Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 2574–2584
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Nonparametric analysis of fingerprint data on large data sets

Jin Chu Wu*, Charles L. Wilson

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

Received 2 May 2006; received in revised form 9 November 2006; accepted 17 November 2006

Abstract

By executing different fingerprint-image matching algorithms on large data sets, it reveals that the match and non-match similarity scores have no specific underlying distribution function. Thus, it requires a nonparametric analysis for fingerprint-image matching algorithms on large data sets without any assumption about such irregularly discrete distribution functions. A precise receiver operating characteristic (ROC) curve based on the true accept rate (TAR) of the match similarity scores and the false accept rate (FAR) of the non-match similarity scores can be constructed. The area under such an ROC curve computed using the trapezoidal rule is equivalent to the Mann–Whitney statistic directly formed from the match and non-match similarity scores. Thereafter, the Z statistic formulated using the areas under ROC curves along with their variances and the correlation coefficient is applied to test the significance of the difference between two ROC curves. Four examples from the extensive testing of commercial fingerprint systems at the National Institute of Standards and Technology are provided. The nonparametric approach presented in this article can also be employed in the analysis of other large biometric data sets.

Keywords: Fingerprint matching; Nonparametric analysis; Receiver operating characteristic (ROC) curve; Mann–Whitney statistic; Significance test

1. Introduction

Recently, the National Institute of Standards and Technology (NIST) has evaluated the fingerprint-image matching algorithms from different vendors[1] [1,2], using large samples of fingerprint data from a wide range of government sources. Several types of fingerprints (such as flat, rolled, and slap fingerprint images) and the fingerprint collection methods (e.g., using live scan devices or paper fingerprint cards) are included in these data sets. In the SDK tests [2], 6000 subjects' fingerprint images are used as a probe and 6000 second fingerprint images of the same subjects are used as a gallery. The probe is matched against the gallery. The evaluations were conducted on 19 vendor's fingerprint-image matching algorithms.[2]

Comparing two different fingerprint images of the same subject who appears both in the probe and in the gallery generates match similarity score (i.e., genuine-match score). Matching two fingerprint images of two different subjects creates non-match similarity score (i.e., impostor-match score). The fingerprint-image matching algorithms tested in Ref. [1,2] are designed in such a way that the higher values of similarity scores tend to indicate that two fingerprint images are more similar. Hence, the distribution function of the match similarity scores will be centered at higher score than the distribution function of the non-match similarity scores does.

The true accept rate (TAR) and the false accept rate (FAR) are defined, respectively, as the cumulative probability of the match and non-match similarity scores from the highest similarity score to a specified similarity score, i.e., threshold, in relation to their distribution functions. Based on TAR and FAR,

* Corresponding author. Tel: + 301 975 6996; fax: +301 975 5287.
E-mail address: jinchu.wu@nist.gov (J.C. Wu).

[1] These tests were performed for the Department of Homeland Security in accordance with Section 303 of the Border Security Act, codified at 8 U.S.C. 1732. 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.

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

0031-3203/$30.00
doi:10.1016/j.patcog.2006.11.021

a receiver operating characteristic (ROC) curve can be constructed. The evaluations of fingerprint-image matching algorithms can be carried out by measuring and comparing the corresponding ROC curves.

The similarity scores generated directly from different fingerprint-image matching algorithms can be either integers or real numbers in different ranges. But real numbers with certain number of significant decimal places can be converted into integers. As a result, the similarity scores can be treated as discrete random variables rather than continuous random variables. Thus, a precise ROC curve can be built by moving the threshold one integral score at a time from the highest similarity score to the lowest similarity score for large size of data sets.

By executing different fingerprint-image matching algorithms on large data sets, it reveals that the match and non-match similarity scores for large samples have no definite underlying distribution function, and their distribution functions vary substantially from algorithm to algorithm. This suggests that a nonparametric analysis is pertinent to evaluating fingerprint-image matching algorithms on large data sets. Therefore, the evaluation of an ROC curve can be made without any assumption about such irregularly discrete distribution functions.

Evaluation of ROC curves had been studied in depth in the literature. In some approaches, the TARs at a specific FAR or within a region of FARs are chosen to be criteria [1–4]. However, in other approaches, the area under an ROC curve is invoked [4–10]. In the cited references and references therein, the studies of the area under an ROC curve were mainly focused on medical practice over small data sets. The biometric evaluations of fingerprint technology cited in the references [1,2] used large data sets, but the area under an ROC curve was not employed. Thus, the technique of using the area under an ROC curve as a criterion has never been explored for large amount of data.

The motivations behind using the area under an ROC curve as the criterion are twofold. First, the area under an ROC curve is a very important index in the analysis of ROC curves. This area is equal to the probability of correctly identifying which is more likely than the other in the two stimuli under investigation [9–11], and it measures the overall ROC curve as a whole. Second, the area under an ROC curve computed using the trapezoidal rule is equivalent to the Mann–Whitney statistic directly formed, in our case, by the match similarity scores and the non-match similarity scores [9,10,12,13].

The above second point has two consequences. First, the variance of the Mann–Whitney statistic can be utilized as the variance of the area. Second, because the Mann–Whitney statistic is asymptotically normally distributed regardless of the distributions of the match and non-match similarity scores thanks to the Central Limit Theorem [8,13,14], the Z statistic formulated in terms of areas under two ROC curves along with their variances and the correlation coefficient is subject to the standard normal distribution and can be used to test the significance of the difference of these two ROC curves.

The discrete distribution functions of the match and non-match similarity scores from the large-size fingerprint data set are explored in Section 2. Based on these distributions, a precise ROC curve is created, as discussed in Section 3. The area under such an ROC curve is studied in Section 4. Thereafter, the Z statistic is applied to the significance test of the difference between two ROC curves, as presented in Section 5. As the contents are presented, the test results of four fingerprint-image matching algorithms will be given as examples. The detailed formulas for constructing an ROC curve, computing the area under an ROC curve using the trapezoidal approach, and calculating its variance are provided in the forms that can be coded easily. Finally, conclusions are presented in Section 6.

2. The discrete distribution functions of the match and non-match similarity scores

The similarity scores are supposed to be represented in integers, as discussed earlier. Different fingerprint-image matching algorithms invoked different scoring systems. Without loss of generality, for a matching algorithm, let the integral score set be $\{s\} = \{s_{\min}, s_{\min+1}, \ldots, s_{\max}\}$, running consecutively from the minimum score $s_{\min}$ to the maximum score $s_{\max}$. To make the presentation clear, from here on, the symbol "$\forall s \in \{s\}$" indicates that s takes all integral scores from $s_{\min}$ up to $s_{\max}$ in the ascending order, and the symbol "$\forall s \in \{\bar{s}\}$" means that s takes all integral scores from $s_{\max}$ down to $s_{\min}$ in the descending order.

