

NISTIR 7495

**Operational Measures and Accuracies
of ROC Curve
on Large Fingerprint Data Sets**

Jin Chu Wu

NIST

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

NISTIR 7495

Operational Measures and Accuracies of ROC Curve on Large Fingerprint Data Sets

Jin Chu Wu

National Institute of Standards and Technology
Gaithersburg, MD 20899

May 2008



U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
James M. Turner, Deputy Director

Operational Measures and Accuracies of ROC Curve on Large Fingerprint Data Sets

Jin Chu Wu^{*}

Image Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

At any point on a receiver operating characteristic (ROC) curve in combination with two distributions of genuine scores and impostor scores, there are three related variables: the true accept rate (TAR) of the genuine scores, the false accept rate (FAR) of the impostor scores, and the threshold. Any one of these three variables determines the other two variables. The measures and accuracies of TAR and threshold while FAR is specified and the measures and accuracies of TAR and FAR once threshold is fixed are all investigated. In addition, the measures and accuracies of the equal error rate (EER) and the corresponding threshold are also explored. From the operational perspective, these exhaust all possible scenarios. The nonparametric two-sample bootstrap based on our previous studies of bootstrap variability on large fingerprint data sets is employed to compute the measurement accuracies. Four high-accuracy and two low-accuracy fingerprint-image matching algorithms invoking different types of scoring systems are taken as examples.

Keywords: Receiver operating characteristic (ROC) curve; Fingerprint matching; Nonparametric two-sample bootstrap; Standard errors; Confidence interval; Equal error rate

^{*} Tel: + 301-975-6996; fax: + 301-975-5287. E-mail address: jinchu.wu@nist.gov.

1. Introduction

A receiver operating characteristic (ROC) curve can be used to evaluate the performances of algorithms in many biometric applications and especially in the applications of analyzing fingerprint data on large data sets. Genuine scores are generated by comparing two different fingerprint images of the same subject, and impostor scores are created by matching two fingerprint images of two different subjects. Both scores may be referred to as similarity scores in this article. These two sets of similarity scores constitute two distributions, respectively, as schematically depicted in Figure 1 (A) for continuous similarity scores. The cumulative probabilities of genuine and impostor scores from the highest similarity score to a specified similarity score (i.e., threshold) are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. Thus, in the FAR-and-TAR coordinate system, while the threshold moves from the highest similarity score down to the lowest similarity score, an ROC curve is constructed as drawn in Figure 1 (B).

An ROC curve is characterized by the relative relationship between these two distributions [1, 2]. As extensively explored in our previous studies [1], it was revealed that 1) there is usually no underlying parametric distribution function for genuine scores and impostor scores, 2) the distribution of genuine scores and the distribution of impostor scores are considerably different in general, and 3) the distributions vary substantially from algorithm to algorithm in a way that differentiates algorithms in terms of qualities. This suggests that the nonparametric analysis be pertinent to evaluating fingerprint-image matching algorithms on large-scale data sets.

An ROC curve can be measured by invoking the area under ROC curve [1, and references therein]. However, from the operational perspective in the fingerprint applications, at any point on an ROC curve in combination with two distributions of genuine scores and impostor scores, the three variables, which are FAR, TAR, and threshold, are related to each other, as illustrated in Figure 1 (A) and (B). Any one of these three variables can determine the other two variables. Then, questions arise. How accurate are the measures of TAR and threshold, if the FAR is specified [3]? And how accurate are the measures of TAR and FAR, while the threshold is fixed? In practice, it is never required that TAR be specified in the first place.

In addition, the equal error rate (EER) is defined to be $1 - \text{TAR}$ (i.e., the type I error) or FAR (i.e., the type II error) when they are equal. As is well-known, these two errors are traded-off of each other. The EER can be used as a metric to evaluate the performance of fingerprint-image matching algorithms. Generally speaking, the smaller the EER is, the more apart the two distributions of genuine scores and impostor scores are, thus the higher the ROC curve is and the more accurate the fingerprint-image matching algorithm is [1, 2]. Then, how accurate are the measures of EER and the corresponding threshold? The above three scenarios exhaust all possible cases from the operational perspective.

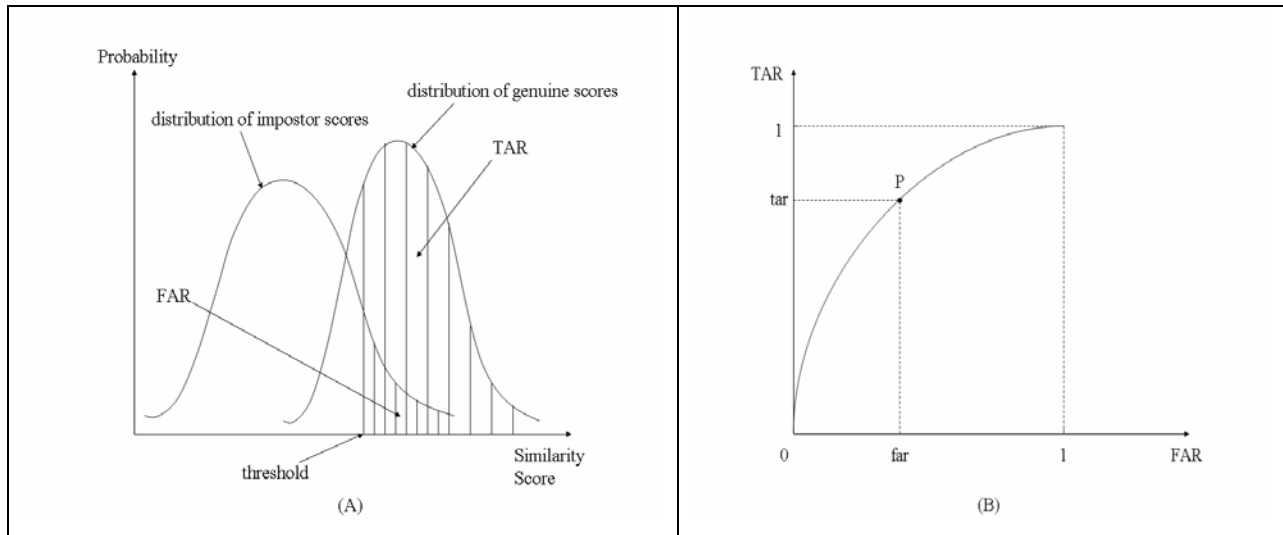


Figure 1 (A): A schematic diagram of distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. (B): A schematic drawing of an ROC curve constructed by moving the threshold from the highest similarity score down to the lowest similarity score. Any point P on an ROC curve has two coordinates FAR and TAR and is associated with threshold through two distributions.

Different fingerprint-image matching algorithms employ different scoring systems, such as integers, or real numbers, or less-than-one real numbers with different numbers of decimal places. As studied before [1, 4, 5], different scoring systems can be converted to integer scores, if they are not. Certainly, at the end of the calculation, all thresholds must be converted back using the original scoring system.

