# Computational Methods for Determining Individuality 

Sargur N. Srihari and Chang Su<br>Center of Excellence for Document Analysis and Recognition (CEDAR)<br>University at Buffalo, State University of New York<br>Amherst, New York 14228, U.S.A<br>\{srihari, changsu\}@cedar.buffalo.edu


#### Abstract

Individuality is the state or quality of being an individual. We establish a computational methodology to determine whether a particular modality of data is sufficient to establish the individuality of every individual or even a demographic group. To test the individuality, generative models are given or learned to represent the distribution of certain characteristics such as birthday, human heights and fingerprints. Given the individuality assessments of different characteristic, the models based on multiple characteristics are proven to get strengthen individuality.


Key words: individuality, generative model, fingerprint

## 1 Introduction

As commonly used, individual refers to a person or to any specific object in a collection. From the seventeenth century on, individual indicates separateness, as in individualism. Individuality is the state or quality of being an individual; a person separate from other persons. Individuality study is important in forensic identification since we need to assess whether a particular input such as gender, height, weight or fingerprints can be used to identify specific person from the trace evidence they leave, often at a crime scene or the scene of an accident.

Studies on the individuality started from the late 1800s. About 20 models have been proposed since then trying to establish the improbability of two random people having the same certain characteristics. All the models try to quantify the uniqueness property to be able to defend forensic identification as a legitimate proof of identification in the courts. Each of these models try to find out the probability of false correspondence, i.e. probability that a wrong person is identified given a certain evidence collected from a crime scene from a set of previously recorded whole database. i.e., the probability that the features of two fingerprints match though they are taken from different individuals. A match here does not necessarily mean an exact match but a match within given tolerance levels. All the models establish the probability of two different people being identified as the same based on their features, namely, the probability of random correspondence (PRC). The models have been classified based on the different approaches that have been taken through a century of individuality studies. The
latest class of models is called generative models. Generative models are statistical models that represent the distribution of the feature. In these models, a distribution of the features is learnt through a training dataset. Features are then generated from this distribution to test their individuality. What training set is used is immaterial as long as it is representative of the entire population.

The following of this paper is organized as follows: We discuss individuality computational method of birthday in Section 2 and heights in Section 3. Section 4.1 introduces a new generative model for both minutiae and ridges and individuality computation are also given. The paper concludes with a summary in Section 5.

## 2 Individuality of Birthday

Generative models for determining individuality can be understood by considering the trivial example of using the birthday of a person. The birthday problem asks whether any of the $k$ people have a matching birthday with any of the others.

To compute the approximate probability that in a room of $k$ people, at least two have the same birthday, we disregard variations in the distribution, such as leap years, twins, seasonal or weekday variations, and assume that the 365 possible birthdays are equally likely. Therefore the PRC value for birthday problem is $1 / 356$. Uniform density is used here to model the birthday distribution.

It is easier to first calculate the probability $p(k)$ that all n birthdays are different. If $k>365$, by the pigeonhole principle this probability is 0 . On the other hand, if $k \leq 365$, it is by the pigeonhole principle this probability is 0 . On the other hand, if k 365 , it is

$$
\begin{equation*}
\bar{p}(k)=1 \times\left(1-\frac{1}{365}\right) \times\left(1-\frac{2}{365}\right) \cdots\left(1-\frac{k-1}{365}\right)=\frac{365!}{365^{k}(365-k)!} \tag{1}
\end{equation*}
$$

The event of at least two of the $k$ persons having the same birthday is complementary to all $k$ birthdays being different. Therefore, its probability $p(k)$ is

$$
\begin{equation*}
p(k)=1-\bar{p}(k) \tag{2}
\end{equation*}
$$

This probability surpasses $1 / 2$ for $k=23$ (with value about $50.7 \%$ ).

## 3 Individuality of Human Height

The goal of the generative model for height, is to come up with an analytical value for the probability of two individuals having the same height within some tolerance $\pm \epsilon$. Different with the birthday problem, PRC value can not be computed directly. We have to learn the parameters of the generative model firstly. The steps in studying individuality using a generative model are given below.