While executing a matching algorithm over a fingerprint-image data set, comparing two different fingerprint images of the same subject generates match similarity score. The match similarity score set is denoted as

$$\mathbf{T} = \{s_i | \forall i \in \{1, \ldots, N_T\}\}, \tag{1}$$

where N_T is the total number of match similarity scores. Here, the similarity scores s_i take values from the integral score set $\{s\}$, i.e., $s_i \in \{s\}$. But s_i may not exhaust all members in the integral score set $\{s\}$. In addition, some of the comparisons may very well share the same integral value. Therefore, the match similarity score set $\mathbf{T}$ can be partitioned into pairwise-disjoint subsets $\{\mathbf{T}_s\}$. In each subset $\mathbf{T}_s$, members have the same integral score $s \in \{s\}$. The match similarity score set $\mathbf{T}$ is the union of all these subsets $\{\mathbf{T}_s\}$.

With respect to each subset $\mathbf{T}_s$, the frequency $f_T(s)$ of the similarity score s is the size of $\mathbf{T}_s$ and the corresponding probability is $p_T(s) = f_T(s)/N_T$. Therefore, to deal with the whole spectrum of the scores by including zero frequencies, the discrete frequency and probability distribution functions of the match similarity scores can be expressed, respectively, as

$$\mathbf{F_T} = \left\{ f_T(s) | \forall s \in \{s\} \text{ and } \sum_{\tau = s\,\min}^{s\,\max} f_T(\tau) = N_T \right\}, \tag{2}$$

$$\mathbf{P_T} = \left\{ p_T(s) | \forall s \in \{s\} \text{ and } \sum_{\tau = s\,\min}^{s\,\max} p_T(\tau) = 1 \right\}. \tag{3}$$

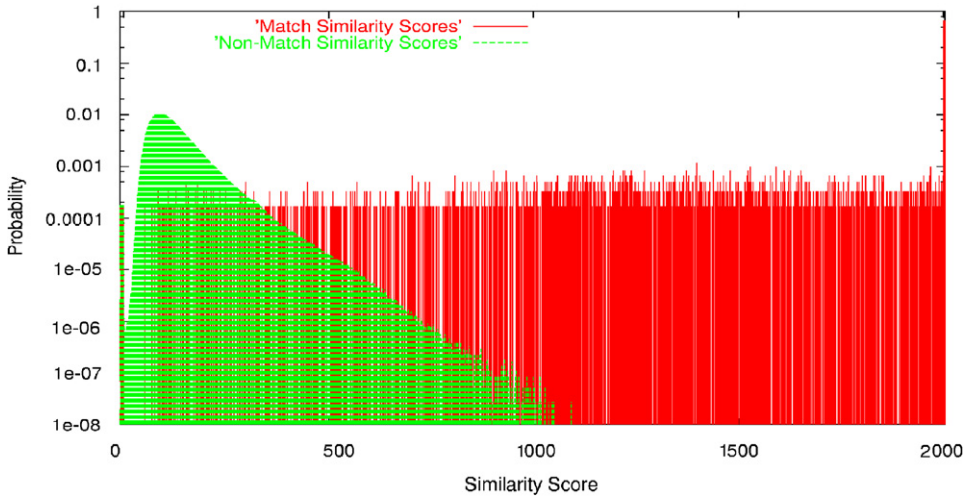

Fig. 1. The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 1. The integral similarity scores run from 0 to 2000. The widths of peaks at the highest score and at the lowest score are enlarged to show the characteristics of the distributions.

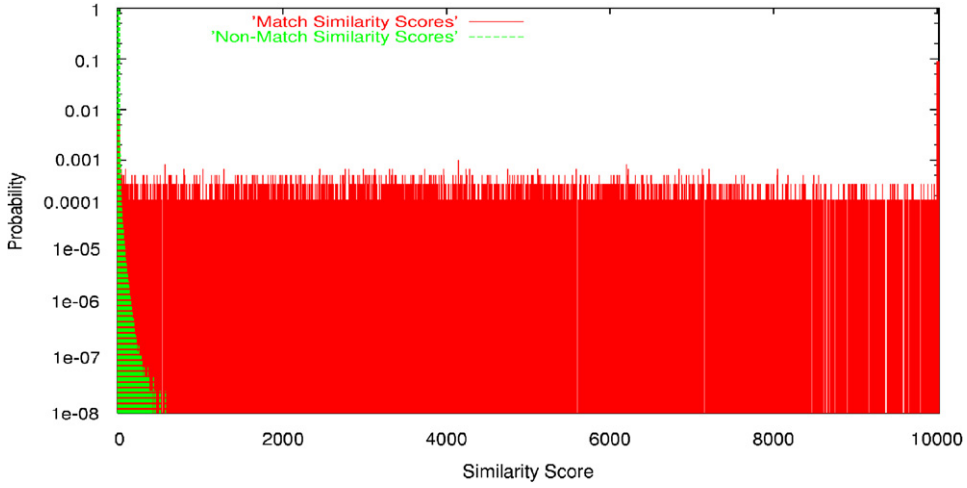

Fig. 2. The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 2. The integral similarity scores run from 0 to 9999. The widths of peaks at the highest score and at the lowest score are enlarged to show the characteristics of the distributions.

Matching two fingerprint images of two different subjects creates non-match similarity score. The non-match similarity score set is denoted as

$$\mathbf{F} = \{s_i | \forall i \in \{1, \ldots, N_F\}\} \tag{4}$$

where N_F is the total number of non-match similarity scores. By analogy with the match similarity scores, the discrete frequency and probability distribution functions of the non-match similarity scores can be formulated in terms of the frequency $f_F(s)$ and the probability $p_F(s) = f_F(s)/N_F$, respectively, as

$$\mathbf{F_F} = \left\{ f_F(s) | \forall s \in \{s\} \text{ and } \sum_{\tau=s\,\min}^{s\,\max} f_F(\tau) = N_F \right\}, \tag{5}$$

$$\mathbf{P_F} = \left\{ p_F(s) | \forall s \in \{s\} \text{ and } \sum_{\tau=s\,\min}^{s\,\max} p_F(\tau) = 1 \right\}. \tag{6}$$

By executing fingerprint-image matching Algorithms 1–4 on a large-size fingerprint database, it is found that the match and non-match similarity scores for large samples have no definite underlying distribution functions, and different algorithms have different characteristics of probability distribution functions of the match and non-match similarity scores, as demonstrated in Figs. 1–4. Thus, a nonparametric analysis must be employed in order to deal with such fingerprint data. In our studies, N_T is 6000 and N_F is 35 994 000. This means that the least probabilities of the match and non-match similarity scores are on the order of 10^{-4} and 10^{-8}, respectively. Hence, the probability is depicted in logarithmic scale.

