j) = e^{-\lambda(C_i,C_j)} \frac{\lambda(C_i,C_j)^w}{w!}. \quad (25)$$

To obtain the $100(1-\alpha)\%$ trimmed mean, we denote the lower and upper $100\alpha/2$th percentiles of $\{p^*(w; C_i, C_j), 1 \le i, j \le N^*\}$ by $p^*_C(w; \alpha/2)$ and $p^*_C(w; 1-\alpha/2)$. Also, define the set of all trimmed $p^*(w; C_i, C_j)$ probabilities as $\mathcal{T} \equiv \{(i,j) : p^*_C(w;\alpha/2) \le p^*(w; C_i, C_j) \le p^*_C(w; 1-\alpha/2)\}$. Then, the $100(1-\alpha)\%$ trimmed mean PRC is

$$\overline{\mathrm{PRC}}_\alpha = \frac{\sum_{(i,j)\in\mathcal{T}} |C_i||C_j| p^*(w; C_i, C_j)}{\sum_{(i,j)\in\mathcal{T}} |C_i||C_j|}. \quad (26)$$

### C. Estimation of $(r_0, d_0)$

Parameters $(r_0, d_0)$ determine the matching region for a query minutiae. In the ideal situation, a genuine pair of matching minutiae in the query and file will correspond exactly leading to the choice of $(r_0, d_0)$ as (0,0). However, factors, such as skin elasticity and nonuniform 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 nonmatches, nonzero values of $r_0$ and $d_0$ need to be specified for matching pairs of genuine minutiae. We adopt the procedure taken by Pankanti *et al.* [18] for selecting a reasonable value for the pair $(r_0, d_0)$ such that only a small prespecified proportion of genuine matches will be rejected. The value of $r_0$ is determined based on the distribution of Euclidean distances between every pair of matched minutiae in the genuine case. We align pairs of genuine fingerprints and find the corresponding pairs of minutiae. The value of $r_0$ is selected based on the distribution of the Euclidean distance between the locations of the minutiae pairs. The value of $r_0$ is selected so that only the upper 2.5% of the genuine matching distances (corresponding to large values of $r$) are rejected.

In a similar fashion, we can compute the distribution of the angular distance between the directions for minutiae pairs. The value of $d_0$ is determined to be the 97.5%th percentile of this distribution (again, the upper 2.5% of the genuine matching angular distances will be rejected).

Pankanti *et al.* [18] used a database of 450 mated fingerprint pairs where the true minutiae locations and the correspondences between minutiae in each genuine fingerprint pair were determined by a fingerprint expert. Using this ground truth corre-

spondence, they estimated $r_0$ and $d_0$ to be 15 and 22.5, respectively. These values will be used in our experiments to estimate the probability of random correspondence.

## IV. Experimental Results

Our methodology for assessing the individuality of fingerprints is validated on three target populations, namely, the NIST Special Database 4 [17], the FVC2002 DB1, and the FVC2002 DB2 [13] fingerprint databases. The NIST fingerprint database [17] is publicly available and contains 2 000 8-b grayscale fingerprint image pairs measuring $512 \times 512$ pixels. Due to the relatively large size of the images in the NIST database, we used the first image of each pair for statistical modeling. Minutiae could not be automatically extracted from two images of the NIST database due to poor quality. Thus, the total number of NIST fingerprints used in our experiments was $F = 1{,}998$.

For the FVC2002 database, also available in the public domain, we used two of its subsets DB1 and DB2. The DB1 impressions (images size $= 388 \times 374$) are acquired using the optical sensor "TouchView II" by Identix, while the DB2 impressions (image size $= 296 \times 560$) are acquired using the optical sensor "FX2000" by Biometrika. Each database consists of $F = 100$ different fingers with eight impressions $(L = 8)$ per finger. Due to the small size of the images in the DB1 and DB2 databases, the minutiae consolidation procedure was adopted to obtain a master. The mixture models were subsequently fitted to each master.

