Evaluation techniques for biometrics-based authentication systems (FRR)

Ruud M. Bolle, Sharath Pankanti and Nalini K. Ratha
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598
{bolle, ratha, sharat}@us.ibm.com

Abstract

Biometrics-based authentication is becoming popular because of increasing ease-of-use and reliability. Performance evaluation of such systems is an important issue. We attempt to address two aspects of performance evaluation that have been conventionally neglected. First, the "difficulty" of the data that is used in a study influences the evaluation results. We propose some measures to characterize the data set so that the performance of a given system on different data sets can be compared. Second, conventional studies often have reported the false reject and false accept rates in the form of match score distributions. However, no confidence intervals are computed for these distributions, hence no indication of the significance of the estimates is given. In this paper, we systematically study and compare parametric and nonparametric (bootstrap) methods for measuring confidence intervals. We give special attention to false reject rate estimates.

1 Introduction

Automated biometrics is the science of authenticating or identifying subjects based on their physiological or behavioral characteristics such as fingerprint, face, voice and signature. In an authentication system, we are interested in confirming whether the presented biometrics is in some sense "close" to the enrolled biometrics of the same user. Authentication involves $1:1$ matching in contrast to identification which involves $1:N$ matching.

A biometrics signal is a pattern and authenticating a person with a fingerprint or other biometrics is in essence an exercise in pattern recognition. Let the stored fingerprint be presented by template P' and the acquired fingerprint by P. In terms of hypothesis testing, we have

$$\begin{aligned} H_0: &\quad P = P', \quad \text{the person is genuine} \\ H_1: &\quad P \neq P', \quad \text{the person is an impostor.} \end{aligned}$$

Often some similarity measure $s = Sim(P, P')$ is defined and H_0 is decided if $s \geq Th$ and H_1 is decided if $s < Th$. Deciding H_0 when H_1 is true gives a false accept; deciding H_1 when H_0 is true results in a false reject.

Reporting the matching performance of biometric systems is a controversial issue and perhaps an issue that is not well understood. There are two components in performance evaluation: (i) the test data set and (ii) the matcher. When reporting the performance accuracy, the size of the set of fingerprint impressions used to estimate the error rates, is a parameter that definitely should be given. It should also be required to provide other details about the test data set. Such details include a characterization of the subject population.

False Accept Rates (FAR) and False Reject Rates (FRR) are important intrinsic characteristics of a matcher. Given a matcher, in theory the FRR and FAR could be determined analytically. For instance, in a fingerprint authentication scenario, if all sources of noise, such as sensor noise, feature noise and distortions between pairs of matching finger templates could be modeled, the error rates could be computed. Probably the most difficult issue here are the sources of noise introduced by imaging the fingerprints. Clearly, it is impossible to model all noise sources and one has to resort to statistical techniques to estimate the error rates. Here it is important to remember that the reported error rates are only estimates of the true error rates, and only that.

The threshold Th is an important parameter in an authentication system. It determines the tradeoff between false reject rate and false accept rate of a system. Statistical performance evaluation of biometrics authentication or identification systems in terms of FRR and FAR as a function of Th is a difficult issue, if addressed at all. All too often, error rates are not reported or poorly reported, or, worse, the systems are claimed to be 100% accurate. At the very least, the Equal Error Rate (where FRR = FAR) should be reported, it is desirable, however, to report system accuracy with a Receiver Operating Curve (ROC) [3, 4]. This is a graph that expresses the relation between FRR and FAR when the matching threshold Th is varied. However, the ROC in itself does not mean much. As stated before, what also needs to be reported is the data set size of fingerprint impressions that is used to compute these statistics. In addition, some indication should be given of the quality of the impressions, e.g., the conditions under which the prints are collected and a description of the subjects that are used for acquiring the database. Finally, it should be reported how accurate the

0-7695-0750-6/00 $10.00 © 2000 IEEE

estimates of the above statistics really are. Given that the samples in the database are representative, the accuracy of the estimates depends on the database size—the larger the size, the more accurate the estimates. These issues can be addressed by computing confidence intervals

Other methods for evaluating authentication systems can be found in [1, 8]. In our earlier work [7], we presented our results of confidence interval estimation for authentication systems both on distributions (to construct ROCs with confidence intervals) and on distribution parameters. An interesting observation has been made in [5] for correlating authentication performance to identification performance.

