NISTIR 7226

Nonparametric Analysis of Fingerprint Data

Jin Chu Wu
Charles L. Wilson
U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards and Technology
Information Access Division
Gaithersburg, MD 20899

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

NISTIR 7226

Nonparametric Analysis of Fingerprint Data

Jin Chu Wu
Charles L. Wilson
U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards and Technology
Information Access Division
Gaithersburg, MD 20899

May 2005

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary
TECHNOLOGY ADMINISTRATION
Phillip J. Bond, Under Secretary of Commerce for Technology
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
Hratch G. Semerjian, Acting Director

Nonparametric Analysis of Fingerprint Data

Jin Chu Wu* and Charles L. Wilson
Image Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

This paper demonstrates that, for large-scale tests, the match and non-match similarity scores have no specific underlying distribution function. The forms of these distribution functions require a nonparametric approach for the analysis of the fingerprint similarity scores. In this paper, we present an analysis of the discrete distribution functions of the match and non-match similarity scores of the fingerprint data. This analysis demonstrates that 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 without any assumption regarding operating thresholds and the forms of the distribution functions. 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 defined using the areas under ROC curves along with their variances is applied to test the significance of the difference between two ROC curves. Four examples from NIST's extensive testing of commercial fingerprint systems are provided. The nonparametric approach presented in this article can also be employed in the analysis of other biometric data.

Keywords: Fingerprint matching; Nonparametric analysis; Receiver Operating Characteristic (ROC) curve; Mann-Whitney statistic; Significance test

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

1. Introduction

Recently, the National Institute of Standards and Technology (NIST) has evaluated the fingerprint matching from different vendors[*] [1,2]. These evaluations of vendors' fingerprint-image matching algorithms were performed 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 datasets. In the SDK tests [2], a probe consisting of 6000 of the subjects' fingerprint images (such as left index finger, etc.), is compared against a gallery built from 6000 of the subjects' fingerprint images (such as left index finger, etc.). This requires 36,000,000 comparisons and is the smallest of the large-scale tests discussed in this paper. These evaluations were conducted on 19 different vendor's fingerprint-image matching algorithms.

A score generated by comparing two different fingerprint images of the same subject who appears both in the probe and in the gallery is called match similarity score (i.e., genuine-match score). A score generated by comparing two fingerprint images of two different subjects is called non-match similarity score (i.e., impostor-match score). The fingerprint-image matching algorithms tested in [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 and the lower values of similarity scores represent that two fingerprint images are less similar. Hence, the distribution function of the match similarity scores will be centered at higher scores than the distribution function of the non-match similarity scores.

The True Accept Rate (TAR) is defined as the cumulative probability of the match similarity scores from the highest match similarity score at a specific similarity score. The False Accept Rate (FAR) is specified as the cumulative probability of the non-match similarity scores from the highest non-match similarity score at a specific similarity score. Based on the TAR and FAR, a

[*] 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.

Receiver Operating Characteristic (ROC) curve can be constructed. The technology evaluations of fingerprint-image matching algorithms can be carried out by using ROC curves for comparing the performances of their corresponding algorithms.

The similarity scores generated by using fingerprint-image matching algorithms are usually represented in integers with different ranges of values. Even though similarity scores of a fingerprint-image matching algorithm are expressed, for example, as real numbers ranging from zero to one inclusively with five significant decimal places, they can be easily converted into integers after multiplying by 10^5. Integers are much easier to be dealt with than real numbers. Therefore, in this article, the similarity scores are treated as discrete random variables rather than continuous random variables. One of benefits that can be obtained from this is that with large samples of fingerprint used to evaluate matching algorithms, precise ROC curves can be calculated. These ROCs are based upon the distribution of the match similarity scores and the distribution of the non-match similarity scores and can be easily constructed by moving one integral score at a time, i.e., without any assumption regarding the threshold of discrete similarity scores.

The match similarity scores as well as the 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 approach is pertinent to the analysis of the fingerprint data. Evaluation of the performance 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 relevant using system design criteria [1-4]. However, in other approaches, the area under the 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 with a small datasets. The biometric evaluations cited in the references [1,2] used large data sets. In the analysis of large amount of fingerprint data and in the evaluation of the fingerprint-image matching algorithms, the technique using the area under an ROC curve has not been explored [1-3].