1. Consider a probabilistic generative model and estimate its parameters from a particular data set.
2. Evaluate analytically the probability of two individuals to have the same height(or other bio-metric), with some tolerance $\pm \epsilon$.

For the study of individuality of height, a Gaussian density is a reasonable model to fit the distribution of heights of individuals. The height statistics is collected from CDC Advance Data No. 361 [1]. Figure 1 shows modeling the heights (inch) for males and females aged 20 years and over using a Gaussian p.d.f. with mean $\mu_{f}=63.8, \mu_{m}=69.3$ and standard deviation $\sigma_{f}=11.1$, $\sigma_{m}=3.3$. Now the probability of two individuals having the same height with some tolerance $\pm \epsilon$ can be derived as follows.


Fig. 1. Gaussian density used to model heights of individuals $\mu=5.5$ and $\sigma=0.5$ i.e. mean 5.5 feet and standard deviation 6 inches.

Probability of one individual having height $a \pm \epsilon$ is $\int_{a-\epsilon}^{a+\epsilon} P(h \mid \mu, \sigma) d h$

$$
\text { where } \quad P(h \mid \mu, \sigma) \sim \mathcal{N}(\mu, \sigma)=\frac{1}{\sqrt{2 \pi} \sigma} e^{-\frac{(h-\mu)^{2}}{2 \sigma^{2}}}
$$

Probability of two individuals having height $a \pm \epsilon$ is $\quad\left(\int_{a-\epsilon}^{a+\epsilon} P(h \mid \mu, \sigma)\right)^{2}$

Probability of two individuals having any same height $\pm \epsilon$ is

$$
\begin{equation*}
p_{\epsilon}=\int_{-\infty}^{\infty}\left(\int_{a-\epsilon}^{a+\epsilon} P(h \mid \mu, \sigma) d h\right)^{2} d a \tag{3}
\end{equation*}
$$

Eq 3 can be numerically evaluated for a given value of $\mu, \sigma$. Figure 2 shows the probability values for fixed $\mu_{f}=63.8, \mu_{m}=69.3$ and varying $\sigma_{f}, \sigma_{m}$. A
tolerance of 0.1 inches was used in the probability calculations. It is obvious to note that, when $\sigma$ decreases, the width of the Gaussian is smaller and hence the probability that two individuals having the same height is more.


Fig. 2. Individuality of heights calculated using a Gaussian as a generative model. For different values of $\sigma$ and fixed $\mu_{f}=63.8, \mu_{m}=69.3$, the probabilities are calculated.

The Probability of Random Correspondence for female height with a mean height of 63.8 inch and standard deviation of 11.1 inches is $p_{\epsilon}=0.0025$. i.e. 25 out of every 10000 female have the same height (tolerance of 0.1 inches) and the Probability of Random Correspondence for male height with a mean height of 69.3 inch and standard deviation of 3.3 inches is $p_{\epsilon}=0.0085$. i.e. 85 out of every 10000 men have the same height (tolerance of 0.1 inches)

Based on this model, the probability of at least two people sharing a height among $k$ individuals can be estimated. This evaluation is similar to that of the birthday problem where the probability of two people having the same birthday among $k$ individuals is calculated. In the case of heights we have a real value instead of 365 date value. This is handled by the implicit discreteness due to the tolerance $\epsilon$.

Assuming that there are $h$ possible height and all $h$ possible heights are equally likely, $\bar{p}(k)$ is given by

$$
\begin{equation*}
\bar{p}(k)=1 \cdot\left(1-\frac{1}{h}\right) \cdot\left(1-\frac{2}{h}\right) \cdots\left(1-\frac{k-1}{h}\right)=\frac{h!}{h^{k}(h-k)!} \tag{4}
\end{equation*}
$$

The event of at least two of the $k$ persons having the same height is complementary to all $k$ heights being different. Therefore, its probability $p(k)$ is

$$
\begin{equation*}
p(k)=1-\bar{p}(k) \tag{5}
\end{equation*}
$$

Thus $p(2)=1-\bar{p}(2)=1-1 \cdot\left(1-\frac{1}{h}\right)$, since $p_{\epsilon}=p(2)$ we have $h=1 / p_{\epsilon}$.
The Table 1 shows the probability of at least two people sharing a same birthday or height amongst a certain number of people, assuming $p_{\epsilon}=0.025$.