In this paper, we describe techniques for characterizing data sets that are used for accuracy estimation and compare results of FRR confidence interval estimation using parametric and bootstrap-based technique. Section 2 describes the parametric and bootstrap techniques and introduces the definitions used in this paper. Some techniques for data set evaluation are presented in Section 3. Procedures for evaluating the validity and accuracy of the FRR estimates for biometrics-based authentication systems are presented in Section 4. In Section 5, we discuss issues with parametric and nonparametric techniques and give some conclusions.

2 Performance evaluation

With fingerprint as an example biometrics, we introduce parametric techniques and nonparametric bootstrap techniques for computing confidence measures. To evaluate a fingerprint authentication system, a set of matching fingerprint pairs $\mathbf{M}_0 = \{a_1, ..., a_m\}$ and a set of nonmatching pairs $\mathbf{M}_1 = \{b_1, ..., b_n\}$ needs to be acquired. Here $m < n$ because collecting prints from subjects always results in more nonmatching than matching pairs. (From this it immediately follows that the FAR can be estimated more accurately.) Matching these sets of pairs results in m scores s of matching fingers and n scores s of nonmatching fingers. We denote these sets of scores by $\mathbf{X} = \{X_1, ..., X_m\}$ and $\mathbf{Y} = \{Y_1, ..., Y_n\}$, respectively.

Let us concentrate on the set $\mathbf{X}$ and assume that this is a sample of m numbers drawn from a population with distribution F. That is, $F(x) = Prob(X \leq x)$, with $x \geq 0$. The function F is the probability distribution or cumulative distribution function of match scores s of matching pairs. We can study this sample $\mathbf{X}$ in order to estimate a certain characteristic, $\theta(F)$, associated with F. Just as the distribution function F of matching scores, the distribution of $\theta(F)$ is of unknown form. A statistic, $T = T(\mathbf{X})$ may be used to estimate $\theta(F)$ from the data $\mathbf{X}$. (Here we assume that T is unbiased.) The $\theta(F)$ that are of interest in this paper are F for the match scores s, i.e., $\theta(F) = F(\mathbf{X})$.

To compute the confidence intervals, we have a choice between using parametric or nonparametric methods.

2.1 Parametric performance evaluation

Parametric evaluation methods impose assumptions about the available data and the distribution underlying the samples. Under the simplifying assumptions, the performance evaluation problem reduces to estimating a few parameters of a "known" distribution. Here we assume that multiple impressions (genuine-genuine) of a given biometric identifier provide i.i.d. match scores.

Remember, we have m match scores $\mathbf{X} = \{X_1, ..., X_m\}$ of our mated pairs. What we need is a confidence interval for the estimate $\hat{F}(x)$ of $F(x)$ at some $x > 0$. Here we have

$$\hat{F}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\,(X_i \leq x) = \frac{1}{m}(\#\, X_i \leq x)$$

The parametric assumption then is that $m\,F(x)$ is a binomial random variable Z with parameters m, $0 < p(x) < 1$ and mass distribution

$$P(Z = z) = \binom{m}{z} p^z (1-p)^{m-z},\ z = 0, ..., m$$

The expectation of Z, $E(Z) = mp(x) = mp$, and an estimate of the mean is $m\hat{F}(x)$. After all, the binomial distribution represents the number of successes in m trials (where a success is $X_i \leq x$). The variance of $mF(x)$ is $mp(x)(1 - p(x))$, so an estimate of the variance is $\mathrm{var}\,(\hat{F}(x)) = \hat{F}(x)(1 - \hat{F}(x))/m$.

A $0 \geq \beta \geq 1$ confidence interval for $\hat{F}(x)$, $[K_1, K_2]$ is given by the K_1 satisfying [8]

$$\beta/2 = Prob\,(z \leq K_1) = \sum_{z=0}^{K_1} \binom{m}{z} p^z (1-p)^{m-z}$$

and $K_2 = \lfloor m - K_1 \rfloor$.