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 performance of an ROC curve as a whole. Second, the area under an ROC curve computed using the trapezoidal rule is exactly the Mann-Whitney statistic directly formed, in our case, by the match similarity scores and the non-match similarity scores [9,10,12,13].

There are two consequences of the above second point. First, the variance of the Mann-Whitney statistic can be utilized as the variance of the area. Second, since the Mann-Whitney statistic is asymptotically normally distributed regardless of the distributions of the match similarity scores and the non-match similarity scores thanks to the Central Limit Theorem [8,13,15], the Z statistic formulated in terms of areas under two independent ROC curves, generated by large-size fingerprint dataset, along with their variances, is subject to the standard normal distribution with zero expectation and a variance of one and can be used to test the significance of the difference of these two areas.

As pointed out in the reference [9] and references therein, there are other ways to calculate the area under an ROC curve and its variance. In this article, the area under a precise ROC curve, generated from large-size fingerprint dataset, is computed using the trapezoidal rule, and thus the variance of the Mann-Whitney statistic is employed. The technique of using the area under an ROC curve as the criterion to evaluate the performance of biometric systems provides a sound ground for conducting statistical significance test for measuring the difference between two fingerprint ROC curves in a nonparametric way. Our analysis will be performed using large data samples without assumptions about the forms of the match and non-match score distributions.

And also in this article, the detailed formulas for constructing an ROC curve, computing the area under an ROC curve using the trapezoidal approach, and calculating the variance of the area under an ROC curve derived from the variance of the Mann-Whitney statistic are provided. These formulas are expressed in a way that allows them to be coded easily by using arrays. This is very useful for dealing with large datasets, such as the data from fingerprints as investigated in this article.

The discrete distribution functions of the match and non-match similarity scores from the fingerprint data 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 computed using the areas under ROC curves along with their variances is applied to test the significance of the difference between two ROC curves. This is presented in Section 5. As the contents are presented, some examples will be provided. Finally, conclusions are presented in Section 6.

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

Studying the distribution function is the usual starting point of analyzing fingerprint data. The match similarity score set is a set of integral scores, as discussed earlier, generated by matching two different fingerprint images of the same subject,

$$\mathbf{T} = \{ s_i \mid \forall\, i \in \{1, \ldots, N_T\}\} \qquad (1)$$

where N_T is the total number of match similarity scores. The match similarity score set **T** can be represented in an array with N_T elements, and each element records a match similarity score.

Let the integral score set be $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$, where s_{min} and s_{max} are the minimum and maximum similarity scores, respectively, with respect to a specific fingerprint-image matching algorithm for a specific application. This score set $\{s\}$ consists of consecutive integers from s_{min} up to s_{max}. In Equation (1), 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 **T** can be partitioned into pairwise-disjoint subsets $\{\mathbf{T}_s\}$. In each of the subsets, $\mathbf{T}_s$, the members have the same integer $s \in \{s\}$. The match similarity score set **T** is the union of all these subsets $\{\mathbf{T}_s\}$.

The frequency $f_T(s)$ of the similarity score s, which appears in the match similarity score set **T**, is the size of the subset $\mathbf{T}_s$ that shares the similarity score s. To deal with the whole spectrum of the scores, the integral scores, that appear in the score set $\{s\}$ but not in the match similarity score set **T**, must be included. The frequencies for these scores are obviously equal to zero. 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. The corresponding probability $p_T(s)$ equals the frequency $f_T(s)$ divided by the total number of match similarity scores, N_T, i.e., $p_T(s) = f_T(s) / N_T$. Therefore, by including zero frequencies, the discrete frequency distribution function of the match similarity scores can be expressed in terms of the frequency $f_T(s)$ as

$$\boldsymbol{F_T} = \{\ f_T(s) \mid \forall\ s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} f_T(\tau) = N_T\ \} \tag{2}$$

And the discrete probability distribution function of the match similarity scores can be represented in terms of the probability $p_T(s)$ as

$$\boldsymbol{P_T} = \{\ p_T(s) \mid \forall\ s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} p_T(\tau) = 1\ \} \tag{3}$$

The discrete frequency distribution function and the discrete probability distribution function of the match similarity scores can indeed be expressed in arrays with $s_{max} - s_{min} + 1$ elements, and each element contains the frequency and the probability of the match similarity score s, i.e., $f_T(s)$ and $p_T(s)$, respectively.