For Algorithm 1, the match similarity scores with relatively high probabilities at higher scores have a stand-alone peak at 2000 occupying 67.52% of the whole population, and the probability distribution of the non-match similarity scores is a normal-like distribution skewed toward higher scores. For

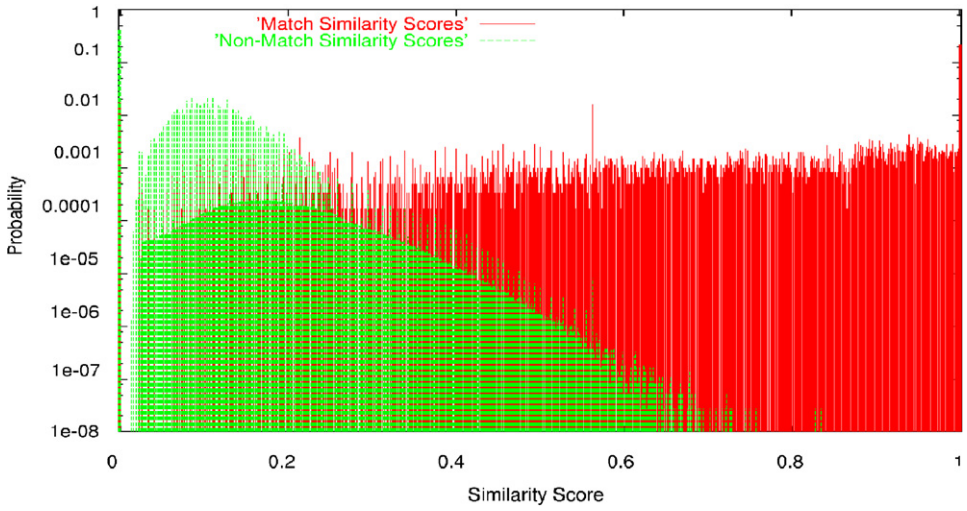

Fig. 3. The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 3. The real-number similarity scores run from 0.0 to 1.0 in five significant decimal places, which can be converted into integers. The widths of peaks at the highest score and at the lowest score are enlarged to show the characteristics of the distributions.

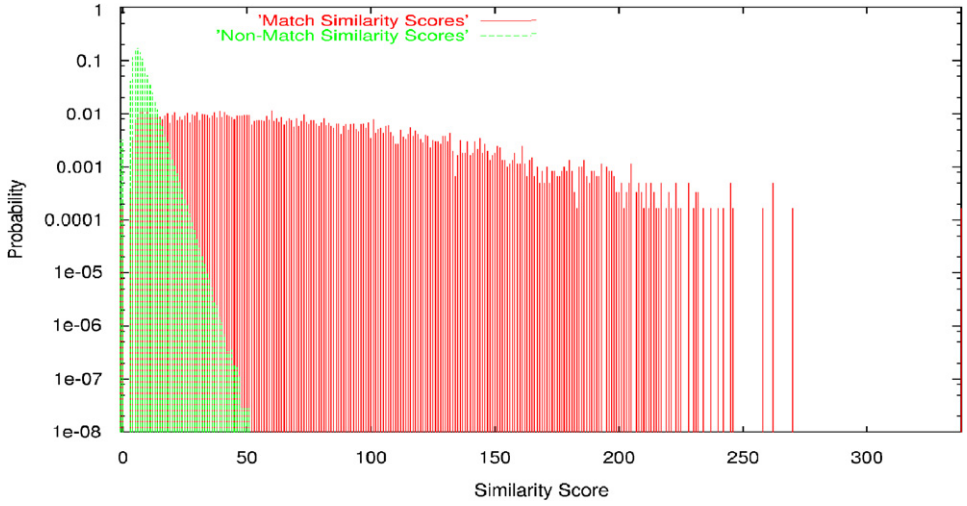

Fig. 4. The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 4. The integral similarity scores run from 0 to 338. The width of peak at the lowest score is enlarged to show the characteristics of the distributions.

Algorithm 2, the match similarity scores are almost uniformly distributed with a small peak at 9999 taking 8.98% of the whole population, but the non-match similarity scores with very steep-decay probabilities have an extremely sharp peak at zero overwhelmingly dominating 97.56% of the whole population.

Algorithms 1 and 2 behave differently in the sense that Algorithm 1 tried to push similarity scores higher and Algorithm 2, on the contrary, tended to push similarity scores lower. However, there is one thing that is common between these two algorithms. That is, both of them attempt to separate the center of the probability distribution of the non-match similarity scores from the center of the probability distribution of the match similarity scores by as wide a margin as possible. In such a way, it can produce a better ROC curve.

For Algorithm 3, the peak of the non-match similarity scores at 0.0 counts 41.33% of the population, and is separated from other scores. For Algorithm 4, the peak of the non-match similarity scores is not as high as that one, but is also separated from other scores. The distances between two centers of the probability distributions of the match and non-match similarity scores, respectively, for Algorithms 3 and 4 are not as wide as those for Algorithms 1 and 2.

3. The ROC curve of the match and non-match similarity scores

Investigating the ROC curve of the match and non-match similarity scores is a way to discover how the discrete probability distribution functions of the match and non-match similarity scores are related to each other, and thus how well/bad the fingerprint-image matching algorithm works. An ROC curve is constructed based on the cumulative discrete probability distribution functions of the match and non-match similarity scores.

From Eqs. (3) and (6) for the discrete match and non-match similarity scores, respectively, the cumulative discrete probability distribution functions can be computed by moving the threshold one integral score at a time from the highest similarity score $s_{\max}$ down to the lowest similarity score $s_{\min}$. They are expressed as

$$\mathbf{C_T} = \left\{ c_T(s) = \sum_{\tau=s}^{s\,\max} p_T(\tau) | \forall s \in \{\bar{s}\} \right\} \tag{7}$$

and

$$\mathbf{C_F} = \left\{ c_F(s) = \sum_{\tau=s}^{s\,\max} p_F(\tau) | \forall s \in \{\bar{s}\} \right\}, \tag{8}$$

where $c_T(s)$ and $c_F(s)$ are the cumulative probabilities of the match and non-match similarity scores, respectively, at the integral score s from the highest similarity score $s_{\max}$. Therefore, in the FAR-and-TAR coordinate system, an ROC curve of the match and non-match similarity scores is a curve connecting $s_{\max} - s_{\min} + 1$ points, $\{(c_F(s), c_T(s)) | \forall s \in \{\bar{s}\}\}$, and extending to the origin of the coordinate system.