In practice, this is computationally intensive and may give some numerical problems. Following the law of large numbers, one could assume that $\hat{F}(x)$ is distributed according to the normal distribution, i.e., $Prob\,(F(x) \leq Z) \sim \mathcal{N}(F(x), \sigma(x))$. Where an estimate of $\sigma(x)$ is given by

$$\hat{\sigma}(x) = \sqrt{\frac{\hat{F}(x)(1 - \hat{F}(x))}{m}} \qquad (1)$$

Confidence intervals can be determined with percentiles of the normal distribution, e.g., a 90% confidence interval is

$$-1.645\,\hat{\sigma}(x) < \hat{F}(x) < 1.645\,\hat{\sigma}(x)$$

2.2 Bootstrap performance evaluation

No parametric form of the distribution of match and nonmatch scores are assumed here. Here we need a measure of the statistical accuracy of the point estimator $T(\mathbf{X})$. This because, in general, the estimator $T = T(\mathbf{X})$ is not equal to $\theta(F)$ and we would like to get some idea of the statistical

properties of the error $T(\mathbf{X}) - \theta(F)$. In other words, we are interested in how much importance should be given to T. One way to achieve this is to compute the $(1-\alpha)100\%$ confidence interval for $T(\mathbf{X})$ in the form $[q^\star(\alpha/2), q^\star(1-\alpha/2)]$ where $q^\star(\alpha/2)$ and $q^\star(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of $Dist_T(x) = Prob\,(T(\mathbf{X}) \leq x)$. The bootstrap principle prescribes sampling, with replacement, the set $\mathbf{X}$ a large number (B) times, amounting to the sets $\mathbf{X}_i^\star, i = 1, ..., B$ and calculating many estimates $T_i^\star = T_i^\star(\mathbf{X}_i^\star)$, $i = 1, ..., B$. Determining confidence intervals amounts to what essentially are counting exercises. One can compute a (say) 90% confidence interval by counting the bottom 5% and top 5% of the estimates $T_i^\star$ and subtracting these estimates from the set $\{T_i^\star\}$. The leftover set determines the confidence interval.

We are interested in estimating the probability distribution $F(x)$ at some point x_o, that is, a characteristic of F, $\theta(F) = F(x_o)$. Here we have to do the best with what we have, i.e., the observed sample $\mathbf{X} = \{X_1, ..., X_m\}$, since we do not have the whole population. The sample population $\mathbf{X}$ has distribution $\hat{F}$, which is called empirical distribution and is defined as

$$\hat{F}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\,(X_i \leq x) = \frac{1}{m}(\#\, X_i \leq x)$$

which puts equal mass $1/m$ at each observation x_i.

What we are interested in is $\theta(F) = F(x_o)$, but all that we can obtain is an estimate $\hat{F}(x_o)$. This estimate *is itself* a random variable which has distribution $\hat{G}(y) = Prob\,(Y \leq y) = Prob\,(F(x_o) \leq y)$. We can obtain a bootstrap confidence interval for $\hat{F}(x_o)$ by sampling with replacement from $\mathbf{X}$ as follows.

1. Calculate the estimate $\hat{F}(x_o)$ from the sample $\mathbf{X}$.
2. *Resampling.* Create a bootstrap sample $\mathbf{X}^\star = \{X_1^\star, ..., X_m^\star\}$ by sampling $\mathbf{X}$ with replacement.
3. *Bootstrap estimate.* Calculate $\hat{F}^\star(x_o)$ from $\mathbf{X}^\star$.
4. *Repetition.* Repeat steps 2-3 B times (B large), resulting in $\hat{F}_1^\star(x_o)$, $\hat{F}_2^\star(x_o)$ ... $\hat{F}_B^\star(x_o)$. The distribution $\hat{G}^\star$ of $\hat{F}(x_o)$ is given by

$$\hat{G}^\star(y) = \frac{1}{B}\sum_{i=1}^{B} \mathbf{1}(\hat{F}_i^\star(x_o) \leq y) = \frac{1}{B}(\#\hat{F}^\star(x_o) \leq y)$$

To obtain a bootstrap estimate of a confidence interval of $\hat{F}(x_o)$, we sort the bootstrap estimates into increasing order to obtain $\hat{F}^\star_{(1)}(x_o) \leq \hat{F}^\star_{(2)}(x_o) \leq ... \leq \hat{F}^\star_{(B)}(x_o)$. A $(1-\alpha)100\%$ bootstrap confidence interval is $(\hat{F}^\star_{(q_1)}(x_o), \hat{F}^\star_{(q_2)}(x_o))$, where $q_1 = \lfloor B\alpha/2 \rfloor$ the integer part of $B\alpha/2$ and $q_2 = B - q_1 + 1$.