All similarity scores are dealt with as integer scores, and thus the probability distribution functions of similarity scores are all discrete and an ROC curve is no longer a smooth curve [1]. Since there is usually no underlying parametric distribution function for similarity scores, the empirical distribution is assumed for each of the observed scores. As opposed to continuous distribution, some concepts and definitions need to be established and modified accordingly. For instance, first, the ties of genuine scores and/or impostor scores at a threshold can often occur on large fingerprint data sets and thus must be taken into account while computing the estimated TAR at a specified FAR. Second, while computing the cumulative discrete probability at a score, the probability at this score must be taken into account [6]. Third, it seems that generally speaking there does not exist such a similarity score (range) at which the type I error is exactly equal to the type II error.

In the medical applications, many studies were done in some respects [7-10], but their data sets are usually far smaller than those in the fingerprint applications. In our previous studies [5], first, the bootstrap variability was extensively investigated empirically on large fingerprint data sets, and thereafter the number of two-sample bootstrap replications in the fingerprint applications was determined to be 2000. Second, the measurement accuracy of TAR while FAR is specified was computed using the nonparametric two-sample bootstrap [11-15]. The two samples are a set of genuine scores and a set of impostor scores. Their corresponding distributions are interrelated with each other by the fingerprint-image matching algorithm that generates them. All statistics of interest are influenced under the combined impact of these two samples.

In this article, the same nonparametric two-sample bootstrap using 2000 bootstrap replications will be employed to compute the measurement accuracies in the above three scenarios from the operational perspective. In the case where the threshold is fixed, the application of two-sample bootstrap is reduced to the application of one-sample bootstrap, as dealt with in Ref. [16]. This is because TAR and FAR can be determined by the threshold individually from the technical viewpoint. In our applications, the total number of genuine scores is a little over 60 000 and the total number of impostor scores is a little over 120 000 [4].

To be self-contained, the formulation of discrete distribution functions of genuine and impostor scores, and ROC curve is presented in Section 2. And also a detection error trade-off (DET) curve is briefly described in Section 2. The methods of computing the measures and accuracies of TAR and threshold if FAR is specified, TAR and FAR if threshold is fixed, as well as the ERR and the corresponding threshold, are presented in Section 3. The nonparametric two-sample bootstrap algorithms in these three scenarios are also provided in Section 3. Four high-accuracy and two low-accuracy fingerprint-image matching algorithms¹ invoking different types of scoring systems are taken as examples. The results are shown in Section 4. Finally, conclusion and discussion can be found in Section 5.

2. The formulation of discrete distribution functions of similarity scores and ROC curve

As indicated in Section 1, without loss of generality, the similarity scores are expressed inclusively using the integer score set $\{s\} = \{s_{\min}, s_{\min}+1, \dots, s_{\max}\}$, running consecutively from the minimum score s_{\min} to the maximum score s_{\max} . The genuine score set and the impostor score set are denoted as

$$\mathbf{G} = \{ m_i \mid m_i \in \{s\} \text{ and } i = 1, \dots, N_G \}, \quad (1)$$

and

$$\mathbf{I} = \{ n_i \mid n_i \in \{s\} \text{ and } i = 1, \dots, N_I \}, \quad (2)$$

where N_G and N_I are the total numbers of genuine scores and impostor scores, respectively. Note that similarity scores may not exhaust all members in the integer score set $\{s\}$ and some image comparisons may very well share the same integer score. Thus, the genuine score set and the impostor score set can be partitioned into pairwise-disjoint subsets, respectively.

Let $P_i(s)$, where $s \in \{s\}$ and $i \in \{G, I\}$, denote the empirical probabilities of the genuine scores and the impostor scores at a score s , respectively. To deal with the whole spectrum of the similarity scores by including zero frequencies, the discrete probability distribution functions of genuine scores and impostor scores can be expressed, respectively, as

¹ Specific hardware and software products identified in this report were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

$$P_i = \{ P_i (s) \mid s = s_{\min}, s_{\min}+1, \dots, s_{\max} \text{ and } \sum_{\tau=s_{\min}}^{s_{\max}} P_i (\tau) = 1 \}, i \in \{G, I\}. \quad (3)$$

And the cumulative discrete probability distribution functions of genuine scores and impostor scores are defined in this article to be the probabilities cumulated from the highest score s_{\max} down to the integer score s , and are expressed as

$$C_i = \{ C_i (s) = \sum_{\tau=s}^{s_{\max}} P_i (\tau) \mid s = s_{\max}, s_{\max}-1, \dots, s_{\min} \}, i \in \{G, I\}, \quad (4)$$

where $C_i (s)$, $i \in \{G, I\}$, are the cumulative probabilities of genuine scores and impostor scores, i.e., the TAR and FAR, respectively.

It is assumed that an ROC curve explored in this article is formed using the trapezoidal rule, from which the Mann-Whitney statistic of the two distributions can be derived and the Z statistic can be applied to the area under an ROC curve for significance test [1, and references therein]. Hence, an ROC curve is a curve connecting $s_{\max} - s_{\min} + 1$ points $\{ (C_I (s), C_G (s)) \mid s = s_{\max}, s_{\max}-1, \dots, s_{\min} \}$ using line segment in the FAR-and-TAR coordinate system, and extending to the origin of the coordinate system. Overlap of points $(C_I (s), C_G (s))$ can occur, while both $P_I (s)$ and $P_G (s)$ are zero. An ROC curve goes horizontally, vertically, or inclined upper-rightwards at a score s , depending on whether only $P_I (s)$ is nonzero, or only $P_G (s)$ is nonzero, or both of them are nonzero, respectively.

A DET curve will be invoked while discussing the EER. As opposed to an ROC curve, $1 - \text{TAR}$ instead of TAR is used as the vertical axis to generate a DET curve. If the discrete probability distribution functions of similarity scores are encountered and if the type I error is cumulated from the lowest similarity score, a DET curve may have extra points in comparison with what the corresponding ROC curve has. These extra points may be created, for instance, by similarity scores with zero probability. In addition, the sum of the type I error and the TAR at a score s is equal to $1 + P_G (s)$, which is no longer equal to 1 if $P_G (s)$ is greater than zero. As a result, unlike the cases in which the continuous distributions are dealt with, the DET curve is no longer to have a reflection relationship with the corresponding ROC curve under the transformation from TAR to $1 - \text{TAR}$.

3. Methods

3.1 The Estimated TAR and Threshold at a Specified FAR

It is assumed that all similarity scores are converted to integer scores. It was shown in our previous studies [5] that given a FAR = f where $0 < f < 1$, without loss of generality, the threshold t is defined to satisfy

$$C_I(t+1) < f \text{ and } C_I(t) \geq f, \quad (5)$$

where both t and $(t+1) \in \{s\}$. Hence, $P_I(t) = C_I(t) - C_I(t+1) > 0$. It indicates that the probability of impostor scores at the threshold t is always positive in the circumstance where FAR is specified in the first place.