The best fitting mixture model [see (5) and (6)] was found for each finger for these three databases. Two types of statistical tests for checking the appropriateness of the mixture model (6) as a distribution on fingerprint minutiae were carried out. The first type of test was to select between two models, either the mixture or the uniform, for the minutiae for each finger based on the likelihood ratio (the mixture and uniform models were fitted to the master whenever the consolidation procedure of Section III-A was adopted for a database). This model selection procedure can decide only between the mixture and the uniform model. However, it may be the case that the true distribution on fingerprint minutiae is neither one of these. Thus, the second type of statistical test carried out was to assess the goodness of fit of the mixture model to the observed distribution of minutiae for each finger. The quality of fit of the mixture distribution was determined via a $p$-value where large $p$-values ($p$-values $> 0.01$) led to the conclusion that the mixture distribution is an adequate model; otherwise, when the $p$-value is smaller than 0.01, the mixture distribution is inadequate. In a similar fashion, we also tested the goodness of fit of the uniform model to the distribution of minutiae for each finger. Based on these statistical tests, we found strong evidence for the appropriateness of the mixture models as a distribution on fingerprint minutiae for all three databases. We refer the reader to the technical report [27] for more details on the tests that were carried out as well as the experimental results.

The distributions of $m$ and $n$ for the three fingerprint databases are shown in Fig. 8(a)–(c), respectively (the distribution of $m$ and the distribution of $n$ are identical and, hence, only one histogram is obtained). The mean $m$ (and $n$) values for the NIST, FVC2002 DB1, and FVC2002 DB2 databases

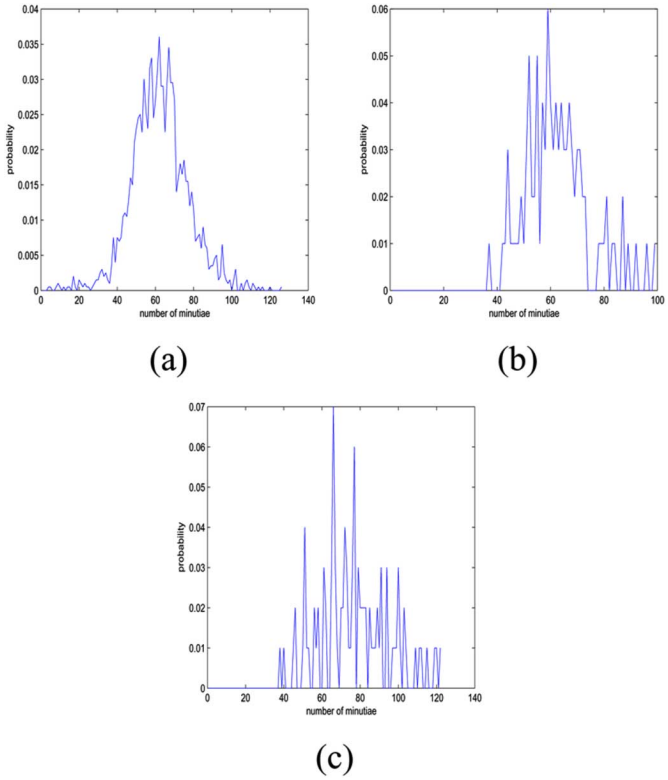



Fig. 8. Empirical distributions of the number of minutiae in the (a) NIST database, (b) master prints constructed from the FVC2002 DB1 database, and (c) master prints constructed from the FVC2002 DB2 database. The average number of minutiae in the three distributions are 62, 63, and 77, respectively.

are approximately 62, 63, and 77, respectively (For the FVC databases, $m$ and $n$ are reported as the mean number of minutiae centers in each master).