The bootstrap supplies very powerful methods for estimating the distribution of estimates of parameters or characteristics of unknown distributions. This, in turn, supplies estimates of confidence intervals. In Sections 3 and 4, we give some examples of the power of the bootstrap for evaluating fingerprint matchers.

2.3 The data may not be independent

The bootstrap is valid only for i.i.d. samples. (Incidentally, this is also true for the parametric estimate.) If the set $\mathbf{X}$ of match scores is obtained by matching only one pair of impressions for each finger, clearly the data are independent. If more than two impression are taken from each available finger, which is often the case, the data are dependent.

If multiple scores per finger are available, the data can be shown to be weakly dependent, because when properly arranged, the data set is *n-dependent* [6]. To account for weak dependence, we use a "moving block" bootstrap instead of the bootstrap described below. Here bootstrap sets are generated in the form of pseudo-series and the arising statistics are pseudo-replications [6]. This particular moving blocks bootstrap is only valid for samples of matching scores. This bootstrap technique and appropriate bootstrap techniques for samples of mismatch scores will be described in a forthcoming paper.

3 Database characterization

To compare and contrast performance figures, we need to understand the data set characteristics. Here standard public data sets will help in a big way. However, one cannot dictate researchers and vendors to use a single database. Hence, we propose several measures to characterize test data sets. For speaker verification, Doddington et al. [2] present a data set characterization in terms of four classes which are called sheeps, goats, lambs and wolves. For fingerprint authentication, we propose similar qualitative measures. This is an attempt to characterize fingerprint test sets in terms of "difficulty," and to characterize such sets for different kinds of populations.

The test set of fingerprints, used in this paper, were gathered from 40 subjects drawn from a sample of 160 volunteers from a large campus population. The volunteers comprised mostly of adults (< 2% were above age 50) and both sexes were well represented (102 male, 58 female). Each volunteer contributed four fingers (left and right index and middle fingers) in two sessions separated by a period of six weeks. In the first session, two impression of each finger were collected. The remaining two impressions per finger were collected in the second session. All prints were collected using the same optical fingerprint scanner with 512 dpi image resolution.

The total set consists of two sets of 320 prints, taken six weeks apart. The totals number of fingers is 160. We have four prints from each finger.

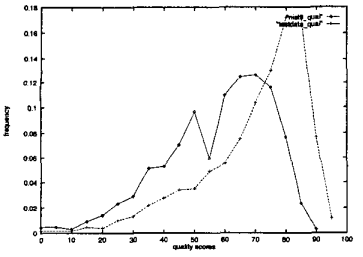

Figure 1. Image quality plots.

3.1 Image quality histograms

We use an in-house fingerprint image-quality measure. (For the purposes of this discussion, the algorithm to compute this quality is not important.) Although methods for automatic quality assessment of a fingerprint image are highly subjective, the average quality of fingerprint images in data sets are a practical index for comparing data sets. Still a better method is to compute the histogram of image qualities for a test data set and compare it with the histogram of the qualities of a standard public database. One can use several measures to compute the distance between histogram of image qualities. For this purpose, half the above test set was used: four sample prints of two fingers of 40 subjects. We plot the quality of 600 NIST-9 database fingerprint images used as the reference data set. The histograms of the two data sets (i.e., NIST-9 and the test data set) are shown in Fig. 1. The multimodal histogram corresponds to the NIST-9 data set. The effect of two peaks is probably due to the fact that the NIST fingerprint images are digitized impressions on paper.

An alternate here is to plot histograms of matching scores of the test data set and the reference data set.