In a group of 10 (or more) randomly chosen people, there is more than $11 \%$ probability that some pair of them will have the same birthday and $10 \%$ for female and $32 \%$ for male probability that some pair of them will have the same height. For 80 or more people, the probability is more than $99 \%$, tending toward $100 \%$ as the pool of people grows.

Table 1. the probability of at least two people sharing a same birthday or height amongst a certain number of people

| $\mathbf{k}$ | $\mathbf{p ( k )}$ |  |  |
| :---: | :---: | :---: | :---: |
|  | Birthday | Female Heights | Male Heights |
| 2 | 0.0028 | 0.0025 | 0.0085 |
| 5 | 0.0277 | 0.0248 | 0.0825 |
| 10 | 0.1194 | 0.1072 | 0.3251 |
| 20 | 0.4184 | 0.3830 | 0.8196 |
| 40 | 0.8966 | 0.8670 | 0.9995 |
| 80 | 0.9999 | 0.9998 | $1-4 \times 10^{-13}$ |
| 120 | $1-1.9 \times 10^{-6}$ | $1-2.2 \times 10^{-6}$ | 1 |
| 200 | $1-2.7 \times 10^{-11}$ | $1-3.1 \times 10^{-11}$ | 1 |
| 300 | $1-7.0 \times 10^{-22}$ | $1-7.2 \times 10^{-22}$ | 1 |
| 400 | 1 | 1 | 1 |

## 4 Individuality of Fingerprints

Fingerprints have been used for identification from the early 1900s. Their use for uniquely identifying a person has been based on two premises, that, (i) they do not change with time and (ii) they are unique for each individual. Until recently, fingerprints had been accepted by courts as a legitimate proof of identification. But, after the 1999 case US vs Byron Mitchell, fingerprint identification has been challenged under the basis that the premises stated above have not been objectively tested and the error rates have not been scientifically established. Though the first premise has been accepted, the second one on individuality is the widely challenged one.

### 4.1 Generative Models for Fingerprint Individuality

Studies on the individuality of fingerprints date back to the late 1800s. All previous models can be classified into five different categories, namely, gridbased models, ridge-based models, fixed probability models, relative measurement models and generative models. Grid-based models include Galton [2] and Osterburgh [3] which were proposed in the late 80 s and the early 90 s respectively. One instance of ridge-based models is introduced by Roxburgh [4]. Fixed probability models contain the class of Henry-Balthazard models [5]. Relative
measurement models include the Champod model [6] and the Trauring model [7]. The latest class of models, namely, the generative models aim at being flexible to represent observed distributions through different fingerprint databases and then ascertained uncertainties from models. Based on the the assumed non-independence of minutia locations and orientations, various mixture models could be used [8] and [9].


Fig. 3. (a) Minutiae: ridge ending and ridge bifurcation (b) detected minutiae and ridge points on a skeleton fingerprint image

In existing generative models only minutiae have been modeled without considering ridge features. Minutiae means small details in the fingerprints, it refers to the ridge endings and ridge bifurcation. See Figure 3 for the examples of two kinds of minutiae and the minutiae on a skeleton fingerprint image. We further embed ridge information into existing generative models by using the distribution for ridge points. The proposed model offers more reasonable and accurate fingerprint representation and therefore a more reliable probability of random correspondence $(P R C)$. In this model, the ridge is represented as a set of ridge points sampled at equal interval of inter ridge width. Ridge length is defined as the number of ridge points that could be sampled from the ridge. Three types of ridges are defined as (i) short ridges: $l(r) \leq L / 3$, (ii) medium ridges: $L / 3<l(r)<2 L / 3$ and (iii) long ridges: $2 L / 3 \leq l(r) \leq L$, where $L$ is the maxima ridge length which was collected from the FVC 2002 database.