The fingerprint-image matching algorithm for identifying the similarity of fingerprint images is designed in such a way that the probability distribution of the match similarity scores is centered at higher scores than the probability distribution of the non-match similarity scores. At the highest similarity score, the probability of the non-match similarity score is always zero. Thus, an ROC curve always starts from the origin of the FAR-and-TAR coordinate system, ends at the point (1, 1), and is above the straight line from the origin to (1, 1). Overlap of points $(c_F(s), c_T(s))$ can occur, while both $p_F(s)$ and $p_T(s)$ are zero. An ROC curve goes horizontally, vertically, or inclined upper rightwards at the score s, depending on whether only $p_F(s)$ is nonzero, or only $p_T(s)$ is nonzero, or both of them are nonzero, respectively.

Except at scores at which both $p_F(s)$ and $p_T(s)$ are zero, such a precise ROC curve provides the same information as that nonzero $p_F(s)$ and nonzero $p_T(s)$ provide. The precise ROC curve uniquely and accurately represents the cumulative discrete probability distribution functions of the match and non-match similarity scores. Moreover, such an ROC curve is constructed directly from the original data, after converting to integral scores if necessary, without any assumption regarding their distributions.

The ROC curves, corresponding to the four fingerprint-image matching algorithms presented in the previous section, are shown in Figs. 5 and 6. In Fig. 5 In a logarithmic scale is used for the FAR to show ROC curves at the higher-score region of the non-match similarity scores. In Fig. 6 a linear scale is used for the FAR to show ROC curves at the lower-score region of the non-match similarity scores.

For Algorithm 1, the second point on the ROC curve, i.e., one point above the origin (0, 0), is at (0, 0.6752), because the peak of the match similarity scores at the highest similarity score 2000 dominates 67.52% of the population (see Figs. 1 and 6). Then, the ROC curve does not leave the TAR coordinate axis until the highest non-match similarity score is reached. At that point, the highest non-match similarity score appears only once with probability of a little above 10^{-8}, and the cumulative probability of the match similarity scores from the highest similarity score has already reached 89.05% (see Figs. 1 and 5). After that, the ROC curve gradually reaches the point (1, 1) (see Figs. 1 and 6), because of the shape of the probability distribution of the non-match similarity scores and the relative position of two probability distributions of the match and non-match similarity scores.

In contrast, for Algorithm 2, on one hand, the ROC curve leaves the TAR coordinate axis when the cumulative probability of the match similarity scores gets to 93.47% (see Figs. 2 and 5). This is higher than 89.05% for Algorithm 1. On the other hand, it is intriguing to see that its ROC curve jumps from one side of the FAR-and-TAR coordinate system to the final point (1, 1) (see Fig. 6). This is because the probability distribution of the non-match similarity scores has a peak at the lowest similarity score zero, and it overwhelmingly occupies 97.56% of the population (see Fig. 2). Therefore, the ROC curve of Algorithm 1 starts to be higher than the ROC curve of Algorithm 2 after the FAR reaches about 20%, as evidenced by Fig. 7. This indicates that an ROC curve may be higher than the other one in a region but lower than the other one in other region.

The same qualitative analyses can be applied to the ROC curves of Algorithms 3 and 4. The ROC curve of Algorithm 3 connects many more points in the FAR-and-TAR coordinate system than the one of Algorithm 4 does, because of different scoring systems (see Figs. 3–5).

For large data sets, very little computation power is needed to create a precise ROC curve. For instance, even for a scoring system using real-number scores ranging from zero through one with five significant decimal places, the total number of integer scores is just 10^5 plus one. Thus, the total number of points in the FAR-and-TAR coordinate system, which the ROC curve needs to connect, is not very large when compared to the computing power of the current desk-top computers.

4. The area under an ROC curve

An ROC curve can be quantitatively assessed using the area under the ROC curve. The area under an ROC curve is a very important index, which represents the probability that the score obtained for the genuine match, s_G, is higher than the score assigned for the impostor match, s_I, i.e., **Prob** $(s_G > s_I)$, given both genuine match and impostor match, assuming the score is a continuous random variable [11]. Moreover, the area under an ROC curve computed using the trapezoidal rule is equivalent to the Mann–Whitney statistic [9,10,12,13], directly formed from the discrete match and non-match similarity scores in our case. Therefore, the variance of the Mann–Whitney statistic can be utilized as the variance of the area.

As shown earlier, an ROC curve can go horizontally, vertically, inclined toward upper right, or stay where it is for each increment of the two cumulative probabilities, depending on whether $p_F(s)$ and/or $p_T(s)$ are nonzero or not. Thus, the area under an ROC curve consists of a set of trapezoids, each of which is built by a rectangle and a triangle in general. But the