It was proven using ROC curve that the estimated TAR at a specified FAR = f is given by [5]

$$\hat{TAR}(f) = C_G(t+1) + P_G(t) * \frac{f - C_I(t+1)}{P_I(t)}. \quad (6)$$

In other words, if $P_G(t)$ is not equal to zero, the ratio of $(\hat{TAR}(f) - C_G(t+1))$ to $P_G(t)$ must be equal to the ratio of $(f - C_I(t+1))$ to $P_I(t)$; otherwise, the TAR is the same as $C_G(t+1)$. And the ratio is always in $(0, 1]$ because of Eq. (5). This formula takes into account the ties of genuine scores and impostor scores. The ties of similarity scores not only can often occur but also can be large, in the practice of testing and evaluating different fingerprint-image matching algorithms, which generates large sizes of genuine scores and impostor scores. Neglecting the impact of the ties can cause error of evaluation.

3.2 The Estimated TAR and FAR at a Specified Threshold

The distributions of similarity scores dealt in this article are all discrete. The cumulative discrete probability at any legitimate score is the probability cumulated from the highest similarity score down to this score [6]. The estimated TAR and FAR at a specified threshold t , that might not be a legitimate score, are,

$$\begin{aligned} \hat{TAR}(t) &= C_G(s) \\ \hat{FAR}(t) &= C_I(s) \end{aligned} \quad \text{for } t \in (s-1, s] \text{ and } s-1 \& s \in \{s\}. \quad (7)$$

When the original scoring system invokes integers, the input threshold t may very well not be a legitimate score. In such a case, the estimated TAR and FAR at the threshold t are the cumulative discrete probabilities of genuine scores and impostor scores, respectively, from the highest similarity score down to the legitimate integer score that is the ceiling of the input threshold t .

3.3 The Estimated EER and the Corresponding Threshold

As indicated in Section 1, for discrete distributions of genuine and impostor scores, the definition of EER needs to be established. According to the convention, it is assumed that $1 - \text{TAR}$, i.e., the type I error denoted by ER_I , is the probability of genuine scores cumulated from the lowest similarity score, and FAR, i.e., the type II error denoted by ER_II , is the probability of impostor scores cumulated from the highest similarity score. Thus, at a similarity scores $s \in \{s\}$, their estimators are expressed as, respectively,

$$\begin{aligned} \hat{\text{ER}}_{\text{I}}(s) &= 1 - C_G(s + 1) \\ \hat{\text{ER}}_{\text{II}}(s) &= C_I(s) \end{aligned} \quad \text{for } s \in \{s\}, \quad (8)$$

in which $C_G(s_{\max} + 1) = 0$ is assumed. Eq. (8) implies that the probabilities of genuine scores and impostor scores at the score s are all taken into account for discrete distributions [6].

As the similarity score s runs from the highest score s_{\max} down to the lowest score s_{\min} , the estimated type I error $\hat{\text{ER}}_{\text{I}}(s)$ decreases from 1 to $P_G(s_{\min})$, but the estimated type II error $\hat{\text{ER}}_{\text{II}}(s)$ increases from $P_I(s_{\max})$ to 1. Both of them are step functions. Hence, the absolute difference $|\hat{\text{ER}}_{\text{I}}(s) - \hat{\text{ER}}_{\text{II}}(s)|$ decreases first, and then increases after reaching its minimum. It seems that for discrete distributions the minimum can rarely reach zero.

Once $|\hat{\text{ER}}_{\text{I}}(s) - \hat{\text{ER}}_{\text{II}}(s)|$ is at its minimum if and only if the score s is in the range $[s_1, s_2]$, the estimated EER is defined to be

$$\hat{\text{EER}} = \frac{\hat{\text{ER}}_{\text{I}}(s) + \hat{\text{ER}}_{\text{II}}(s)}{2}, \text{ iff } s \in [s_1, s_2]. \quad (9)$$

In general, all similarity scores in the range $[s_1, s_2]$ may correspond to the same point on an ROC curve and the same point on a DET curve. The point on a DET curve may be one of extra points

with respect to the corresponding ROC curve as indicated in Section 2. Nonetheless, these similarity scores have equal weight for determining the EER. Thus, the corresponding threshold is simply defined to be

$$T\hat{H}S = \left\lfloor \frac{S_1 + S_2}{2} \right\rfloor. \quad (10)$$

3.4 Measurement Accuracies

The measurement accuracies of the estimated statistics of interest in three different scenarios as discussed in Subsection 3.1 through 3.3 are explored using the nonparametric two-sample bootstrap [11-15] based on our previous empirical studies of the bootstrap variability on large fingerprint data sets [5]. The algorithm is as follows.

Algorithm

- 1: **for** $i = 1$ **to** B **do**
- 2: select N_G scores WR from {original N_G genuine scores} \Rightarrow {new N_G genuine scores} $_i$
- 3: select N_I scores WR from {original N_I impostor scores} \Rightarrow {new N_I impostor scores} $_i$
- 4: {new N_G genuine scores} $_i$ & {new N_I impostor scores} $_i \Rightarrow$ statistics $S\hat{T}_i^k, k = 1, 2$
- 5: **end for**
- 6: { $S\hat{T}_i^k | i = 1, \dots, B$ } $\Rightarrow S\hat{E}_B^k$ and $(\hat{Q}_B^k(\alpha/2), \hat{Q}_B^k(1 - \alpha/2))$, while $k = 1, 2$
- 7: **end**

where B is the number of two-sample bootstrap replications and WR stands for “with replacement”. A set of original N_G genuine scores and a set of original N_I impostor scores are generated by a fingerprint-image matching algorithm. As shown from Step 1 to 5, this algorithm runs B times. In the i -th iteration, N_G scores are selected WR from the original set of N_G genuine scores to form a new set of N_G genuine scores, N_I scores are selected WR from the original set of N_I impostor scores to form a new set of N_I impostor scores, and then from these two new sets of similarity scores the i -th bootstrap replications of two estimated statistics of interest, i.e., $S\hat{T}_i^k, k = 1, 2$, are generated, respectively.

More specifically, while $FAR = f$ is specified, $S\hat{T}_i^1$ stands for the i -th bootstrap replication of the estimated $T\hat{A}R(f)$ derived using Eq. (6) and $S\hat{T}_i^2$ represents the i -th bootstrap replication of the estimated threshold obtained using Eq. (5). If the threshold t is fixed, $S\hat{T}_i^1$ is the i -th replication of the estimated $T\hat{A}R(t)$ and $S\hat{T}_i^2$ is the i -th replication of the estimated $F\hat{A}R(t)$ derived using Eq. (7). Once the EER is dealt with, $S\hat{T}_i^1$ is the i -th replication of the estimated $E\hat{E}R$ obtained using Eq. (9) and $S\hat{T}_i^2$ is the i -th replication of the estimated threshold derived using Eq. (10).