These three possible ridge length types can be associated with any minutiae. Without loss of generality, we can assume that there exist only three possible ridge length types corresponding to a minutiae. For the generative model, the ridge length type is modeled as a uniform distribution $F^{l}\left(l_{r} \mid a, b\right)$, where $[a, b]$ is the interval of the uniform distribution. For ridges with different lengths, different ridge points are picked as anchors. The index to be used for ridge point selection should satisfy following two conditions: 1) The index should be
large so as to infer as many other ridge points as possible. 2) The index should not be too large to overstep the ridge length. A tradeoff has to be balanced between the two conditions [10]. For medium ridges, $(L / 3)^{t h}$ ridge point is picked and for long ridges, both $(L / 3)^{t h}$ and $(2 L / 3)^{t h}$ are picked. None ridge point will be chosen for short ridges. For the generative model, the ridge points are modeled as a combining distribution of the ridge point location and the direction. The proposed joint distribution model for fingerprint presentation is based on a mixture consisting of G components. Each components is distributed according the density of the minutiae and the $i^{t h}$ ridge points: $F_{g}^{m} F_{g}^{i}$. The equation of the generative model is given in Eq. 6.

$$
f\left(\cdot \mid \Theta_{G}\right)= \begin{cases}F^{l}\left(l_{r}\right) \cdot \sum_{g=1}^{G_{1}} \pi_{g} F_{g}^{m}\left(s_{m}, \theta_{m} \mid \Theta_{G}\right) & l_{r} \leq L / 3  \tag{6}\\ F^{l}\left(l_{r}\right) \cdot \sum_{g=1}^{G_{2}} \pi_{g} F_{g}^{m}\left(s_{m}, \theta_{m} \mid \Theta_{G}\right) & L / 3<l_{r}<2 L / 3 \\ \cdot F_{g}^{L / 3}\left(r_{L / 3}, \phi_{L / 3}, \theta_{L / 3} \mid \Theta_{G}\right) & \\ F^{l}\left(l_{r}\right) \cdot \sum_{g=1}^{G_{3}} \pi_{g} F_{g}^{m}\left(s_{m}, \theta_{m} \mid \Theta_{G}\right) \cdot F_{g}^{L / 3}\left(r_{L / 3}, \phi_{L / 3}, \theta_{L / 3} \mid \Theta_{G}\right) & \\ \cdot F_{g}^{2 L / 3}\left(r_{2 L / 3}, \phi_{2 L / 3}, \theta_{2 L / 3} \mid \Theta_{G}\right) & l_{r} \geq 2 L / 3\end{cases}
$$

In Eq. $6, F_{g}^{m}(\cdot)$ represents the distribution of the minutiae location $s_{m}$ and the direction $\theta_{m} . F_{g}^{i}(\cdot)$ presents the distribution of the $i^{t h}$ ridge points. To estimate the unknown parameters in the generative model, we develop an algorithm based on the EM algorithm. Different numbers of the components $G$ for the mixture model were validated using $k$-means clustering. The one with the best $k$-means clustering results was chosen.

### 4.2 Fingerprint Individuality Computation

Given a template $T$ with $n$ minutiae and an input/query $Q$ with $m$ minutiae and corresponding ridge points pairs and $w$ out of them match, the probability of Random Correspondence is given by

$$
\begin{align*}
& P R C_{0}=p^{*}(w ; Q, T)= \\
= & \binom{n}{w} \cdot\left(p_{m}(Q, T)\right)^{w}\left(1-p_{m}(Q, T)\right)^{n-w} \tag{7}
\end{align*}
$$

The probability is a binomial probability whose parameters are n and $p_{m}(Q, T)$. The latter is the probability that a random minutiae and corresponding ridge points pair from $Q$ will match a pair from T. Since most of the matchers try to maximize the number of matchings (i.e. they would find a matching even in a fingerprint that are totally different, we calculate the conditional expectation, conditioned on that fact that the number of matches is always greater than zero and equating this to the number of pair matches between Q and T , the estimation can be written as

$$
\begin{equation*}
\frac{n \cdot p_{m}(Q, T)}{\left(1-\left(1-p_{m}(Q, T)\right)^{n}\right)}=w_{0} \tag{8}
\end{equation*}
$$