For the three databases, NIST 4, FVC2002 DB1, and FVC2002 DB2, the agglomerative clustering procedure in Section III-B was carried out for the fitted mixture models to find the number of clusters $N^*$. Table I gives the value of $N^*$ and means of the following quantities for each database: $m$ and $n$, the whole fingerprint area, and $\lambda$ for the mixture models representing the theoretical mean number of impostor matches. The last column gives the mean PRC, $\overline{\mathrm{PRC}}_\alpha$, corresponding to the "12-point match" criteria (see [1] and [18]) based on the mixture models (i.e., obtaining 12 or more matches or $w = 12$). We choose $\alpha = 0.05$ to correspond to the 5% trimmed mean of the probabilities. Note that the mean values of $m$ and $n$ for the NIST and DB1 databases are similar but mean $\lambda$ of DB1 is larger than that of NIST, resulting in a much larger mean PRC for DB1 compared to NIST. Considering DB1 and DB2 now, the mean $\lambda$ remains the same but the mean number of minutiae in DB2 is much larger compared to DB1. A larger number of minutiae implies a greater chance of obtaining a random match and, hence, a larger mean PRC value.

A comparison of $\overline{\mathrm{PRC}}_\alpha$ $(\alpha = 0.05)$ was carried out for two different choices of $\lambda$ for the Poisson model: 1) $\lambda$ derived from the cluster of mixture models [see (14), (25), and (26)] and 2) $\lambda$ derived from the uniform model [see (17) and (26)]. The values of $m$ and $n$ are taken to be the mean in each database. Table II gives the results of this comparison for the NIST

TABLE I
NUMBER OF CLUSTERS, $N^*$, AS WELL AS MEAN $\lambda$ AND $\overline{\mathrm{PRC}}_\alpha$ BASED ON THE MIXTURE MODELS FOR THE THREE DATABASES. THE 2.5% GENUINE NONMATCH CRITERIA ARE CHOSEN WHEN ASSESSING THE PRCS AT $\alpha = 5\%$

| Database | $(m, n, w)$ | $N^*$ | Mean fingerprint area | Mean $\lambda$ | PRC |
|---|---|---|---|---|---|
| NIST | $(62, 62, 12)$ | 33 | $2.5 \times 10^5$ | 2.5 | $4.1 \times 10^{-4}$ |
| FVC2002 DB1 | $(63, 63, 12)$ | 9 | $1.2 \times 10^5$ | 5.09 | $5.9 \times 10^{-3}$ |
| FVC2002 DB2 | $(77, 77, 12)$ | 12 | $1.8 \times 10^5$ | 5.14 | $8.4 \times 10^{-3}$ |

TABLE II
A COMPARISON BETWEEN THE $\overline{\mathrm{PRC}}_\alpha$ OBTAINED FROM THE MIXTURE AND UNIFORM MODELS BASED ON MEAN $m$, $n$ AND THE "12-POINT MATCH" CRITERIA WITH EMPIRICAL VALUES. THE 2.5% GENUINE NONMATCH CRITERIA ARE CHOSEN WHEN ASSESSING THE PRCS AT $\alpha = 5\%$

| Database | $(m, n, w)$ | Mixture | | Uniform | | Empirical | |
|---|---|---|---|---|---|---|---|
| | | Mean $\lambda$ | PRC | Mean $\lambda$ | PRC | Mean no. of matches | PRC |
| NIST | $(62, 62, 12)$ | 2.5 | $4.1 \times 10^{-4}$ | 1.46 | $2.9 \times 10^{-7}$ | 7.1 | $3.4 \times 10^{-3}$ |
| FVC2002 DB1 | $(63, 63, 12)$ | 5.09 | $5.9 \times 10^{-3}$ | 2.95 | $1.0 \times 10^{-4}$ | 8.0 | $1.4 \times 10^{-2}$ |
| FVC2002 DB2 | $(77, 77, 12)$ | 5.14 | $8.4 \times 10^{-3}$ | 2.96 | $8.4 \times 10^{-5}$ | 8.6 | $1.9 \times 10^{-2}$ |