Finally, as indicated in Step 6, from the two sets $\{S\hat{T}_i^k | i=1, \dots, B\}$, while $k=1, 2$, respectively, the estimators of the standard errors (SE), i.e., the unbiased standard deviations $S\hat{E}_B^k$, and the estimators of the $\alpha/2$ 100 % and $(1 - \alpha/2)$ 100 % quantiles of the bootstrap distributions $\hat{Q}_B^k(\alpha/2)$ and $\hat{Q}_B^k(1-\alpha/2)$ at the significance level α can be calculated. The Definition 2 of quantile in Ref. [17] is adopted. That is, the sample quantile is obtained by inverting the empirical distribution function with averaging at discontinuities. Thus, $(\hat{Q}_B^k(\alpha/2), \hat{Q}_B^k(1-\alpha/2))$ stand for the estimated bootstrap $(1 - \alpha)$ 100 % confidence intervals (CI). If 95 % confidence interval is of interest, then α is set to be 0.05.

In the scenarios as described in Subsection 3.1 and 3.3, at the end of the computation, the threshold must be converted back using the original scoring system. If the original scoring system invokes integers, to be more conservative, the lower bound and upper bound of 95 % confidence interval of the threshold are rounded to integers using floor and ceiling functions, respectively. In addition, as it will be shown in Section 4, the distribution of bootstrap replications of threshold is quite asymmetric. Thus, to describe the measurement accuracy of threshold, it is better to use 95 % confidence interval than standard error. As indicated in Section 1, the number of the nonparametric two-sample bootstrap replications B is set to be 2000, and N_G is a litter over 60 000 and N_I is a little over 120 000.

4. Results

Six fingerprint-image matching algorithms were taken to be examples.² Among them, Algorithms 1 to 4 were of high accuracy and Algorithms 5 and 6 were of low accuracy. These algorithms were using different types of scoring systems. Algorithms 1 to 3 all employed integers but in different ranges of integers. Algorithm 4 invoked less-than-100 real numbers. And Algorithms 5 and 6 used less-than-1 real numbers but with different numbers of decimal places.

4.1 Measures and Accuracies of TAR and Threshold at a Specified FAR

If FAR is specified, the estimated $\hat{TAR}(f)$ and threshold can be computed using Eqs. (6) and (5), respectively. The FAR was set to be 0.001 [2, 4, 5]. In Table 1 are shown the estimates of TARs, standard errors (SE), and 95 % CIs for six fingerprint-image matching algorithms. As indicated in Subsection 3.4, the 95 % CIs were calculated using the Definition 2 of quantile in Ref. [17]. However, generally speaking, they do match the 95 % CIs up to the fourth decimal place for higher-accuracy algorithms and the third decimal place for lower-accuracy algorithms, which were calculated if the distribution of 2000 bootstrap replications of the statistic $\hat{TAR}(f)$ for each algorithm is assumed to be normal. For example, for high-accuracy Algorithm 1, the 95 % CI as shown in Table 1 is (0.992622, 0.993922), and the 95 % CI assuming normal distribution is (0.992618, 0.993892) using the estimated \hat{SE} 0.000325. It is also found that the higher the accuracy of the algorithm is, the smaller the standard error is, and thus the narrower the 95 % CI is. These observations are consistent with those in Ref. [1, 4, 5].

Algorithm	$\hat{TAR}(f)$	\hat{SE}	95 % Confidence interval
1	0.993255	0.000325	(0.992622, 0.993922)
2	0.994322	0.000324	(0.993662, 0.994918)
3	0.989263	0.000470	(0.988307, 0.990159)
4	0.972307	0.000913	(0.970592, 0.974176)
5	0.929970	0.001522	(0.926869, 0.932899)
6	0.796753	0.003503	(0.789545, 0.803961)

Table 1 The estimates of TARs, standard errors (SE), and 95 % confidence intervals for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6, respectively, while FAR was specified at 0.001.

² The algorithms are proprietary. Hence, they cannot be disclosed.

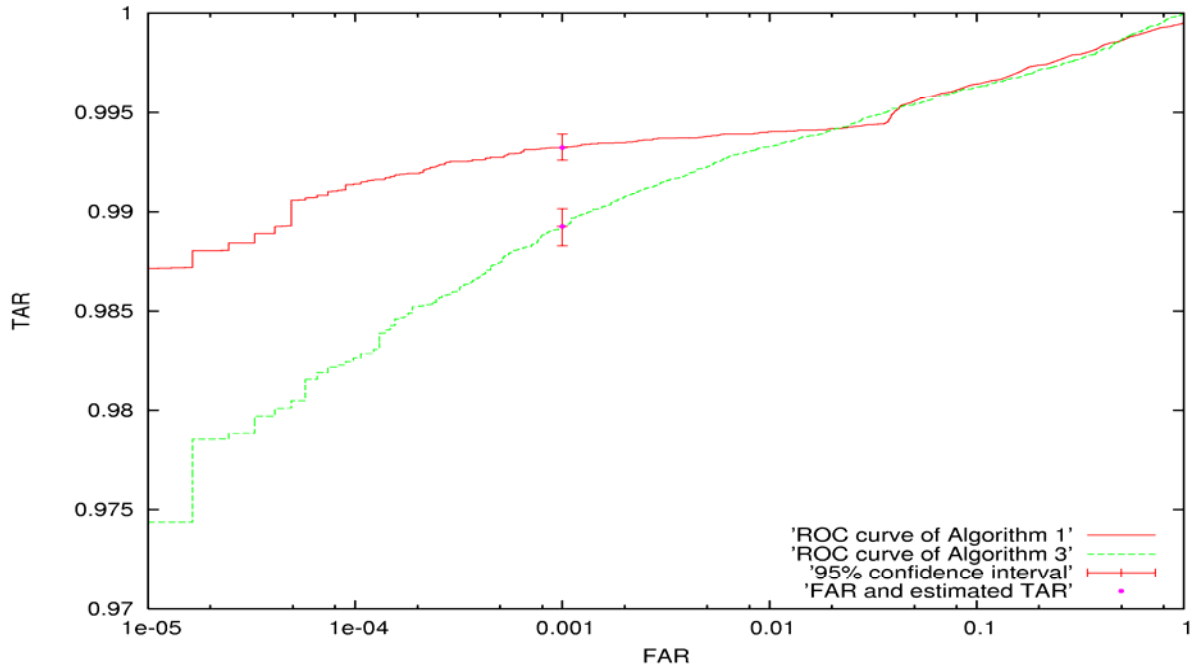


Figure 2 ROC curves of Algorithms 1 and 3 along with their 95 % confidence intervals of the estimated $\hat{TAR}(f)$ if the FAR is specified at 0.001.