$w_{0}$ is found out by the proposed models fit into Q and T and determining the number of matches by $k$-plet [11] matching algorithm. Value of $p_{m}(Q, T)$ can be found from Eq 8 . In a database contains N different fingers with L impressions of the same finger, $(Q, T)_{\text {impostor }}$ is used to denote all the $N(N-1) L^{2} / 2$ impostor pairs.

$$
\begin{equation*}
\overline{P R C}=\frac{1}{N(N-1) L^{2} / 2} \sum_{(Q, T)_{\text {impostor }}} p^{*}(w ; Q, T) \tag{9}
\end{equation*}
$$

### 4.3 Experiments and Results

Generative models without ridge information and with the ridge information model introduced in 4.1 have been implemented and experiments have been conducted on FVC2002 DB1 [12]. The number of components $G$ for the mixture model was found after validation using $k$-means clustering. The database has 100 different fingerprints with 8 impressions of the same finger. Thus, there are a total of 800 fingerprints using which the model has been developed.

Table 2. PRC for different fingerprint matches with varying m (number of minutiae in template), $n$ (number of minutiae in input) and $w$ (number of matched minutiae or minutiae and corresponding ridge points pairs) - With ridge information and without ridge information. $\mathbf{P R C}_{\mathbf{0}}$ is PRC for the general population and $\overline{\mathbf{P R C}}$ is for PRC for FVC2002-DB1.


We compare the results to that of [9]. Random fingerprints are generated from the model. Values of $P R C_{0}$ and $\overline{P R C}$ are calculated using the formulae introduced in Section 4.1. The results are presented in Table 2. The PRCs are calculated through varying number of minutiae in template $(m)$, $\operatorname{Input}(n)$ and the number of ridges matched $(w)$. Table 2 shows that more the number of minutiae in the template and the input, the higher the PRC. In experiments conducted on the FVC2002 DB1, there are some differences between the results obtained here and the results in Jain et al. [9]. This may result from use of different matching algorithms, which $w_{0}$ depends on. Our highlight is that the PRC values embedded with ridge information model are never greater than PRC values without ridge information. Table 2 also shows the PRC values corresponding to use of ridge information model in the generative model and these probabilities are lesser when compared to those without ridge information which indicates that ridge information strengthens individuality of fingerprints. The PRCs for the different $m$ and $n$ with $6,16,26$ and 36 matching ridges are shown in Figure 4. It is obvious to note that, when $w$ decreases or $m$ and $n$ increase, the probability that two random fingerprints matching is more.


Fig. 4. PRCs with different number of the matched ridges for (a) $w=6$, (b) $w=16$, (c) $w=26$, and (d) $w=36$.

Because the ridge information models are independent on generative models, other recently proposed generative models such as [13] could also be embedded
with ridge information in similar manner and are also expected to offer more reliable PRC values.

The probability of at least two fingerprints matched among a certain number of fingerprints is computed similarly as Eq. 5 as well and given by Table 3. In 100,000 randomly chosen fingerprints, there is only $7.72 \times 10^{-15}$ probability that some pair of them will match if we consider both minutiae and ridge in matching. This probability is much smaller than previous minutiae only model which is $5.90 \times 10^{-6}$. The probability of at least two fingerprints matched among U.S. and world population are $7.09 \times 10^{-8}$ and $3.42 \times 10^{-5}$ respectively.

Table 3. the probability of at least two fingerprints matched among a certain number of fingerprints with average number of minutiae $m=n=39$ and average number of matched minutiae $w=27$

|  | Minutiae and Ridge Information only Minutiae |  |
| :---: | :---: | :---: |
| $\mathbf{k}$ | $\mathbf{p}(\mathbf{k})$ | $\mathbf{p}(\mathbf{k})$ |
| 2 | $1.54 \times 10^{-24}$ | $1.18 \times 10^{-15}$ |
| 5 | $1.54 \times 10^{-23}$ | $1.18 \times 10^{-14}$ |
| 10 | $6.93 \times 10^{-23}$ | $5.31 \times 10^{-14}$ |
| 100 | $7.64 \times 10^{-21}$ | $5.84 \times 10^{-12}$ |
| $10^{5}$ | $7.72 \times 10^{-15}$ | $5.90 \times 10^{-6}$ |
| $3.03 \times 10^{8}$ | $7.09 \times 10^{-8}$ | $1.12 \times 10^{-2}$ |
| $6.66 \times 10^{9}$ | $3.42 \times 10^{-5}$ | $6.77 \times 10^{-2}$ |
| $10^{10}$ | $7.72 \times 10^{-5}$ | $7.02 \times 10^{-1}$ |
| $8.48 \times 10^{14}$ | $1.03 \times 10^{-1}$ | 1 |
| $10^{20}$ | 0.99999999999999946 | 1 |
| $6.48 \times 10^{23}$ | 1 | 1 |