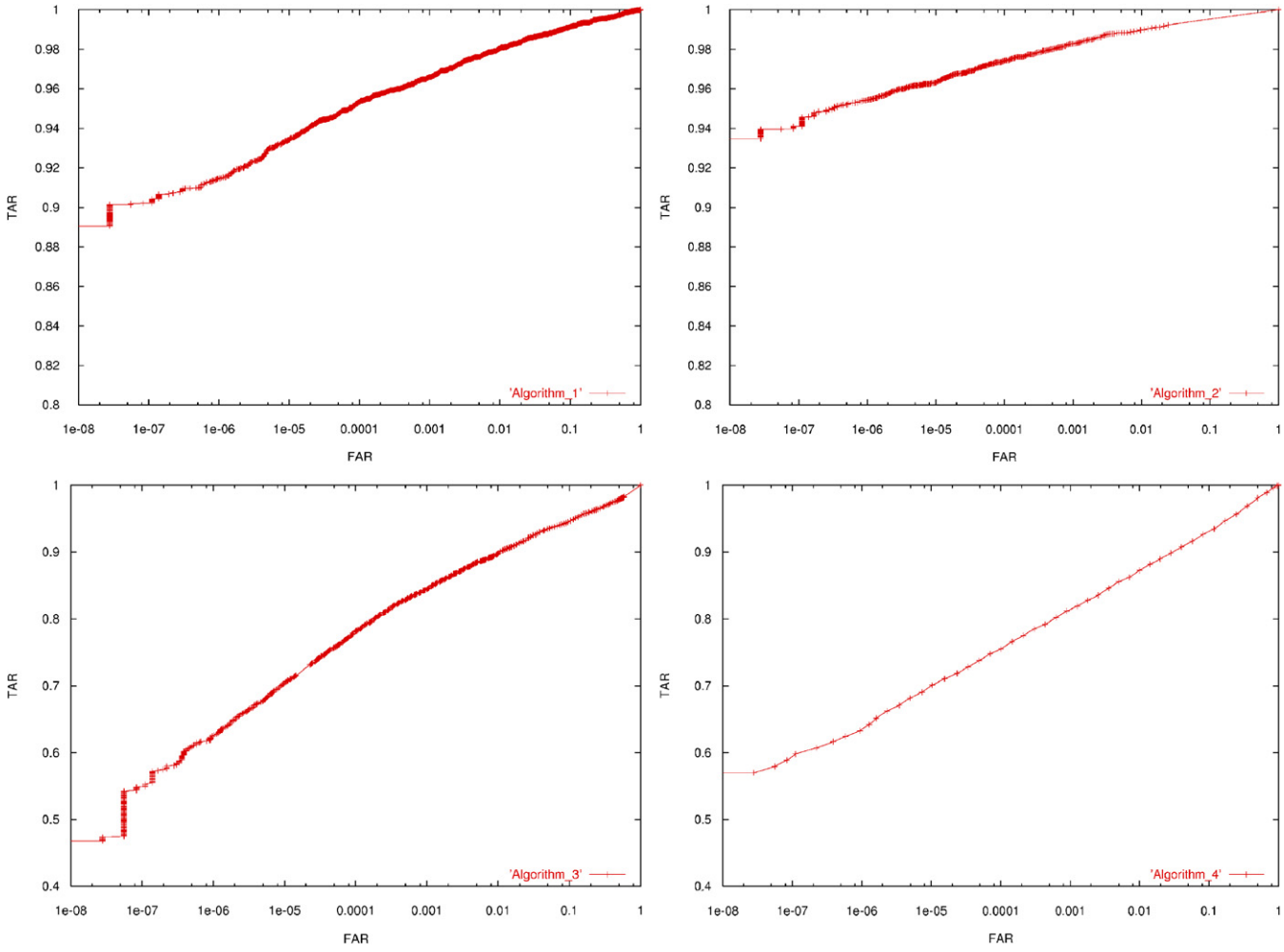

Fig. 5. The four ROC curves of Algorithms 1–4, respectively, where the FAR is in a logarithmic scale to show ROC curves at the higher-score region of the non-match similarity scores. The cross points represent the points on which the ROC curves are constructed.

trapezoid can be reduced to a rectangle, a vertical line, or a point.

As shown in Fig. 8, without loss of generality, in the FAR-and-TAR coordinate system, at the score $s \in \{\bar{s}\}$, by including zero-frequency scores, a trapezoid is constructed by four points: A $(c_F(s+1), 0)$, B $(c_F(s+1), c_T(s+1))$, C $(c_F(s), c_T(s))$, and D $(c_F(s), 0)$, in clockwise direction, assuming $c_F(s_{max}+1) = c_T(s_{max}+1) = 0$. This boundary condition corresponds to the origin of the FAR-and-TAR coordinate system, and will be applied throughout the following discussion. The lengths $(c_F(s) - c_F(s+1))$ and $(c_T(s) - c_T(s+1))$ form a triangle, and the lengths $(c_F(s) - c_F(s+1))$ and $c_T(s+1)$ create a rectangle.

From Eqs. (7) and (8), it follows that at the score $s \in \{\bar{s}\}$, the above three lengths are

$$c_F(s) - c_F(s+1) = \frac{f_F(s)}{N_F} \tag{9}$$

and

$$c_T(s) - c_T(s+1) = \frac{f_T(s)}{N_T} \tag{10}$$

and

$$c_T(s+1) = \sum_{\tau=s+1}^{s\,\max} \frac{f_T(\tau)}{N_T} \tag{11}$$

where $\sum_{\tau=s\,\max+1}^{s\,\max} = 0$ according to the above boundary condition. This notation will be applied throughout the following discussion. Hence, the area under an ROC curve can be computed as

$$\begin{aligned}\hat{\mathbf{A}} &= \sum_{s=s\,\max}^{s\,\min} \text{trapezoid}(s) \\ &= \sum_{s=s\,\max}^{s\,\min} \text{triangle}(s) + \sum_{s=s\,\max}^{s\,\min} \text{rectangle}(s) \\ &= \frac{1}{N_T N_F} * \sum_{s=s\,\max}^{s\,\min} f_F(s) * \left[\frac{1}{2} * f_T(s) + \sum_{\tau=s+1}^{s\,\max} f_T(\tau)\right].\end{aligned} \tag{12}$$