The non-match similarity score set is a set of integral scores created by comparing two fingerprint images of two different subjects,

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

where N_F is the total number of non-match similarity scores. By analogy with the match similarity scores, the discrete frequency distribution function of the non-match similarity scores can be formulated as

$$\boldsymbol{F_F} = \{\ f_F(s) \mid \forall\ s \in \{s\} \text{ and } \sum_{\tau=s\min}^{s\max} f_F(\tau) = N_F\ \} \tag{5}$$

where $f_F(s)$ is the frequency of the score s occurring in the non-match similarity score set **F**, including zero frequencies. And the discrete probability distribution function of the non-match similarity scores can be expressed as

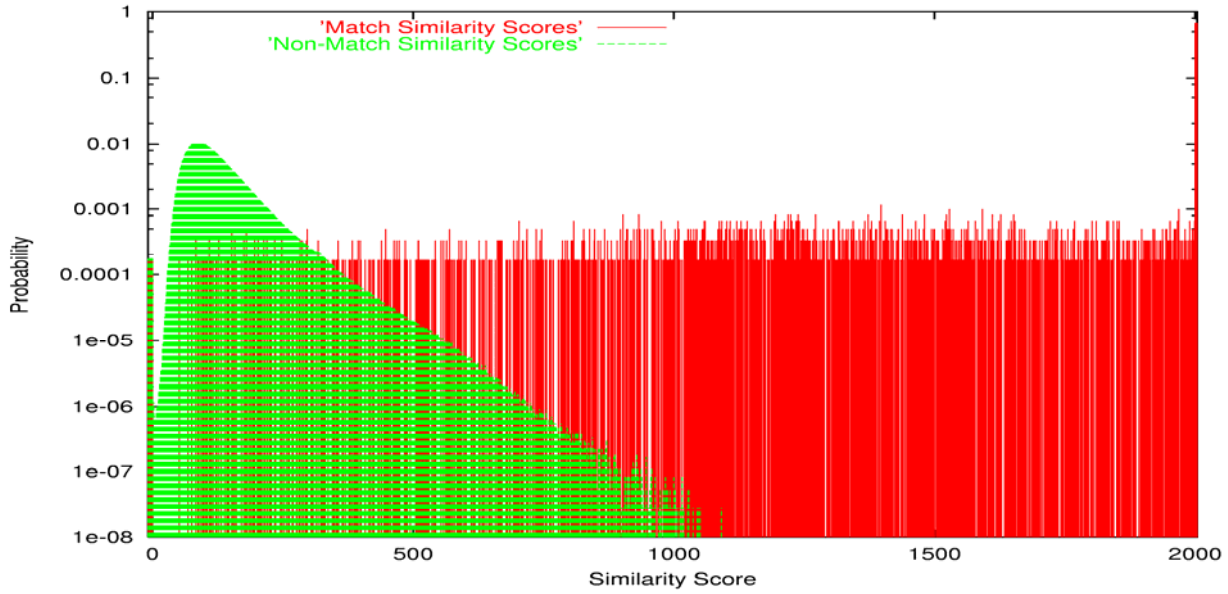

Figure 1 The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 1. The integral similarity scores run from 0 to 2000. The widths of peaks at the highest score and at the lowest score are enlarged to show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