## 5 Summary

Generative models of individuality attempt to model the distribution of features and then use the models to determine the probability of random correspondence. This paper provides a detailed survey of individuality models. We have analyzed 3 models based on birthday, heights and fingerprints. For birthday model, generative model fits uniform distribution and PRCs can be gotten directly. Human height model uses Gaussian density to present the height distribution. PRCs are computed from the generative model leaned form statistic data. Fingerprint individuality computation is the most complicated one. A generative model with an mixture distribution is used to model both minutiae and ridge information. The new generative models are learned and then compared by the experiments with the generative model without ridge information on the FVC2002 DB1. The PRCs obtained for a fingerprint template and input with 36 minutiae each with 16 matching minutiae is $1.4 \times 10^{-10}$ (or 14 in a 100,000 million, or equivalently,

1 in 7,000 million). This is a much stronger result than without using ridge information which is 1 in 200,000 . With 20 matching minutiae this probability is one in 300 trillion, as opposed to the earlier result of 1 in 100 million in [9]. Since proposed ridge information model offers a more reasonable and more accurate fingerprint representation, PRC values with ridge information are much smaller than PRC values without ridge information.

Acknowledgments. This work was supported by a grant from the Department of Justice, Office of Justice Programs Grant NIJ 2005-DD-BX-K012. The opinions expressed are those of the authors and not of the DOJ.

## References

1. McDowell, M.A., Fryar, C.D., Hirsch, R., Ogden, C.L.: Anthropometric reference data for children and adults: U.s. population, 1999c2002. National Health and Nutrition Examination Survey (2003)
2. Galton, F.: FingerPrints. McMillan, London (1892)
3. Osterburg, J.: Development of a mathematical formula for the calculation of fingerprint probabilities based on individual charectiristics. Journal of American Statistical Association 772 (1997) 72
4. Roxburgh, T.: Galton's work on the evidential value of fingerprints. Indian Journal of Statistics 1 (1933) 62
5. Henry, E.: Classification and Uses of FingerPrints. Routledge \& Sons London (1900)
6. Champod, C., Margot, P.: Computer assisted analysis of minutiae occurrences on fingerprints. Proc. International Symposium on Fingerprint Detection and Identification, J. Almog and E. Spinger, editors, Israel National Police, Jerusalem (1996) 305
7. Trauring, M.: Automatic comparison of finger-ridge patterns. Nature (1963) 197
8. Pankanti, S., Prabhakar, S., Jain, A.K.: On the individuality of fingerprints. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(8) (2002)
9. Dass, S., Zhu, Y., Jain, A.K.: Statistical models for assessing the individuality of fingerprints. Fourth IEEE Workshop on Automatic Identification Advanced Technologies (2005) 3-9
10. Fang, G., Srihari, S.N., Srinivasan, H., Phatak, P.: Use of ridge points in partial fingerprint matching. Proc. of SPIE: Biometric Technology for Human Identification IV (2007) 65390D1-65390D9
11. Chikkerur, S., Cartwright, A.N., Govindaraju, V.: K-plet and cbfs: A graph based fingerprint representation and matching algorithm. International Conference on Biometrics (2006) 309-315
12. Maio, D., Maltoni, D., Cappelli, R.: Fingerprint verification competition. http://bias.csr.unibo.it/fvc2002/ (2002)
13. Zhu, Y., Dass, S., Jain, A.: Compound stochastic models for fingerprint individuality. Proc. of International Conference on Pattern Recognition 3 (2006) 532-535