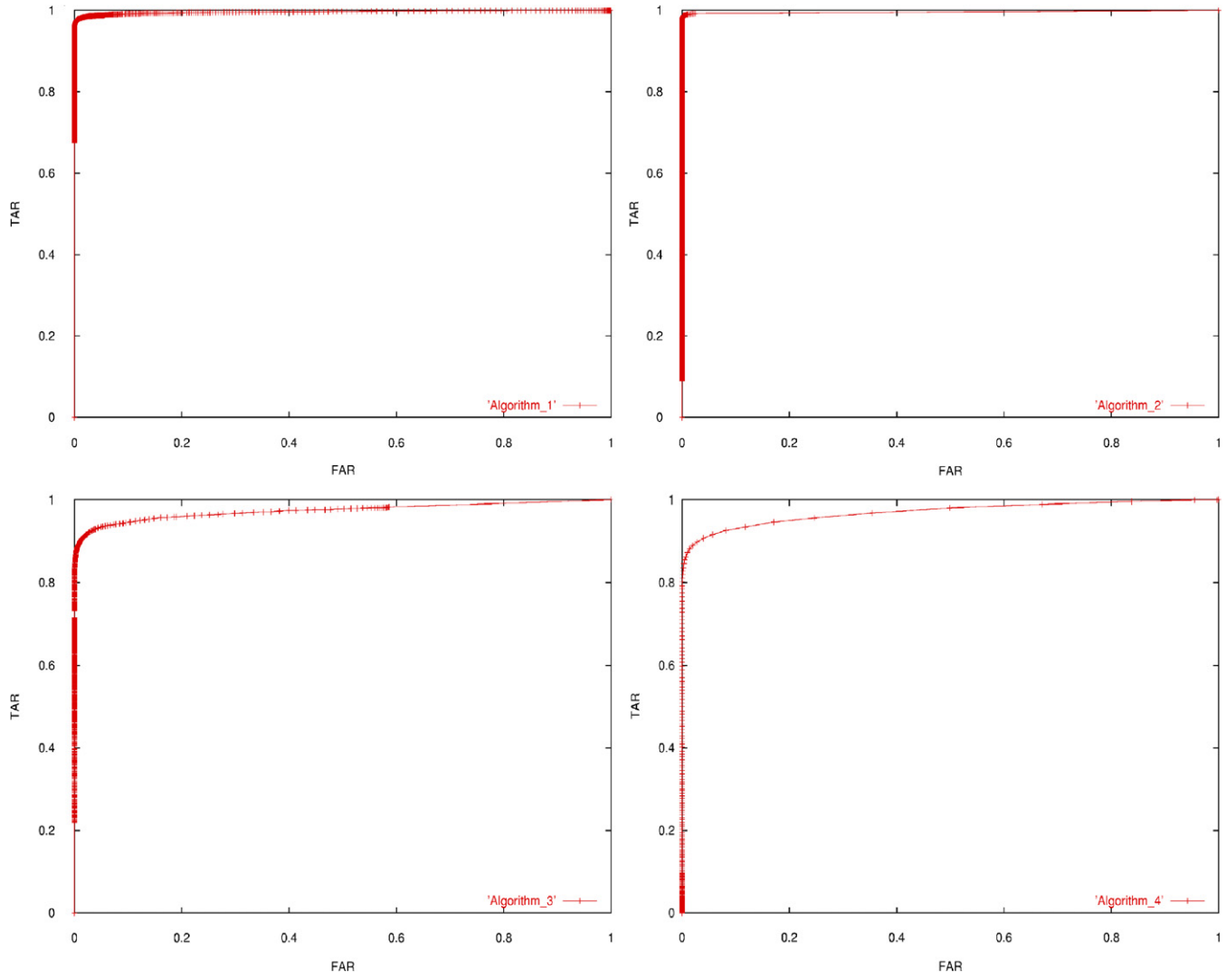

Fig. 6. The four ROC curves of Algorithms 1–4, respectively, where the FAR is in a linear scale to show ROC curves at the lower-score region of the non-match similarity scores. The cross points represent the points on which the ROC curves are constructed.

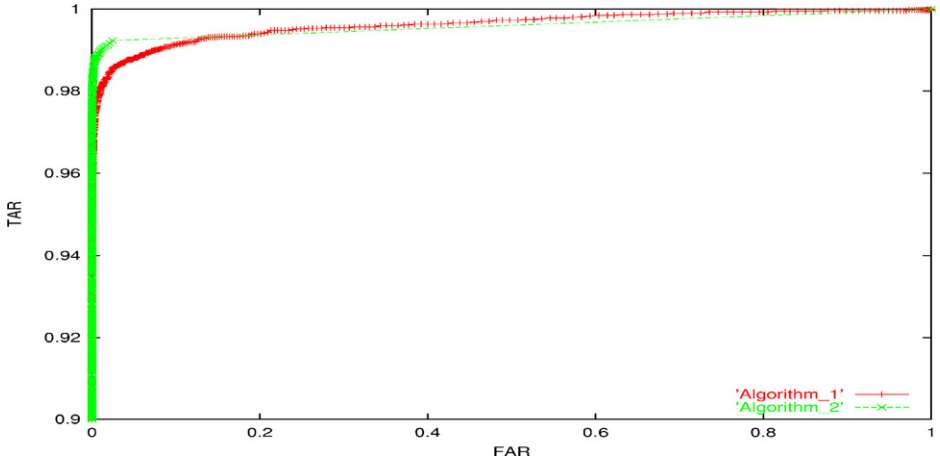

Fig. 7. Enlarged parts of ROC curves of Algorithms 1 and 2, where the non-match similarity scores are more significant. In this region, the ROC curve of Algorithm 1 is generally higher than the one of Algorithm 2. The cross points represent the points on which the ROC curves are constructed.

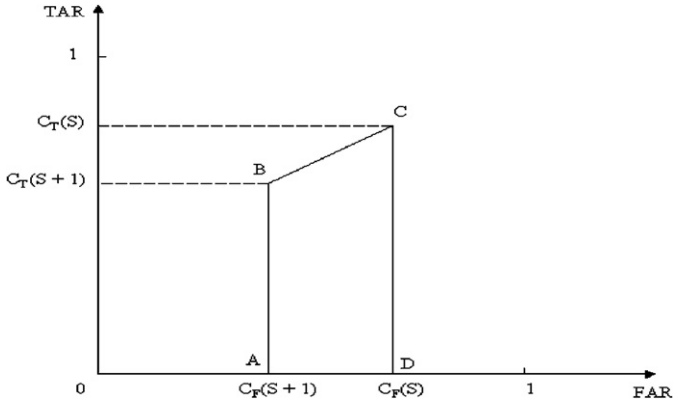

Fig. 8. A schematic drawing of four points A–D along with their coordinates in the FAR-and-TAR coordinate system. They form a trapezoid at the score s, and BC is a segment of an ROC curve.

The summation runs consecutively in the descending order from $s_{\max}$ to $s_{\min}$, including zero-frequency scores.

In order to relate the area under an ROC curve to the Mann–Whitney statistic, a nonparametric approach proceeds as follows. All the N_F scores in the non-match similarity score set **F** are compared with all the N_T scores in the match similarity score set **T**. If a non-match similarity score s_F is less than a match similarity score s_T, it counts 1; if equal, it counts $\frac{1}{2}$; and if greater, it counts zero. That is, for discrete scoring, this rule can be expressed as [10]

$$\mathbf{R}(s_T, s_F) = \begin{cases} 1 & \text{if } s_F < s_T, \\ \frac{1}{2} & \text{if } s_F = s_T, \\ 0 & \text{if } s_F > s_T. \end{cases} \tag{13}$$

By including zero-frequency scores, the first term in Eq. (12) shows the total number of score pairs in which the non-match similarity score is equal to the match similarity score, weighted by $\frac{1}{2}$ and divided by $N_T N_F$. And the second term in Eq. (12) represents the total number of score pairs in which the non-match similarity score is less than the match similarity score, weighted by 1 and divided by $N_T N_F$. This term is the so-called "the number of inversions" in a sequence formed by non-match and match similarity scores [14]. In other words, the area under an ROC curve can be re-written as

$$\hat{\mathbf{A}} = \frac{1}{N_T N_F} * \sum_{s_{T=1}}^{N_T} \sum_{s_{F=1}}^{N_F} \mathbf{R}(s_T, s_F). \tag{14}$$

Except for the coefficient, this is exactly the Mann–Whitney statistic formed by the match and non-match similarity scores. As a consequence, the variance of the area under an ROC curve can be obtained by computing the variance of the Mann–Whitney statistic.

In order to calculate the variance of the area under an ROC curve, two more cumulative probability distribution functions are required [10]. One of them cumulates the probabilities of match similarity scores from the highest similarity score down to the score that is one score higher than the current score,

$$\mathbf{Q}_T = \left\{ q_T(s) = \sum_{\tau=s+1}^{s\,\max} p_T(\tau) | \forall s \in \{s\} \right\}. \tag{15}$$

And the other one cumulates the probabilities of non-match similarity scores from the lowest similarity score up to the score that is one score lower than the current score,

$$\mathbf{Q}_F = \left\{ q_F(s) = \sum_{\tau=s\,\min}^{s-1} p_F(\tau) | \forall s \in \{s\} \right\}, \tag{16}$$

where another boundary condition $\sum_{\tau=s\,\min}^{s\,\min-1} = 0$ is assumed. Thereafter, using Eqs. (3) and (6), the probability $\mathbf{B_{TTF}}$, that two randomly chosen genuine matches will obtain higher similarity scores than one randomly chosen impostor match, can be written as

$$\mathbf{B_{TTF}} = \sum_{s=s\,\min}^{s\,\max} p_F(s) * \left[q_T^2(s) + q_T(s) * p_T(s) + \frac{1}{3} * p_T^2(s) \right]. \tag{17}$$