3.2 Mu-sigma plots

There are several other methods of characterizing the test set. For example, if we have $i = 1, ..., N$ subjects and $j = 1, ..., M$ fingerprints per subject, we can generate $m \leq M(M-1)/2$ matching scores $S_i = \{s_{i1}, s_{i2}, ..., s_{im}\}$ per individual. Now, for each subject i, we compute estimates of the mean μ and σ, $\hat{\mu}_i^*$ and $\hat{\sigma}_i^*$, respectively. Additionally, confidence intervals $[\hat{\mu}_i^-, \hat{\mu}_i^+]$ and $[\hat{\sigma}_i^-, \hat{\sigma}_i^+]$ could be established.

A graphical analysis of $\hat{\mu}$ versus $\hat{\sigma}$, or $\hat{\mu}$ versus the $\hat{\mu}$-confidence intervals or versus the $\hat{\sigma}$-confidence intervals will give some indication of the quality characteristics of the individual fingerprint images. In Fig. 2, we show a plot of $\hat{\mu}$ versus $\hat{\sigma}$.

An ideal matcher will generate perfect scores with zero standard deviation for all matching pairs of the same finger. In practice, poor quality matching pairs result in smaller scores and a significant spread in the matching scores. The spread itself may be due to poor imaging conditions and/or impression to impression variation. A statistic based on

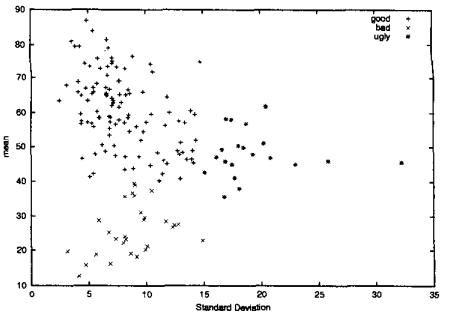

Figure 2. Test score subset mean vs. standard deviation plot.

μ and σ (e.g., $\sum \frac{\mu}{\sigma}$) could characterize the quality of the database. Using this plot, we can characterize the database into several more or less separated clusters: (i) 'good' (ii) 'bad' and (iii) 'ugly.' Figure 2 references these as '+' good, '$\times$' bad and '$\star$' ugly, 111 fingers belong to the good class, 29 to the bad class and 20 to the ugly class.

The 'good' class lies in the upper-left corner of the plot, the area of small standard deviation and small mean of the match scores. The 'bad' class in the lower-left has small standard deviation but small mean of the scores. The points in the center belong to the 'ugly' class, the mean is average and there is a large spread. Visual inspection to discover the properties of the prints to explain the existence of these three classes is labor-intensive. However, we speculate that the existence of 'ugly' class is the result of poor reproduction of the initial prints after six weeks. Habituation should reduce this problems.

4 Authentication accuracy characterization

In this section we look at the accuracy of FRR estimates. Confidence intervals are determined using both parametric and (moving blocks) bootstrap techniques. Further, it is shown that the subsets of match scores in (Fig. 2) result in match score distributions are quite different. Finally, we determine the validity of the measure d' which is often used and advocated measure of fingerprint matcher accuracy.

The number of times the set of match scores in the experiments below are sampled with replacement $B = 1{,}000$ times. The confidence interval are always 90%; that is, with 90% certainty, the estimated value of the probability distribution or probability mass distribution is correct.

4.1 Bootstrap vs. parametric confidence intervals

This section uses all the match scores of 160 all-against-all matches of the four prints from each finger for generating confidence intervals. The all-against-all matching resulted in 12 matched pair scores per finger because the matcher is nonsymmetric, with a total number of match scores of

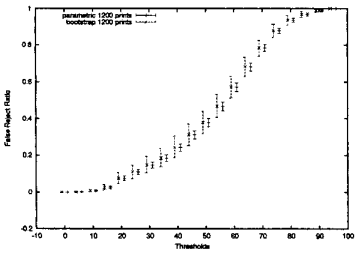

Figure 3. Score distribution and confidence intervals for all fingerprints.

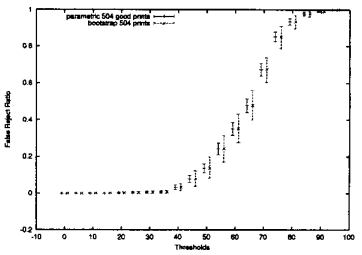

Figure 4. Score distribution and confidence intervals for good fingerprints.