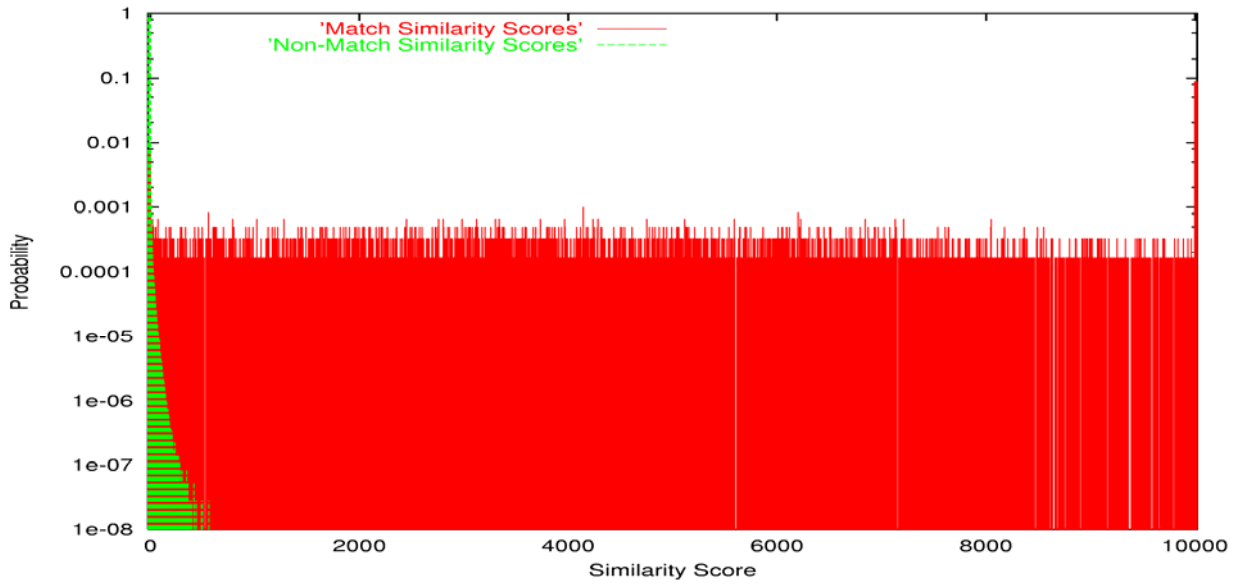

Figure 2 The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 2. The integral similarity scores run from 0 to 9999. The widths of peaks at the highest score and at the lowest score are enlarged to show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

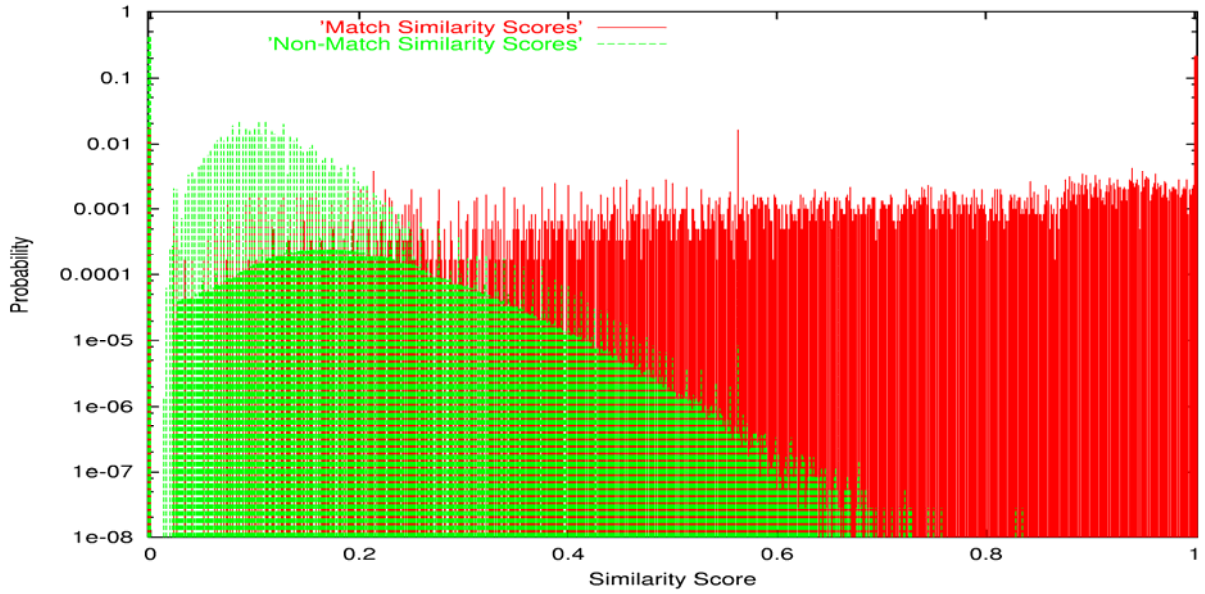

Figure 3 The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 3. The real-number similarity scores run from 0.0 to 1.0 in five significant decimal places, which can be converted into integers. The widths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