And the probability $\mathbf{B_{FFT}}$, that one randomly chosen genuine match will get higher similarity score than two randomly chosen impostor matches, can be expressed as

$$\mathbf{B_{FFT}} = \sum_{s=s\,\min}^{s\,\max} p_T(s) * \left[q_F^2(s) + q_F(s) * p_F(s) + \frac{1}{3} * p_F^2(s) \right]. \tag{18}$$

Finally, the variance of the area under an ROC curve is presented as [10]

$$\mathbf{Var}(\hat{A}) = \frac{1}{N_T N_F} * [\hat{A}(1-\hat{A}) + (N_T - 1)(\mathbf{B_{TTF}} - \hat{A}^2) + (N_F - 1)(\mathbf{B_{FFT}} - \hat{A}^2)]. \tag{19}$$

The standard error of the area under an ROC curve, SE $(\hat{A})$, is defined as the square root of its variance. Since the Mann–Whitney statistic is asymptotically normally distributed due to the Central Limit Theorem, the margin of error and the confidence interval with certain confidence level for each area under an ROC curve can be computed for large-size fingerprint data sets.

Table 1
The areas under ROC curves and their standard errors for four algorithms

Algorithms	Areas $(\hat{A})$	Standard errors (SE $(\hat{A})$)
1	0.996228	0.000544
2	0.996002	0.000659
3	0.974103	0.001535
4	0.970838	0.001492

The areas under ROC curves generated by four fingerprint-image matching algorithms along with their corresponding standard errors are shown in Table 1. All the standard errors are very small, because the areas are all very close to 1 and the sizes of the match and non-match similarity scores are very large [10]. Algorithm 1 has slightly larger area, i.e., higher

matching power than Algorithm 2. Is this difference by chance or real? By the same token, the matching power of Algorithm 3 is quite close to that of Algorithm 4, but both Algorithms 1 and 2 are better than Algorithms 3 and 4. The same questions arise. All these questions can be resolved by the statistical significance test of the difference between two areas under ROC curves.

5. *Z*-test of areas under two ROC curves

As discussed before, the Mann–Whitney statistic is asymptotically normally distributed regardless of the distributions of the match and non-match similarity scores thanks to the Central Limit Theorem. Thus, the straightforward way to test the significance of the difference between two areas under ROC curves is the Z-test. The Z statistic is defined as the difference of two areas divided by the square root of the variance of two-area difference [9], and it is subject to the standard normal distribution with zero expectation and a variance of one. The Z statistic can be expressed as

$$Z = \frac{\hat{A}_1 - \hat{A}_2}{\sqrt{\mathrm{SE}^2(\hat{A}_1) + \mathrm{SE}^2(\hat{A}_2) - 2r\mathrm{SE}(\hat{A}_1)\mathrm{SE}(\hat{A}_2)}} \tag{20}$$

where $\hat{A}_1$ and $\hat{A}_2$ are two areas under ROC curves, SE $(\hat{A}_1)$ and SE $(\hat{A}_2)$ are two standard errors of areas, respectively, and r is the correlation coefficient between two areas. While comparing performances of two fingerprint-image matching algorithms, for two areas with very close values, we have no reason to believe *a priori* that one algorithm is likely to be better than the other. In such cases, the two-tailed test needs to be invoked. Otherwise, the one-tailed test should be employed.

The two areas under ROC curves may or may not be correlated, depending on how the two ROC curves are constructed. For many applications in the analysis of fingerprint data, the two ROC curves are built based on different data sets, or different portions of the same data set, and so on. Under such circumstances, two sets of match similarity scores and two sets of non-match similarity scores that construct the two ROC curves, respectively, do not co-vary. And thus the two areas are not correlated.

However, in the tests discussed in this article, where two fingerprint-image matching algorithms are compared on the same fingerprint data set, the two areas under ROC curves are correlated. They are correlated through elements of the matrix that is formed by the probe and the gallery. Each matrix element is either match or non-match similarity score for two different algorithms, respectively, depending on whether or not the subject in the probe is the same as the subject in the gallery. Thus, such matrix elements establish the correlation between two sets of match and non-match similarity scores of two algorithms, respectively, and thereafter the correlation between two ROC curves.

As shown in the literature [14], the nonparametric Kendall's τ is asymptotically normally distributed, in the null hypothesis of no association between two sets of random variables, with expectation zero and a variance of $(4N+10)/9N(N-1)$ where N is the size of the data set. For example, if N equals 6000, there is only 5% probability for the absolute value of the Kendall's τ to be greater than 0.0169. However, for two matches of fingerprint images, all fingerprint-image matching algorithms have the same tendency to assign a higher (or lower) similarity score to the match where two fingerprint images are more (or less) similar. Such a characteristic of fingerprint-image matching algorithms may cause high positive correlation between two sets of match and non-match similarity scores of two algorithms, respectively. On the other side of the coin, this high correlation may be reduced due to the large magnitude of the size of the fingerprint data sets.

Table 2
The two-tailed p-values of two areas under ROC curves generated by four fingerprint-image matching algorithms

Algorithms	1	2	3	4
1	1.0000	0.7714	0.0000	0.0000
2		1.0000	0.0000	0.0000
3			1.0000	0.0862
4				1.0000

For the four fingerprint-image matching algorithms, the six correlation coefficients between two sets of 6000 match similarity scores range from 0.56 to 0.67. The size of non-match similarity score data is as large as 36 million. It is impractical to compute the Kendall's τ for this size of data sets, since its complexity is $O(N^2)$. Thus, the stochastic approach is invoked. A simple random sample with size of 360 000 non-match similarity scores is selected without replacement out of 35 994 000 data for one iteration, and the average Kendall's τ is computed from such 10 iterations. The six correlation coefficients between two sets of non-match similarity scores lie between 0.07 and 0.25. Using the table provided in Ref. [9], the six resultant correlation coefficients between two areas under ROC curves are from 0.17 through 0.24.

The pairwise two-tailed p-values of two areas under ROC curves for the four fingerprint-image matching algorithms are presented in Table 2. This table is symmetric. So the other part of the table is left blank. And obviously, all diagonal elements in Table 2 are identically equal to one.

For Algorithms 1 and 2, the two-tailed p-value is 0.7714, which is much greater than 5%. According to the table shown in article [9], the resultant correlation coefficient cannot be greater than the largest one of the two correlation coefficients between two sets of match and non-match similarity scores, respectively. Therefore, conservatively, even if using the largest one, i.e., 0.60 in this case, which is the Kendall's τ between two sets of 6000 match similarity scores and is computed without sampling, the two-tailed p-value is 0.6797, which is also much greater than 5%. This indicates that the difference between two areas under ROC curves of Algorithms 1 and 2 is not real but by chance. In other words, it is strongly assured that the performance of Algorithm 1 is most likely the same as the performance of Algorithm 2 at the significance level 77.14% and 67.97% in a conservative way.