1,920. Figure 3 shows two distributions of these match scores with confidence intervals. The distribution with the smaller confidence intervals are parametric estimates as described in Sect. 2.1. That is, at any x, $\hat{F}(x)$ is assumed to be distributed according to a binomial distribution and to establish a confidence interval, it is further assumed that a Gaussian distribution is appropriate to compute the confidence intervals. The distribution with the larger intervals is computed with a moving block bootstrap with $B = 1{,}000$. That is, at any x the confidence interval for $\hat{F}$ in Fig. 3 is established by generating a thousand sample-and-replace bootstrap sets, and computing a thousand estimates $\hat{F}$. These estimates are sorted, and by counting top and bottom 5% the confidence interval is determined.

The confidence intervals, computed with the bootstrap are *larger* than those computed with a parametric estimator. The reason for this is that more than two impressions of each finger are used and hence there is dependence in the data set of scores. Let i_a, i_b and i_c be three images of the same finger, then, obviously, the matches $m(i_a, i_b)$, $m(i_a, i_c)$ and $m(i_b, i_c)$ are dependent random variables.

The confidence interval of an estimate of linear functions of random variables is related to the variance of the random variable. Dependence of the random variables has a large influence on the variance.

Let X and Y be independent, $X \equiv Z$, $\mathrm{E}(X) = \mathrm{E}(Y) = \mathrm{E}(Z) = 0$ and $\mathrm{E}(X^2) = \mathrm{E}(Y^2)$. Then:

$$\mathrm{E}[(X+Y)^2] = 2\mathrm{E}(X^2)$$
$$\mathrm{E}[(X+Z)^2] = 4\mathrm{E}(X^2)$$

From this, it is seen that the sum of dependent random variables, X and Z, has larger variance than the sum of independent random variables X and Y. Remember that the estimates $\hat{F}$ are obtained as a weighted sum of the match scores X_i. These match scores are dependent and hence the variance of $\hat{F}$ is increased.

Very incidentally, confidence intervals using bootstrap estimates for independent random variables amount to confidence intervals that are about the same length as confidence intervals estimated using the parametric model. The difference, however, is that bootstrap confidence intervals are not symmetric around the empirical estimate at x, $\hat{F}(x)$. For score sets that are obtained from fingerprint image databases that only contain two impressions for each finger, traditional bootstrap techniques for confidence interval estimation can be used. If during fingerprint collection more than two prints are acquired from each finger, it is recommended that more than four prints per finger are taken, otherwise the moving block bootstrap is not valid. Six prints per finger is perhaps best, then 16 mated pairs are available per finger. Using parametric estimation techniques in such cases are dubious at least and should be used with care.

4.2 FRR distributions for subsets of the data

Figure 4 shows the distribution of the scores of matching pairs of fingers from the 'good' class. This class corresponds to the cluster with high mean and low variance match scores, indicated by '+' in Fig. 2. The number of fingers in this class is 111. So with 12 matches per finger, we have 1,332 match scores of mated pairs. Again the 90% confidence intervals using the parametric estimator are larger than the 90% confidence intervals computed using bootstrap estimator (using $B = 1{,}000$).

This distribution is better behaved than the one in Fig. 3 in the sense that the distribution is closer to a step function at some match value. After all, what one would really like to see is $m(i_a, i_b) \geq Th$ if i_a and i_b are impressions of the same finger, and $m(i_a, i_b) < Th$ if this is not the case. The only probability distribution that satisfies such a requirement is a step function at some T. In fact, the probability mass distribution corresponding to the set of match scores of the good prints (shown in Fig. 7) seams to be converging to a Gaussian distribution.

Note that the bootstrap confidence intervals in Figs. 3 and 4 are about the same. The lower variances in the class 'good' compensate for the fact that fewer samples are available for confidence interval estimation.

Figure 5 shows the probability distributions with confidence intervals of the classes 'bad' and 'ugly' from Fig. 2. (The parametric estimates of confidence intervals are omitted.) These are probability distributions that are every bio-

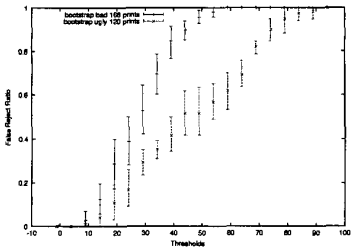

Figure 5. Score distribution and confidence intervals for bad and ugly fingerprints.