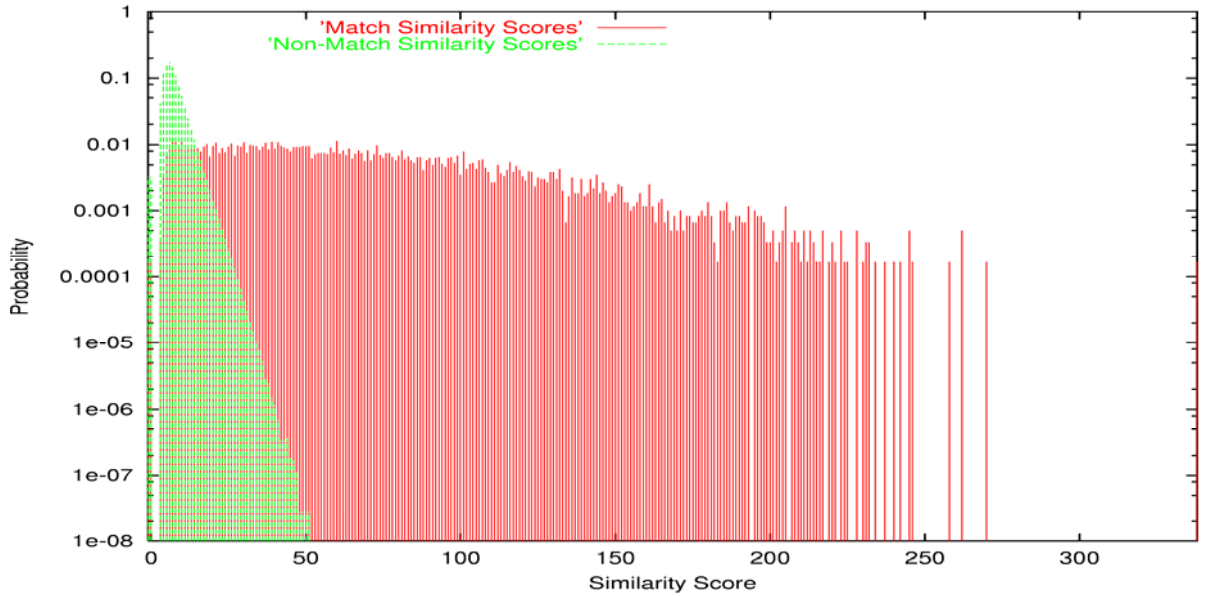

Figure 4 The discrete probability distribution functions of the match and non-match similarity scores generated by using the fingerprint-image matching Algorithm 4. The integral similarity scores run from 0 to 338. The widths of peaks at the lowest score are enlarged to show the characteristics of the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

$$\boldsymbol{P}_{\mathbf{F}} = \{\, p_F(s) \mid \forall\, s \in \{s\} \text{ and } \sum_{\tau=s\,\min}^{s\,\max} p_F(\tau) = 1 \,\} \qquad (6)$$

where $p_F(s)$ is the corresponding probability, i.e., $p_F(s) = f_F(s) / N_F$. By the same token, the non-match similarity score set **F,** its discrete frequency distribution function $\boldsymbol{F}_{\mathbf{F}}$, and its discrete probability distribution function $\boldsymbol{P}_{\mathbf{F}}$ can all be represented in arrays.

Figure 1 to Figure 4 show the discrete probability distribution functions of the match and non-match similarity scores generated by using fingerprint-image matching Algorithms 1, 2, 3 and 4 on the same fingerprint dataset, respectively. These distributions are based upon the integer characteristic of similarity scores. The total number of non-match similarity scores N_F is much greater than the total number of match similarity scores N_T. In our studies, N_T is 6000, and N_F is as large as about 36 million. This means that the least probability of the match similarity scores is on the order of 10^{-4}, whereas the least probability of the non-match similarity scores is on the order of 10^{-8}. In order to show such small probabilities for the large amount of fingerprint data, the probability is depicted in logarithmic scale.

In these figures many match and non-match similarity scores appear only once or twice out of thousands and millions of integral scores, respectively. Most importantly, despite that different fingerprint-image matching algorithms invoked different scoring systems, as far as the discrete probability distribution functions of the random similarity scores are concerned, these four figures show that different algorithms have different characteristics of probability distribution functions of the match and non-match similarity scores.