For Algorithms 3 and 4, the two-tailed p-value is 0.0862 that is greater than 5% by 3.62%. By the same approach, the conservative two-tailed p-value is 0.0161 that is lower than 5% by 3.39%. Thus, the performance of Algorithm 3 is likely the same as the performance of Algorithm 4. In all other cases, as shown in Table 2, the two-tailed p-values are less than 0.00005 in four significant decimal places, which is way below 5%. As mentioned above, in all these cases, the values of areas are not quite close. Thus, the one-tailed test should be invoked. The one-tailed p-value is half of the two-tailed p-value. Hence, it unequivocally indicates that the differences between the areas under ROC curves in these cases are significantly real. In other words, the performances of the corresponding algorithms are most likely different—one is significantly better (or worse) than the other.

Even though the sizes of the fingerprint data sets are large, the Z statistic hypothesis test of using the areas under ROC curves along with their variances and the correlation coefficient can be implemented. This provides a sound statistical ground for testing the significance of the differences between two areas under ROC curves, and thus for evaluating the performances of fingerprint-image matching algorithms.

6. Conclusions

As illustrated in this paper, the discrete probability distribution functions of the match and non-match similarity scores, generated by using fingerprint-image matching algorithms on the large-size data sets, have no definite underlying distribution functions. These distributions vary considerably from algorithm to algorithm. As a consequence, the nonparametric approach must be employed in the analysis of the large size of fingerprint similarity scores.

Although the sizes of fingerprint data sets are much larger than the sizes of the data sets that are dealt with in the medical practice, a precise ROC curve can still be realistically constructed by moving the threshold one integral score at a time from the highest similarity score down to the lowest similarity score. Then, by invoking the trapezoidal rule, the area under an ROC curve can be calculated. This area is equivalent to the Mann–Whitney statistic directly formed from the match and non-match similarity scores.

The area under an ROC curve stands for the probability that the score obtained for the genuine match is higher than the score assigned for the impostor match given both genuine match and impostor match assuming the score is a continuous random variable. Therefore, to evaluate a fingerprint-image matching algorithm, an ROC curve as a whole rather than an ROC curve at a specific point or within a chosen region should be taken into account. Even if a part of an ROC curve produces higher TAR values, this does not guarantee that the ROC curve as a whole is better. The examples shown in this article demonstrated such relationship.

Furthermore, the variance of the area under an ROC curve can be obtained by calculating the variance of the Mann–Whitney statistic. In addition, the Mann–Whitney statistic is asymptotically normally distributed regardless of the distributions of the match and non-match similarity scores thanks to the Central Limit Theorem. Thus, the Z statistic can be formulated. Two-tailed test and/or one-tailed test are conducted based on how much close the values of two areas under ROC curves are. The Z statistic can be computed in a conservative way depending on how to deal with the correlation coefficient.

The fingerprint data sets are large-size data sets. Even on the same data set, different fingerprint-image matching algorithms generate a wide variety of match and non-match distributions. Moreover, uncertainties can arise from processing and comparing fingerprint system test results. Under such circumstances, the Z statistic hypothesis test offers a systematic way to detect the statistical significance of differences between two underlying ROC curves, namely, differences between performances of two fingerprint-image matching algorithms. The method investigated in this article provides the information on which algorithm produces better results, as well as the information about whether the difference is real or just by chance at a quantified significance level.

The approach of analyzing ROC curves using areas under ROC curves has been successfully applied to the analysis of large samples of fingerprint data. In this article, this methodology is applied to comparing two fingerprint-image matching algorithms on the same data set. It can also be applied, for instance, to evaluating the relationship among different fingerprint image qualities. As a matter of fact, in general, the nonparametric approach presented in this article can be employed in the analysis of many kinds of biometric data.

References

[1] C.L. Wilson, et al., Fingerprint vendor technology evaluation 2003: summary of results and analysis report, NISTIR 7123, National Institute of Standards and Technology, June 2004.
[2] C. Watson, C. Wilson, K. Marshall, M. Indovina, R. Snelick, Studies of one-to-one fingerprint matching with vendor SDK matchers, NISTIR 7119, National Institute of Standards and Technology, May 2004.
[3] D. Maio, D. Maltoni, R. Cappelli, J.L. Wayman, A.K. Jain, FVC2000: fingerprint verification competition, IEEE Trans. PAMI 24 (3) (2002) 402–412.
[4] S. Wieand, M.H. Gail, B.R. James, K.L. James, A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data, Biometrika 76 (3) (1989) 585–592.
[5] K.H. Zou, Comparison of correlated receiver operating characteristic curves derived from repeated diagnostic test data, Acad. Radiol. 8 (3) (2001) 225–233.
[6] C.T. Le, B.R. Lindgren, Construction and comparison of two receiver operating characteristic curves derived from the same samples, Biom. J. 37 (7) (1995) 869–877.
[7] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics 44 (1988) 837–845.
[8] D.K. McClish, Comparing the areas under more than two independent ROC curves, Med. Decision Making 7 (1987) 149–155.
[9] J.A. Hanley, B.J. McNeil, A method of comparing the area under two ROC curves derived from the same cases, Radiology 148 (1983) 839–843.
[10] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982) 29–36.

[11] D. Green, J. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966 pp. 45–49.
[12] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, J. Math. Psychol. 12 (1975) 387–415.
[13] G.E. Noether, Elements of Nonparametric Statistics, Wiley, New York, 1967 pp. 31–32.
[14] B.L. van der Waerden, Mathematical Statistics, Springer, Berlin, 1969 p. 274 and pp. 333–335.

About the Author—DR. JIN CHU WU received his Ph.D. in theoretical high energy physics from the University of Pittsburgh. His research focused on grand unification theories (GUTs) and lattice gauge theory. He joined the Superconducting Super Collider Laboratory in Dallas, Texas. Currently, he works at the National Institute of Standards and Technology. Dr. Wu's research interests are nonparametric data analysis, sample sizes, and significance test in biometric evaluation.

About the Author—CHARLES L. WILSON has worked in various areas of computer modeling ranging from semiconductor device simulation, for which he received a DOC gold medal in 1983, and computer-aided design to neural network pattern recognition at Los Alamos National Laboratory, AT & T Bell Laboratories and for the last 27 years at NIST. He is the manager of the Image Group in the Information Access Division. His current research interests are in application of statistical pattern recognition, neural network methods, and dynamic training methods for image recognition, image compression, optical information processing systems, and in standards used to evaluate recognition systems.