TABLE III
A COMPARISON BETWEEN $\overline{\mathrm{PRC}}_\alpha$ OBTAINED FROM THE MIXTURE AND UNIFORM MODELS FOR $m = n = 46$ AND $w = 12$. THE 2.5% GENUINE NONMATCH CRITERIA ARE CHOSEN WHEN ASSESSING THE PRCS AT $\alpha = 5\%$

| Database | $(m, n, w)$ | $N^*$ | Mean $\lambda$ for Mixture | Mixture | Uniform |
|---|---|---|---|---|---|
| NIST | $(46, 46, 12)$ | 33 | 1.87 | $2.25 \times 10^{-6}$ | $5.0 \times 10^{-10}$ |
| FVC2002 DB1 | $(46, 46, 12)$ | 9 | 2.72 | $5.6 \times 10^{-5}$ | $2.8 \times 10^{-7}$ |
| FVC2002 DB2 | $(46, 46, 12)$ | 12 | 1.84 | $4.1 \times 10^{-6}$ | $3.2 \times 10^{-9}$ |

and FVC2002 databases based on the whole fingerprint area. Note that the fingerprint individuality estimates using the mixture models are orders of magnitude higher compared to the uniform model. The reason for this is that when minutiae from the query and file have similar clustering tendencies, a larger number of random matches will arise compared to the uniform model. Table II also gives the empirical number of matches for every pair of impostor fingerprints in each database based on the matcher reported in [28]. This matcher optimally aligns the $m$ query minutiae $\left(X_i^Q, D_i^Q\right), i = 1, 2, \ldots, m$ with the $n$ file minutiae $\left(X_j^T, D_j^T\right), j = 1, 2, \ldots, n$ to obtain the best number of matches. The mean number of matches as well as the PRC for each database are reported in Table II in the 'Empirical' column. The empirical PRC is the proportion of impostor pairs with 12 or greater matches among all pairs with $m$ and $n$ values within $\pm 5$ of the mean. Note that the empirical number of matches and the PRCs are closer to the values derived from the mixture model compared to the uniform model, suggesting the appropriateness of the mixture model in representing the distribution of minutiae.

Since the mathematical model for the PRC was developed for any combination of $m$, $n$, and $w$, we found the trimmed mean PRC value corresponding to $m = n = 46$ and $w = 12$ for the three databases as an example. The PRCs are given in Table III for the mixture and uniform distributions. Note again that the PRCs derived from the mixture model are orders of magnitude higher compared to the uniform. Note that the PRCs depend on the choice of $(r_0, d_0)$ used. If $R_0$ (and $d_0$) is increased, we increase the chance of obtaining spurious matches and, as a result, the PRCs increase. The reverse occurs when $r_0$ (and $d_0$) is decreased. We believe the proposed methodology based on the mixture model gives a more realistic estimate of fingerprint individuality since it is a better model for minutiae clustering tendencies and intraclass variability observed in fingerprint databases.

TABLE IV
VALUES OF MEAN $m$ AND $n$ IN THE OVERLAPPING AREA, THE MEAN OVERLAPPING AREA, AND THE VALUE OF $M$ FOR EACH DATABASE

| Database | (m,n) | Mean Overlapping Area (pixel$^2$) | $M$ |
|---|---|---|---|
| NIST | (52,52) | 112,840 | 413 |
| FVC2002 DB1 | (51,51) | 71,000 | 259 |
| FVC2002 DB2 | (63,63) | 110,470 | 405 |

In the following paragraphs, we will compare our results with those of Pankanti *et al.* [18]. There are two main differences between the experiments presented in this section with the ones discussed in the previous paragraphs (i.e., Tables II and Table III). First, we consider the "corrected" uniform model of Pankanti *et al.* instead of the fully uniform model; differences between the two models will be discussed subsequently. Second, we consider now the area of overlap between the query and file instead of the whole fingerprint area. In [18], the uniform distribution was assumed for obtaining the probabilities corresponding to a location match between a query and file minutiae pair. It follows from this assumption that the number of location matches is distributed according to a hypergeometric distribution (see [18] for details). Further, Pankanti *et al.* [18] assume that the probability of a match between a pair of minutiae directions

$$P\{|D_i - D_j|_a \leq d_0\} \equiv l_0 \tag{27}$$

is independent of $i$ and $j$, and independent of the minutiae location matches. If we assume a uniform distribution on the minutiae directions and consider $d_0 = 22.5°$, the value of $l_0$ is now