For Algorithm 1, the highest match similarity score is 2000, that dominates 67.52% of the whole population of the match similarity scores. And other match similarity scores are distributed between 0 and 1999 with relatively high probabilities at higher scores and very low probabilities at lower scores. But the probability distribution of the non-match similarity scores is a normal-like distribution, skewed towards higher scores. The highest non-match similarity score occurs almost in the middle of its score range.

For Algorithm 2, the highest match similarity score is 9999, which takes 8.98% of the whole population of the match similarity scores. The lowest one is 0, which is 0.77% of the population. Other match similarity scores are almost uniformly distributed between the lowest score and the highest score. However, the lowest non-match similarity score, 0, overwhelmingly occupies 97.56% of the whole population of the non-match similarity scores. And other non-match similarity scores are just concentrated in the region from 1 through 557 with very steep-decay probabilities. The highest non-match similarity score appears at the very low end of its score range.

Algorithm 1 and Algorithm 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.

For Algorithm 3, the real-number match and non-match similarity scores can be easily converted into integers as discussed before. The non-match similarity scores have a peak at the lowest similarity score 0.0, which counts 41.33% of the population of the non-match similarity scores and is separated from a normal-like probability distribution. For Algorithm 4, only very a few match similarity scores appear in the high-score range, the score of which is greater than 250, and there is also a gap between the lowest non-match similarity score zero and the second lowest one.

All in all, from these discrete probability distribution functions, it is very important to notice that the match and non-match similarity scores generated by the fingerprint-image matching algorithms have no definite underlying distribution functions. As a consequence, a nonparametric analysis must be employed in order to deal with such fingerprint data.

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. The ROC curve can be used in fingerprint systems to select an operating point that gives an acceptable trade-off between system accuracy and reliability. An ROC curve is constructed based on the cumulative discrete probability distribution functions of the match and non-match similarity scores.

From Equations (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

$$\boldsymbol{C_T} = \{\, c_T(s) = \sum_{\tau=s}^{s_{max}} p_T(\tau) \mid \forall\, s \in \{\bar{s}\} \,\} \tag{7}$$

and

$$\boldsymbol{C_F} = \{\, c_F(s) = \sum_{\tau=s}^{s_{max}} p_F(\tau) \mid \forall\, s \in \{\bar{s}\} \,\} \tag{8}$$

where $c_T(s)$ and $c_F(s)$ are the cumulative probabilities of the match and non-match similarity scores at each 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, $(\, (c_F(s), c_T(s)) \mid \forall\, s \in \{\bar{s}\} \,)$, 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 match similarity score must be greater than the probability of the non-match similarity score (that may very well be zero in our case). An ROC curve starts from the origin of the FAR-and-TAR coordinate system and ends at the point (1, 1)

above the straight line that is 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 the threshold. If any assumption about the threshold is made, some information from the probability distribution functions of the match and non-match similarity scores can be lost. For the discrete similarity scores it is hard to determine the correct partition of the population, if the threshold happens to be in a part of the population that shares the same similarity score. For small datasets, like those found in the medical practice, researchers always employ this type of ROC curve. However, for large datasets, such as data used to test fingerprint-image matching algorithms, there has been little discussion of such a precise ROC curve [1-3].

The ROC curves, corresponding to the four fingerprint-image matching algorithms presented in the previous section, are shown in Figure 5 and Figure 6. In Figure 5 a logarithmic scale is used for the FAR to show the performance of the ROC curve at the higher-score region of the non-match similarity scores. In Figure 6 a linear scale is used to show the performance of the ROC curve at the lower-score region of the non-match similarity scores. The score range of the non-match similarity scores varies from algorithm to algorithm, as illustrated in Figure 1 through Figure 4.