To illustrate it further, the ROC curves of Algorithms 1 and 3 along with their 95 % CIs of the estimated $\hat{TAR}(f)$ if the FAR is specified to be 0.001 are drawn in Figure 2. Not only is the estimated $\hat{TAR}(f)$ of Algorithm 1 while $FAR = 0.001$ larger than the one of Algorithm 3, but the 95 % CI of Algorithm 1 is also completely above the one of Algorithm 3. However, as shown in Table 1, the 95 % CI of Algorithm 1 does overlap the one of Algorithm 2. The relationship between two 95 % CIs could be used to examine the difference of performances of two algorithms to some extent. Nevertheless, the issue of determining whether the difference between the performances of two algorithms is real or by chance must be dealt with using the significance test.

Another issue is that if FAR changes, the 95 % CI will move along an ROC curve accordingly. However, the FAR cannot be changed too small. As indicated in Section 1, the total number of impostor scores is a little over 120 000. Therefore, if FAR is set to be 0.001 or 0.0001, then the number of failures related to the type II error is about 120 or 12. It seems that the number of 12 is too small to have statistical significance. In order to have smaller FAR but maintain the statistical

significance of test, the total number of impostor scores must increase accordingly [4]. On the other hand, it is obvious that the FAR cannot be set too large, such as 0.01, etc.

In Table 2 are presented the estimates of thresholds $\hat{T}HS(f)$ and their 95 % CIs for six fingerprint-image matching algorithms, while FAR was specified at 0.001. First, the representations of the estimates of thresholds and their 95 % CIs are different from algorithm to algorithm because different algorithms employed different scoring systems. Second, if the input FAR changes, all these quantities will change accordingly. Third, it is observed that the 95 % CIs of the estimated thresholds are quite asymmetric. Hence, the measurement accuracy of the estimated $\hat{T}HS(f)$ is expressed only using the 95 % CI.

Algorithm	$\hat{T}HS(f)$	95 % Confidence interval
1	455	(442, 463)
2	34	(30, 39)
3	4534	(4450, 4616)
4	40.8472	(40.2195, 41.3513)
5	0.1053	(0.1011, 0.1096)
6	0.634030	(0.626579, 0.640724)

Table 2 The estimates of thresholds $\hat{T}HS(f)$ and 95 % confidence intervals for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6, respectively, while FAR was specified at 0.001.

4.2 Measures and Accuracies of TAR and FAR at a Specified Threshold

In Table 3 are shown the estimates of TARs and FARs along with their estimated standard errors (SE) and 95 % CIs for six fingerprint-image matching algorithms, while the threshold t is specified. To show the operational significance, for each algorithm, the estimated threshold obtained while FAR was set to be 0.001 in Subsection 4.1 was chosen to be the specified threshold. Certainly, the input threshold can vary.

The 95 % CIs calculated using the Definition 2 of quantile in Ref. [17], as shown in Table 3, do match the 95 % CIs mostly up to the fourth decimal place for both TARs and FARs, which were

calculated if the distributions of 2000 bootstrap replications of the statistics $\hat{TAR}(t)$ and $\hat{FAR}(t)$, respectively, are assumed to be normal. For instance, for high-accuracy Algorithm 1, the 95 % CI of the estimated $\hat{FAR}(t)$ as shown in Table 3 is (0.000820, 0.001184) and the 95 % CI assuming normal distribution is (0.000830, 0.001186) using the estimated \hat{SE} 0.000091.

Algorithm	Threshold t	$\hat{TAR}(t)$	\hat{SE}	95 % Confidence interval
		$\hat{FAR}(t)$		
1	455	0.993255	0.000337	(0.992605, 0.993905)
		0.001008	0.000091	(0.000820, 0.001184)
2	34	0.994393	0.000300	(0.993792, 0.994994)
		0.001057	0.000093	(0.000877, 0.001246)
3	4534	0.989274	0.000408	(0.988486, 0.990094)
		0.001016	0.000093	(0.000836, 0.001213)
4	40.8472	0.972307	0.000676	(0.970966, 0.973672)
		0.001000	0.000091	(0.000828, 0.001180)
5	0.1053	0.929970	0.001030	(0.927971, 0.931994)
		0.001000	0.000089	(0.000836, 0.001180)
6	0.634030	0.796753	0.001590	(0.793641, 0.799792)
		0.001000	0.000092	(0.000836, 0.001189)

Table 3 The estimates of TARs and FARs along with their estimated standard errors (SE) and 95 % confidence intervals for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6, respectively, while the threshold t is specified for each algorithm. The specified threshold t is the estimated threshold obtained while FAR was set to be 0.001 in Subsection 4.1.

More importantly, for each algorithm, since the specified threshold employed here is the estimated threshold obtained while FAR f was set to be 0.001 in Subsection 4.1, the estimated statistic of interest $\hat{TAR}(t)$ calculated here falls in the 95 % CI computed in Subsection 4.1 using the nonparametric two-sample bootstrap. For instance, for Algorithm 1, the estimated statistic of interest $\hat{TAR}(t)$ is 0.993255 while the threshold is set to be 455 as shown in Table 3 falls within the 95 % CI (0.992622, 0.993922) while $FAR = 0.001$ as shown in Table 1. Reversely, for each algorithm, the estimated statistic of interest $\hat{TAR}(f)$ calculated in Subsection 4.1 also falls in the 95 % CI computed here using the nonparametric two-sample bootstrap.