TABLE V
A COMPARISON BETWEEN FINGERPRINT INDIVIDUALITY ESTIMATES USING (A) THE POISSON AND MIXTURE MODELS, AND (B) THE CORRECTED UNIFORM OF PANKANTI *et al.* [18]. THE 2.5% GENUINE NONMATCH CRITERIA ARE CHOSEN WHEN ASSESSING THE PRCS AT $\alpha = 5\%$

| Database | (m,n,w) | Empirical | | Mixture | | Pankanti | |
|---|---|---|---|---|---|---|---|
| | | Mean no. of matches | PRC | Mean $\lambda$ | PRC | Mean $\lambda$ | PRC |
| NIST | (52,52,12) | 7.1 | $3.9 \times 10^{-3}$ | 3.1 | $4.4 \times 10^{-3}$ | 1.2 | $4.3 \times 10^{-8}$ |
| FVC2002 DB1 | (51,51,12) | 8.0 | $2.9 \times 10^{-2}$ | 4.9 | $1.1 \times 10^{-2}$ | 2.4 | $4.1 \times 10^{-6}$ |
| FVC2002 DB2 | (63,63,12) | 8.6 | $6.5 \times 10^{-2}$ | 5.9 | $1.1 \times 10^{-2}$ | 2.5 | $4.3 \times 10^{-6}$ |

0.125. This choice of $l_0$, also noted in [18], is not able to represent the observed clustering characteristics of the minutiae locations and directions. Subsequently, the value of $l_0 = 0.267$ was suggested in [18] based on empirical observations. With the hypergeometric distribution for the number of matches in the minutiae locations and with the choice of $l_0 = 0.267$, the theoretical probability of obtaining exactly $w$ matches (in both minutiae location and direction) obtained in [18] is called the "corrected" uniform model. Pankanti's model can be approximated using the Poisson probabilities in (14) and mean $\lambda_U$ as in (17) with the following choices for $p_L$ and $p_D$:

$$p_L = \frac{1}{M} \text{ and } p_D = 0.267, \tag{28}$$

where $M$ is the number of cells in the overlapping area (see [18] for details). Table IV gives the mean $m$ and $n$ in the overlapping area, the mean overlapping area, and the value of $M$ for each of the three databases.

A comparison between the fingerprint individuality estimates obtained using the proposed methodology and Pankanti's approach corresponding to minutiae matches in the overlapping area is given here. The query and file fingerprints in the NIST and FVC databases are first aligned using the matcher described in [19]. Then, bounding boxes encompassing all minutiae points in the query and file fingerprints are determined. The overlap area between the two bounding boxes is taken to be the area of overlap between the query and file fingerprints. Thus, the fingerprint individuality estimates presented here are dependent on the matcher. In order to compute the Poisson probabilities using (14) and (25), we further restrict the mixture models onto the overlapping area. This is similar to (6) with $A$ representing the overlapping area, instead of the whole fingerprint area. Also, fingerprint individuality estimates based on the corrected uniform model of Pankanti *et al.* [18] are obtained. Table V gives the PRCs corresponding to the mean $m$, mean $n$, and mean overlapping area for the NIST and FVC databases. The empirical mean number of matches is obtained as before. The empirical PRC is computed as the proportion of impostor pairs with 12 or greater matches among all pairs with $m$ and $n$ values within $\pm 5$ of the mean in the overlapping area. Note that as $m$ or $n$ or both increase, the values of PRCs for both the models become large as 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 compared to the corrected uniform. Note that the mean of $\lambda$ s (the theoretical mean number of matches) as well as the PRCs corresponding to the Poisson and mixture models are closer to the empirical counterparts compared to Pankanti's model.