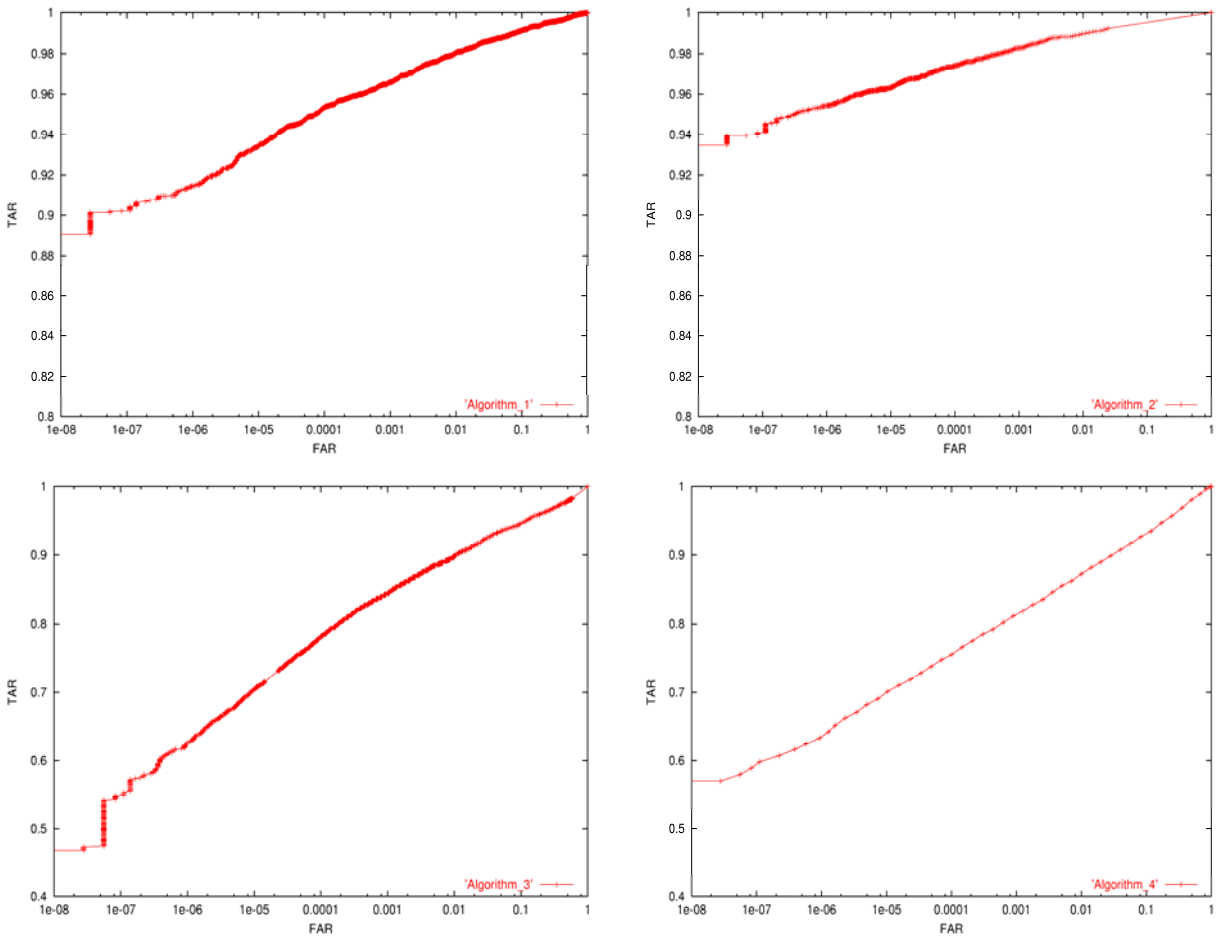

Figure 5 The four ROC curves of Algorithm 1, 2, 3, and 4, respectively, where the FAR is in a logarithmic scale to show the performance of the ROC curve at the higher-score region of the non-match similarity scores. The cross points represent the points on which the ROC curves are constructed.

For Algorithm 1, the second point on the ROC curve, i.e., one point above the origin (0, 0), is at (0, 0.6752) (see Figure 6), due to the peak of the match similarity scores at the highest similarity score 2000, which dominates 67.52% of the population. The ROC curve does not leave the TAR coordinate axis until the highest non-match similarity score is reached. The highest non-match similarity score appears only once in this case, and thus its probability is a little above 10^{-8}. At that point, the cumulative probability of the match similarity scores from the highest similarity score has already reached 89.05% (see Figure 1 and Figure 5). Furthermore, 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, the ROC curve gradually reaches the point (1, 1) from one side of the FAR-and-TAR coordinate system (see Figure 1and Figure 6).