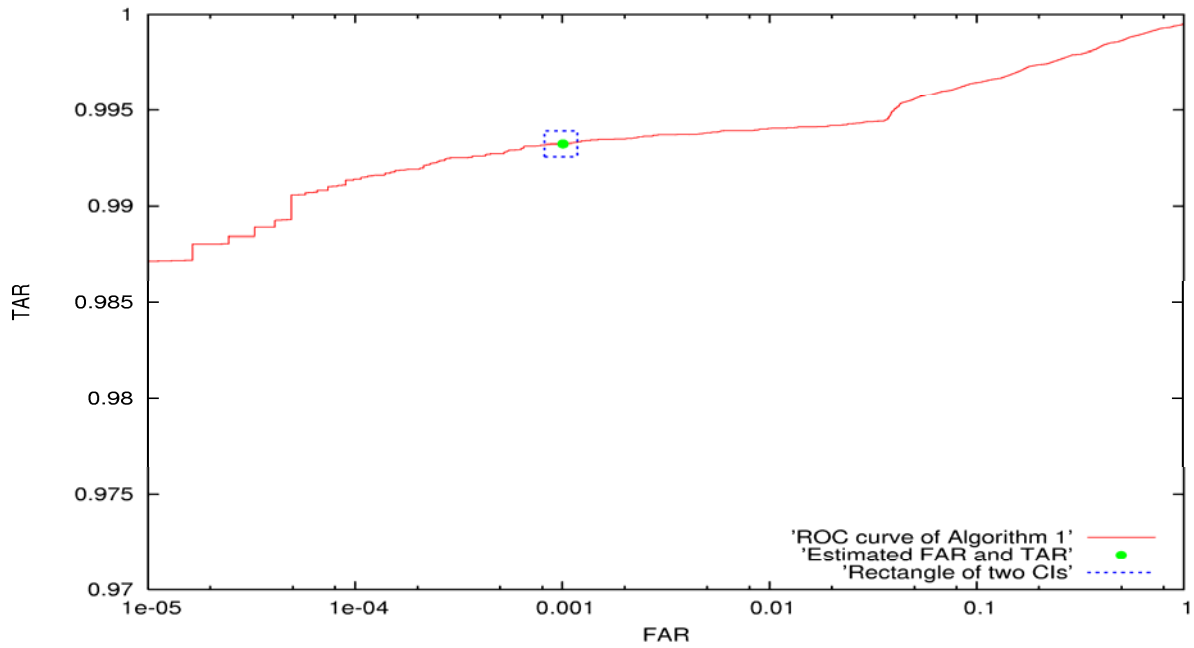


Figure 3 The upper part of the ROC curve of Algorithm 1. The FAR and TAR coordinates of the dot are the estimated $\hat{FAR}(t)$ and $\hat{TAR}(t)$ while the threshold t is specified at 455. The rectangle is formed by two 95 % confidence intervals of FAR and TAR.

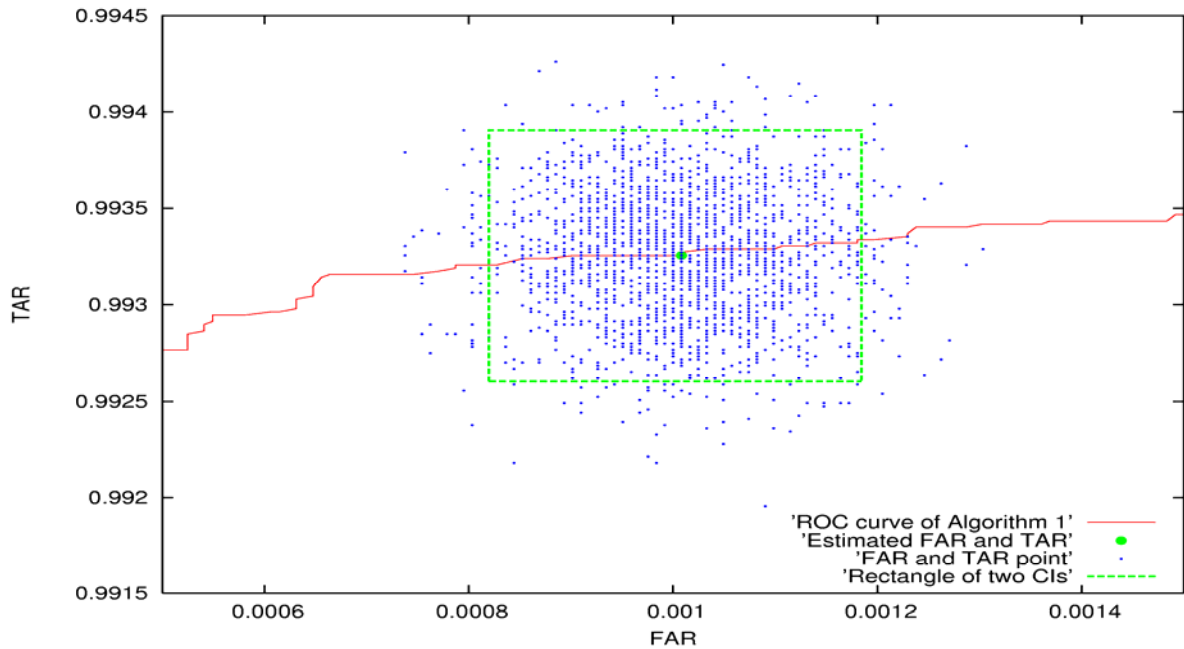


Figure 4 An enlargement of the portion in Figure 3, corresponding to the specified threshold 455. Drawn are 2000 points formed by uncorrelated FAR and TAR that are paired by the bootstrap replications.

Moreover, the 95 % CIs of the estimated $\hat{TAR}(t)$ in Table 3 are equivalent to the 95 % CIs of the estimated $\hat{TAR}(f)$ in Table 1, respectively, especially for high-accuracy fingerprint-image matching algorithms. For example, for high-accuracy Algorithm 1, the 95 % CI of the estimated $\hat{TAR}(t)$, that is (0.992605, 0.993905), matches the 95 % CI of the estimated $\hat{TAR}(f)$, which is (0.992622, 0.993922), up to the fourth decimal place at the both bounds of the interval.

Regarding FAR, all 95 % CIs as shown in Table 3 are mutually compatible as well as the fixed FAR $f = 0.001$ employed in Subsection 4.1 falls into all of the 95 % CIs as listed in Table 3. All these observations indicate that the computation using the nonparametric two-sample bootstrap with 2000 bootstrap replications is quite self-consistent.

To illustrate the measurement accuracies, as an example, the upper part of the ROC curve of Algorithm 1 is depicted in Figure 3. In this figure, the dot on the ROC curve represents the estimated $\hat{FAR}(t)$ and $\hat{TAR}(t)$ while the threshold t is specified to be 455, which are 0.001008 and 0.993255 as shown in Table 3. As a matter of fact, the 95 % CI of the estimated $\hat{FAR}(t)$, i.e., (0.000820, 0.001184), and the 95 % CI of the estimated $\hat{TAR}(t)$, i.e., (0.992605, 0.993905), as presented in Table 3, can constitute a rectangle around the estimated $\hat{FAR}(t)$ and $\hat{TAR}(t)$ as shown in the figure. If the threshold changes, the rectangle can move along the ROC curve.

An enlargement of this portion is shown in Figure 4. In addition, drawn are 2000 FAR-and-TAR points paired by the i -th bootstrap replication of the estimated $\hat{FAR}(t)$, i.e., \hat{ST}_i^2 , with the i -th bootstrap replication of the estimated $\hat{TAR}(t)$, i.e., \hat{ST}_i^1 , while the threshold t is fixed, as shown in Subsection 3.4. In deed, the bootstrap replications of the FAR are not correlated with the bootstrap replications of the TAR when the threshold is specified in the first place. If the rule of forming pairs changes, the distribution of the 2000 points will change accordingly.

Further, in Figure 4, it is obvious that for FAR, besides some points falling on the boundary, less than 50 points are located outside the left boundary and outside the right boundary of the rectangle, respectively. And the same situation occurs for TAR. Thus, by no means, this rectangle is a 95 % confidence rectangle. The rectangle only shows the bounds of the two 95 %

CIs. In order to have 95 % of FAR-TAR points falling inside a rectangle as well as on the boundary, the rectangle may be constructed using two 97.5 % CIs.