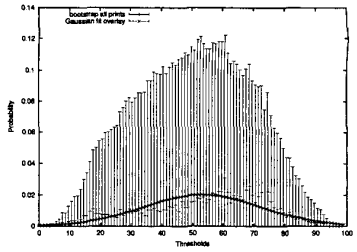

Figure 6. The pdf and confidence intervals for all match scores with Gaussian fit.

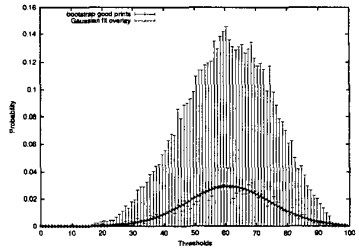

Figure 7. The pdf and confidence intervals for the scores of class 'good' with Gaussian fit.

metrics practitioner nightmare. The probability distribution of mismatched pairs of fingerprints will be a shifted-to-the-left version of the distribution in Fig. 4 but there is not much "room" for such a distribution. Hence, the tradeoff between FRR and FAR will be difficult for populations that generate scores as in the bottom and right of scatter plot Fig. 2. Of course, the FRR-FAR tradeoff depends very much on the particular biometrics application and the users will habituate to the application. It may be that certain subjects that remain in the 'ugly' or 'bad' class will simply have to be handled as exceptions.

The lower of the two curves in Fig. 5 corresponds to the class 'bad,' while the upper curve corresponds to the class 'ugly.'

4.3 The d-prime characteristic

Quite often a measure of "goodness" d' (d-prime) of a matcher is defined as [10]:

$$d' = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)}}. \quad (2)$$

This measure was originally developed to measure the separability of two equivalent normal distributions. Here μ_1 and σ_1 are the mean and variance of the match scores of genuine users and μ_2 and σ_2 the mean and variance of the match scores of mismatching finger prints. This measure is large when μ_1 is large and μ_2 small, i.e., high match scores for genuine users and low match scores otherwise. So, if d' is large, a matcher is claimed to be very accurate,

Here, implicitly, match scores and mismatch scores are assumed to be normally distributed. That is, match scores s_g are distributed as $s_g \sim \mathcal{N}(\mu_1, \sigma_1)$ and mismatch scores s_i are distributed as $s_i \sim \mathcal{N}(\mu_2, \sigma_2)$.

Figure 6 shows the probability mass distribution of all the match scores of 160 all-against-all matches of the four prints from each finger. This resulted in 12 scores per finger because the matcher is nonsymmetric. Hence, the total number of match scores is 1,920. The confidence intervals are also shown in Fig. 6 with the bootstrap number, B = 1,000. Note that the confidence intervals of this probability mass distribution are significantly larger than the confidence intervals of the corresponding graph of the probability distributions in Fig. 3. The reason, of course, is that the probability mass distribution is the derivative of the probability distribution which effectively doubles the confidence intervals.

A Gaussian curve, with mean and variance of the total number of match scores is overlaid on the graph. As can be seen, the probability mass distribution for the entire collection of prints is not very well represented by a normal distribution.

On the other hand, Fig. 7 shows the probability mass distribution of the matching scores of the 'good' fingers. The number of good fingers is 111, so the number of scores is 1,332. Again, the number of bootstrap sets is 1,000.

Apart from some outliers, the Gaussian curve in this figure appears to fit the probability mass distribution of the scores pretty well. So, good fingerprints as defined by the mu-sigma plot of Fig. 2, the Gaussian distribution assumption to justify d', may not be an unreasonable one.

However, it is best to look at this issue a little more objectively. A way toward more objectively determining if a random variable is distributed according to a Gaussian probability distribution is to compute the skew. The skew of a distribution, $-1 \leq S \leq 1$, measures how symmetric a probability distribution is around its mean. This is a necessary condition for a distribution to be Gaussian. With M samples from a distribution, an estimate of the skew is defined

	M	$\bar{X}$	σ^2	S_M
all	1,920	52.7	19.6	-0.40
good	1,332	60.5	13.8	-0.23
bad	348	26.0	11.7	0.40
ugly	240	48.1	20.8	0.05

Table 1. Skew measures of different data subsets.

as

$$S_M = \frac{1}{(M-2)\sigma^3} \sum_{1=1}^{M} (X_i - \bar{X})^3$$

It is the average of $[(X_i - \bar{X})/\sigma]^3$, the variables normalized to a distribution with zero mean and variance 1 and then raised to the power 3.