## V. SUMMARY AND CONCLUSION

A family of finite mixture models is proposed as a flexible and reliable way of representing minutiae variability in fingerprint images. These models are better for representing the clustering property of minutiae features observed in fingerprint images compared to the uniform distribution model assumed in [18]. We believe our model gives rise to more reliable individuality estimates that are, in fact, orders of magnitude larger than the uniform model due to minutiae clustering tendencies in query and file fingerprints pairs. In order to compute the probability of random correspondence, we developed the Poisson distribution with the mean parameter derived from the fitted mixture distributions. The PRCs depend on the choice of bounding boxes that determine a minutiae match. Larger bounding boxes yield higher spurious matches and increase the PRCs as a result. Better matching techniques that reduce the number of spurious matches will decrease the magnitude of the PRCs. Our future work will focus on further improving the models presented here further by considering spatial dependence between the observed minutiae, instead of assuming independence. We will also investigate other mixture distributions on the minutiae locations and directions that are possibly better at capturing the observed variability in the features compared to the Gaussian and Von–Mises distributions.

## ACKNOWLEDGMENT

The authors would like to thank P. K. Pathak, Y. Chen, K. Nandakumar, and S. Prabhakar for helpful advice and support during the preparation of this manuscript.

## REFERENCES

[1] B. Budowle, J. Buscaglia, and R. C. Perlman, "Review of the scientific basis for friction ridge comparisons as a means of identification: Committee findings and recommendations," *Forensic Sci. Commun.*, vol. 8, no. 1, 2006 [Online]. Available: http://www.fbi.gov/hq/lab/fsc/current/descript.html.

[2] S. Cole, "Witnessing identification: Latent fingerprint evidence and expert knowledge," *Social Studies Sci.*, vol. 28, no. 5–6, pp. 687–712, 1998.

[3] S. Cole, "What counts for identity? the historical origins of the methodology of latent fingerprint identification," *Sci. Context*, vol. 27, no. 1, pp. 139–172, 1999.

[4] S. Cole, *A History of Fingerprinting and Criminal Identification*. Cambridge, MA: Harvard University Press, 2001.

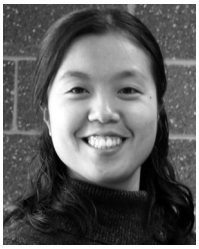
[5] S. Cole, "Is fingerprint identification valid? rhetorics of reliability in fingerprint proponents?Discourse," *Law Policy*, vol. 28, no. 1, pp. 109–135, Jan. 2006.

[6] H. Cummins and C. Midlo, *Fingerprints, Palms and Soles*. Dover, New York: , 1961.