4.3 Measures and Accuracies of EER and the Corresponding Threshold

Besides statistical error, the measurement accuracy of EER also includes systematic error stemming from the discreteness of the distributions of genuine scores and impostor scores realized by the definition of EER, i.e., Eq. (9). Here the systematic error is expressed in terms of the relative error that is equal to half of the minimum of the absolute difference $|\hat{E}R_I(s) - \hat{E}R_{II}(s)|$ divided by the estimated $\hat{E}ER$ derived from Eq. (9). The relative errors of six algorithms are shown in Table 4. It seems that the minima of the absolute difference are not large. However, the relative errors can reach as high as 0.51 %. This happens even for high-accuracy Algorithm 1. Another issue is that whether the minimum occurs and only occurs at a similarity score s or within a score range $[s_1, s_2]$ depends on algorithms. Nonetheless, among the six algorithms, only Algorithm 6 has the situation in which a score range rather than a single score appears, but the score range is formed just by two scores.

Algorithm	Score (range)	Min ($ \hat{E}R_I(s) - \hat{E}R_{II}(s) $)	$\hat{E}ER$	Relative Error
1	346	0.000061	0.006064	0.51 %
2	12	0.000013	0.004346	0.15 %
3	3754	0.000013	0.007076	0.09 %
4	31.3983	0.000002	0.012903	0.01 %
5	0.0343	0.000089	0.036400	0.12 %
6	[0.510836, 0.510837]	0.000003	0.068650	0.00 %

Table 4 The relative errors of all six algorithms. The relative error is equal to half of Min ($|\hat{E}R_I(s) - \hat{E}R_{II}(s)|$) divided by the estimated $\hat{E}ER$ occurring and only occurring at the score (range).

In Table 5 are presented the estimates of EERs along with their estimated standard errors (SE) and 95 % CIs for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6,

respectively. As expected, it shows that generally speaking the higher the accuracy of the algorithm is, the smaller the estimated \hat{EER} is. This is because the two distributions of genuine scores and impostor scores are more apart and thus the ROC curve is higher [1, 2]. And also the 95 % CIs are narrower.

Algorithm	\hat{EER}	\hat{SE}	95 % Confidence interval
1	0.006064	0.000301	(0.005511, 0.006703)
2	0.004346	0.000202	(0.003905, 0.004713)
3	0.007076	0.000295	(0.006496, 0.007655)
4	0.012903	0.000360	(0.012205, 0.013609)
5	0.036400	0.000650	(0.035096, 0.037624)
6	0.068650	0.000743	(0.067174, 0.070162)

Table 5 The estimates of EER, standard errors (SE), and 95 % confidence intervals for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6, respectively.

In addition, the 95 % CIs computed using the Definition 2 of quantile in Ref. [17], as shown in Table 5, do match the 95 % CIs, mostly up to the fourth decimal place for high-accuracy fingerprint-image matching algorithms and the third decimal place for low-accuracy algorithms, which were calculated if the distributions of 2000 bootstrap replications of the statistic EER are assumed to be normal. For example, for high-accuracy Algorithm 1, the 95 % CI of the estimated \hat{EER} as shown in Table 5 is (0.005511, 0.006703) and the 95 % CI assuming normal distribution is (0.005474, 0.006654) using the estimated \hat{SE} 0.000301.

The point corresponding to the EER can be drawn on an ROC curve or a DET curve, as stated in Subsection 3.3. Indeed, the EER is taken from the equality of the type I error and the type II error. In other words, a two-variable issue is reduced to a one-variable issue. Hence, it is simpler and better to use an error bar rather than an error rectangle to show the measurement accuracy. The estimated \hat{EER} s and their 95 % CIs for six algorithms are depicted in Figure 5. In general, this figure shows the relationship of the EER and its 95 % CI with the accuracy of algorithm.

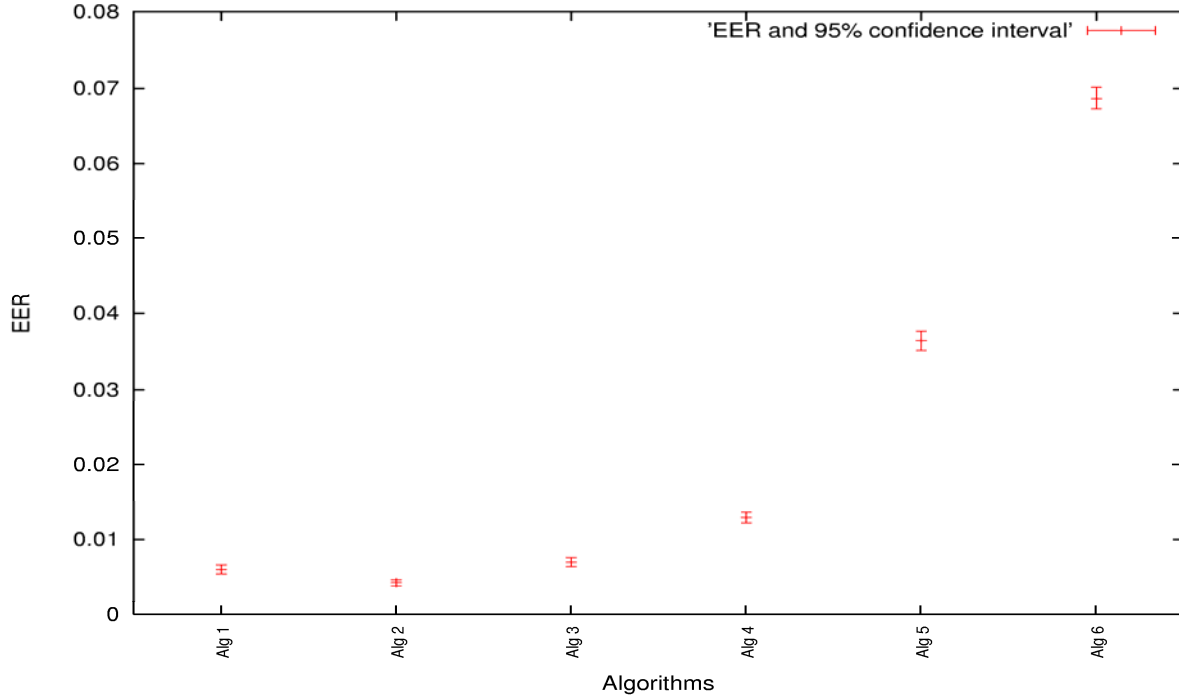


Figure 5 The estimated \hat{EER} s along with their 95 % confidence intervals of high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6.

Algorithm	\hat{THS}	95 % Confidence interval
1	346	(339, 352)
2	12	(11, 14)
3	3754	(3721, 3794)
4	31.3983	(31.1410, 31.6542)
5	0.0343	(0.0337, 0.0350)
6	0.510836	(0.509881, 0.511747)

Table 6 The estimates of thresholds THS and 95 % confidence intervals for high-accuracy Algorithms 1 to 4 and low-accuracy Algorithms 5 and 6, respectively, while EER is dealt with.