Table 1 gives the skew measures of the clusters in the scatter plot of Fig. 2. The skew of the class 'all' is negative, which means that samples $X_i < \bar{X}$ have higher scatter than the samples that are greater than the mean value. Surprisingly, the set 'good' is quite skewed too, which is not immediately obvious from the print of the probability mass distribution (Fig. 7). The skew of all the fingerprints is almost twice (-0.40) as the skew of the good fingerprints (-0.23), though..

The d' as defined by Expression (2) provides some idea of the degree of separability of the impostor and genuine pair match score distributions. If one insists on using d' as an accuracy measure, the skew of both the FRR and FAR should be studied and a confidence intervals should be given of d-prime. Using bootstrap measures described in this paper, estimates $\hat{\mu}$ and $\hat{\sigma}^2$ and confidence intervals are easily established.

5 Discussion and conclusions

A first issue that is addressed in this paper is test data set characterization. Here we look both at measures that objectively quantify the quality of fingerprint images. This allows for comparing databases of test images based on quality histograms. A second method we use is to compute statistics of the matching scores of the mated pairs in a data set. Graphically displaying view distributions of these statistics gives insights into the complexity of the test sets. It also allows for grouping the subjects into classes of difficulty.

The second issue that we explore is confidence intervals of FRR estimates. For FRR estimates, we always have fewer samples than the mismatched pairs, hence, the FRR confidence intervals are wider than for FAR estimates. It is therefore that, in this paper, we are concerned with confidence intervals of the FRR estimates.

Bootstrap techniques give powerful ways to develop confidence intervals. The beauty of these techniques is that they are completely nonparametric. That is, one makes no assumptions about the underlying forms of distribution functions. However, sets of match values are generally dependent. Using [6, 9] we develop a "moving blocks" bootstrap for sets of match scores. This bootstrap lends itself particularly well for computing confidence intervals for the FRR estimates. An important result is that both the parametric method and the traditional bootstrap for computing confidence intervals result in intervals that are smaller than the moving blocks bootstrap.

References

[1] W. Chen, M. Surette, and R. Khanna. Evaluation of automated biometrics-based identification and verification systems. *Proceedings of the IEEE*, 85(9):1464–1478, Septemeber 1997.

[2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in th e NIST 1998 speaker recognition evaluation. In *Proc. of ICSLD'98*, Sidney, Australia, November 1998.

[3] B. Germain et al. Issues in large scale automatic biometric identification. In *IEEE Workshop on Automatic Identification Advanced Technologies*, pages 43–46, Stony Brook, NY, November 1996.

[4] W.W. Peterson, T.G. Birdsall, and W.C. Fox. The theory of sigmal detectability. *Transactions of the IRE*, PGIT-4:171–212, April 1954.

[5] P. J. Phillips. On performance statistics for biomettrics systems. In *Proc. of AutoID'99*, pages 111–116, Summit, NJ, USA, October 1999.

[6] D.M. Politis. Computer-intensive methods in statistical analysis. *IEEE Signal Processing*, 15(1):39–55, January 1998.

[7] R. M. Bolle, N. K. Ratha and S. Pankanti. Evaluating authentication systems using bootstrap confidence intervals. In *Proc. of AutoID'99*, pages 9–13, Summit, NJ, USA, October 1999.

[8] J.L. Wayman. A scientific approach to evaluating biometric systems using mathematical methodology. In *CardTechSecureTech*, pages 477–492, Orlando, FL, May 1997.

[9] A.M. Zoubir and B. Boashash. The boostrap and its application in signal processing. *IEEE Signal Processing*, 15(1):56–76, January 1998.

[10] J. G. Daugman and G. O. Williams, A Proposed Standard For Biometric Decidability, *Proc. CardTech/SecureTech Conference*, Atlanta, GA, pp. 223-234, 1996.