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

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood for incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[9] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[10] J. J. Koehler, "When are peole persuaded by DNA match statistics?," *Law Human Behaviour*, vol. 25, no. 5, pp. 493–513, 2001.
[11] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*. Berlin, Germany: Springer-Verlag, 1986.
[12] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "FVC2000: fingerprint verification competition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 402–411, Mar. 2002.
[13] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "FVC2002: fingerprint verification competition," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 744–747, [Online]. Available: http://bias.csr.unibo.it/fvc2002/databases.asp.
[14] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Berlin, Germany: Springer-Verlag, 2003.
[15] K. V. Mardia, *Statistics of Directional Data*. New York: Academic Press, 1972.
[16] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley, 1997.
[17] NIST: 8-Bit Gray Scale Images of Fingerprint Image Groups (FIGS) [Online]. Available: http://www.nist.gov/srd/nistsd4.htm.
[18] S. Pankanti, S. Prabhakar, and A. K. Jain, "On the individuality of fingerprints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1010–1025, Aug. 2002.
[19] A. Ross, S. Dass, and A. K. Jain, "A deformable model for fingerprint matching," *Pattern Recognit.*, vol. 38, no. 1, pp. 95–103, 2005.
[20] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Sci.*, pp. 309–892, 2005.
[21] S. C. Scolve, "The occurence of fingerprint characteristics as a two dimensional process," *J. Amer. Statist. Assoc.*, vol. 74, no. 367, pp. 588–595, 1979.
[22] D. A. Stoney and J. I. Thornton, "A critical analysis of quantitative fingerprint individuality models," *J. Forensic Sci.*, vol. 31, no. 4, pp. 1187–1216, 1986.
[23] U. S. v. M. Byron, Criminal Action-407 (ED Pa 1999).
[24] U. S. v. Crisp, 324 F 3d 261 (4th Cir 2003).
[25] U. S. v. P. Llera, 179 F Supp 2d 492 (ED Pa 2002).
[26] U. S. v. P. Llera, 188 F Supp 2d 549 (ED Pa 2002).
[27] Y. Zhu, S. C. Dass, and A. K. Jain, Statistical models for assessing the individuality of fingerprints Michigan State University, East Lansing, MI, 2006 [Online]. Available: http://www.cse.msu.edu/cgi-user/web/tech/reports?Year=2006.
[28] Y. Zhu, S. C. Dass, and A. K. Jain, "Statistical models for fingerprint individuality," in *Proc. Int. Conf. Pattern Recognition*, Hong Kong, China, Aug. 2006, vol. 3, pp. 532–535.



**Yongfang Zhu** (S'05) received the B.S. degree from Nankai University, Tianjin, China, in 2002, where she is currently pursuing the M.Sc. and Ph.D. degrees.

Her research interests are in the areas of statistical pattern recognition, biometric authentication, and data mining. She is currently working on various aspects of statistical modeling in biometric authentication systems with applications to assess the extent of uniqueness of fingerprints and developing performance evaluation platforms for biometric systems.

**Sarat C. Dass** received the M.Sc. and Ph.D. degrees in statistics from Purdue University, West Lafayette, IN, in 1995 and 1998, respectively.

Currently, he is Associate Professor in the Department of Statistics and Probability and Adjunct Associate Professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. His research and teaching interests include statistical image processing and pattern recognition, shape analysis, spatial statistics, Bayesian computational methods, foundations of statistics, and nonparametric statistical methods.

**Anil K. Jain** (F'91) received the B.Tech. degree from the Indian Institute of Technology, Kanpur, India, and the M.S. and Ph.D. degrees from The Ohio State University, Columbus, in 1970 and 1973, respectively.

He is a University Distinguished Professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. He received the distinguished alumni award from The Ohio State University. His research interests include statistical pattern recognition, data clustering, and biometric authentication.

Dr. Jain is a fellow of the ACM, IAPR, SPIE, and AAAS. He received a Fulbright Research Award, a Guggenheim Fellowship, the Alexander von Humboldt Research Award, and the 2003 IEEE Computer Society Technical Achievement Award. He holds six patents in the area of fingerprint matching and he is the author of a number of books, including the *Handbook of Multibiometrics* (Springer, 2006); *Biometric Systems, Technology, Design, and Performance Evaluation* (Springer, 2005); *Handbook of Face Recognition* (Springer, 2005); *Handbook of Fingerprint Recognition* (Springer, 2003), which received the PSP award from the Association of American Publishers; *BIOMETRICS: Personal Identification in Networked Society* (Kluwer, 1999); and *Algorithms for Clustering Data* (Prentice-Hall, 1988). ISI has designated him as a highly cited researcher. He is an associate editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and *ACM Transactions on Knowledge Discovery in Data*. He received the 1996 IEEE TRANSACTIONS ON NEURAL NETWORKS Outstanding Paper Award and best paper awards from the Pattern Recognition Society in 1987 and 1991. He was the editor-in-chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is a member of the National Academies panels on Whither Biometrics and Improvised Explosive Device.