In Table 6 are shown the estimates of the thresholds along with their 95 % CIs while EER is dealt with, for all six fingerprint-image matching algorithms. Different algorithms employ different scoring systems. Thus, the representations of the estimates of thresholds and their 95 % CIs are different from algorithm to algorithm as shown in Table 6. Moreover, it is the same observation as in Subsection 4.1 that the 95 % CIs of thresholds are quite asymmetric.

5. Conclusion and Discussion

From the operational perspective, the measures and their accuracies of an ROC curve on large fingerprint data sets were studied. The measurement accuracies were investigated generally in terms of SEs and 95 % CIs using the nonparametric two-sample bootstrap. The number of bootstrap replications was set to be 2000 based on our previous empirical studies of the variability of two-sample bootstrap with respect to large fingerprint data sets [5]. In addition to dealing with the three variables TAR, FAR and threshold, the EER along with the corresponding threshold was also explored. In the case where the TAR is assumed to be specified in the first place, the same approach presented in this article can be applied. In practice from the operational perspective, the cases studied in this article exhaust all possible scenarios.

The estimated $\hat{TAR}(t)$ is not correlated with the estimated $\hat{FAR}(t)$, while the threshold t is specified. The 2000 FAR-and-TAR pairs in Figure 4 were formed according to the order of iteration occurred in the bootstrap algorithm. For these 2000 points, the bivariate normality test using the generalization of Shapiro-Wilk test for multivariate variables in R [18] revealed that the p-value was 0.1041. It indicates that the bivariate distribution of these 2000 points is quite normal. However, if the set of bootstrap replications of TAR and the set of bootstrap replications of FAR were re-paired, for instance, sorted respectively first and paired second, then these new 2000 points will distribute along a straight line, which is by no means a bivariate normal distribution. As a consequence, using ellipse instead of rectangle to describe the confidence region is inappropriate. In practice, using two 95 % CIs to describe the measurement accuracies in such a circumstance is more useful.

Due to discreteness of distributions of genuine scores and impostor scores, regarding the EER, besides statistical error, the systematic error can occur. For example, for Algorithm 1, the estimated relative error is $\frac{1}{2} * 0.000061 / 0.006064 = 0.51 \%$ as shown in Table 4. And the estimated coefficient of variation due to the statistical error as shown in Table 5 is $0.000301 / 0.006064 = 4.96 \%$. Thus, in this case the systematic error is estimated to be about 10 % of the statistical error. In comparison, all other algorithms taken as examples in this article have less

relative errors as shown in Table 4, smaller coefficients of variation because of the statistical error, and smaller ratio of the systematic error to the statistical error. Nonetheless, it must be aware that the systematic error exists while EER is employed.

In this article, one of the statistics of interest is TAR at a specified FAR. Corresponding to this statistic, in some literature [3], the false non-match rate (FNMR), which is equal to $1 - \text{TAR}$ given the same FAR, was employed. As pointed out in our previous studies [5], it is trivial to show that under the same conditions (i.e., with respect to the same series of two bootstrap samples selected with replacement from genuine scores and impostor scores, respectively, as described in the two-sample bootstrap algorithm in Subsection 3.4) the standard error of FNMR is the same as that of TAR, and the lower bound and upper bound of 95 % confidence interval for FNMR can be obtained by interchanging two bounds for TAR and subtracting them from 1, respectively. Everything else related to TAR can be transferred in parallel to FNMR.

As pointed out in our previous studies [5], using the nonparametric two-sample bootstrap to compute the measurement accuracies is a stochastic process of a Monte Carlo simulation. Therefore, unlike a deterministic process, the measurement accuracies of the statistic of interest may fluctuate every time when they are calculated by a random run of two-sample bootstrap. Nonetheless, as demonstrated in our previous studies [5], such standard error, lower bound and upper bound of 95 % confidence interval may fall into the confidence intervals with 95 % probability, which are generated by, for instance, 500 iterations of executions of two-sample bootstrap with 2000 bootstrap replications. Moreover, these confidence intervals are so narrow from the practical point of view.

The area under an ROC curve can be used to measure the performance of fingerprint-image matching algorithms as indicated in Section 1, and then the Z statistic can be used to test the significance of the difference between two ROC curves [1, and references therein]. In this article, the operational measurement accuracies in different scenarios are explored in terms of the standard errors and 95 % confidence intervals. As pointed out in Subsection 4.1, the issue of determining whether the difference between two algorithms is real or by chance must be dealt with using the significance test. The related significance test has not been investigated yet, even

though the relationship between two 95 % confidence intervals could be employed in this regard to some extent. The research of the related significance test is underway.

References

1. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, *Pattern Recognition* 40 (9) (2007) 2574-2584.
2. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in *Biometric Technology for Human Identification IV*, Proceedings of SPIE Vol. 6539, 65390N (2007).
3. R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, A.K. Jain, Performance evaluation of fingerprint verification systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (1) 2006 3-18.
4. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in *Biometric Technology for Human Identification III*, Proceedings of SPIE Vol. 6202, 620207 (2006).
5. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, September, 2007.
6. B. Ostle, L.C. Malone, *Statistics in Research: Basic Concepts and Techniques for Research Workers*, fourth ed., Iowa State University Press, Ames, 1988.
7. D. Mossman, Resampling techniques in the analysis of non-binormal ROC data, *Medical Decision Making*, 15 (4) (1995) 358-366.
8. R.W. Platt, J.A. Hanley, H. Yang, Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test, *Statistics in Medicine*, 19 (3) (2000) 313-322.
9. G. Campbell, General methodology I: Advances in statistical methodology for the evaluation of diagnostic and laboratory tests, *Statistics in Medicine*, 13 (1994) 499-508.
10. K. Jensen, H.-H. Muller, H. Schafer, Regional confidence bands for ROC curves, *Statistics in Medicine*, 19 (4) (2000) 493-509.
11. B. Efron, Bootstrap methods: Another look at the Jackknife. *Ann. Statistics*, 7:1-26, 1979.

12. P. Hall, On the number of bootstrap simulations required to construct a confidence interval, *Ann. Statist.* 14 (4) (1986) 1453-1462.
13. B. Efron, Better bootstrap confidence intervals, *J. Amer. Statist. Assoc.* 82 (397) (1987) 171-185.
14. B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993 pp. 271-282.
15. A.C. Davison, D.V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, 2003.
16. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, *Guide to Biometrics*, Springer, New York, 2003 pp. 269-292.
17. R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, *American Statistician* 50 (1996) 361-365.
18. S. Jarek, The *mvnormtest* Package, Version 0.1-6, 2005, in *R: A Language and Environment for Statistical Computing*, The R Development Core Team, Version 2.2.0, 2005, at <http://www.r-project.org/>.