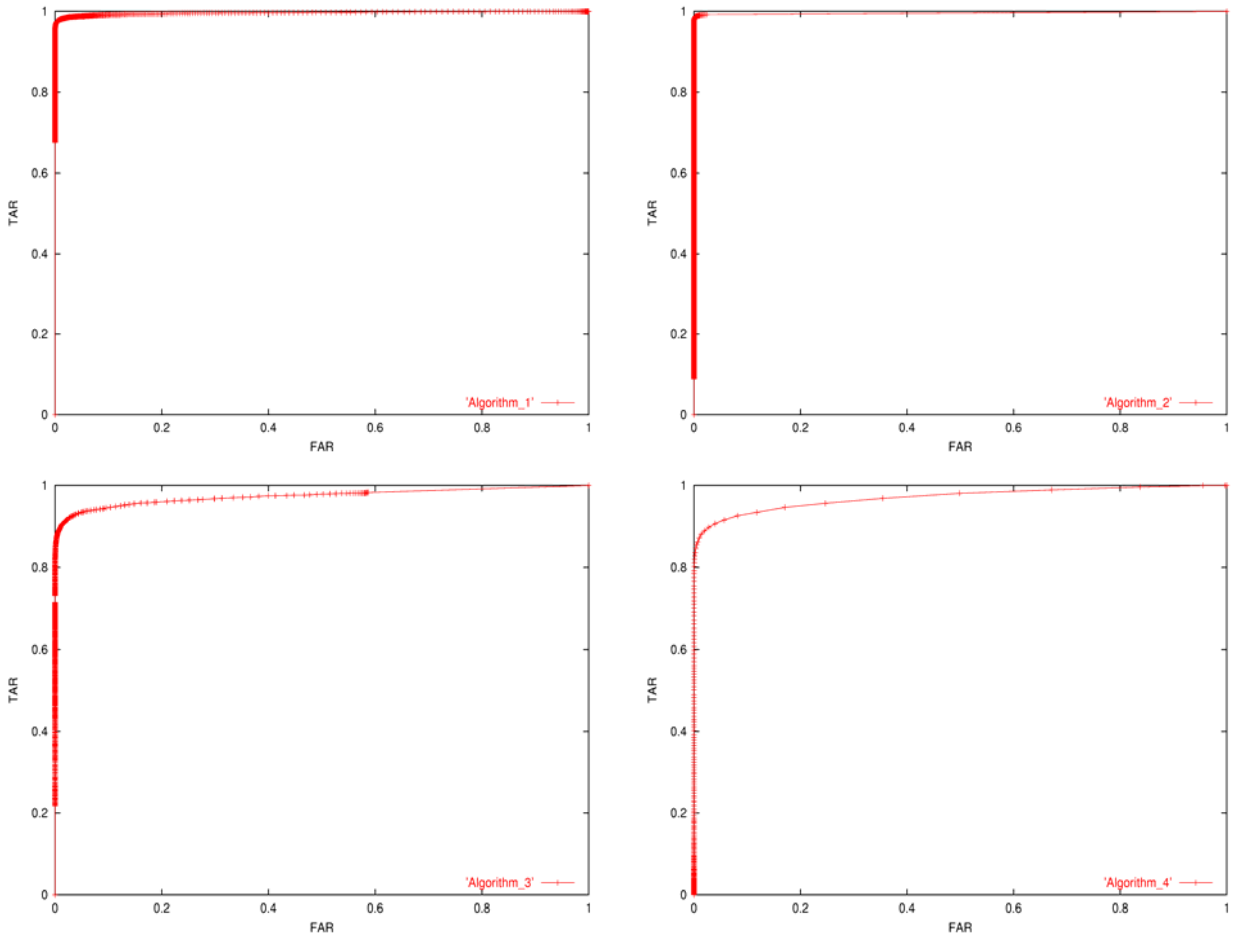

Figure 6 The four ROC curves of Algorithm 1, 2, 3, and 4, respectively, where the FAR is in a linear scale to show the performance of the ROC curve at the lower-score region of the non-match similarity scores. The cross points represent the points on which the ROC curves are constructed.

In contrast, for Algorithm 2, the ROC curve leaves the TAR coordinate axis when the cumulative probability of the match similarity scores from the highest similarity score gets to 93.47% (see Figure 2 and Figure 5). This is higher than 89.05% for Algorithm 1. However, it is intriguing to see that the ROC curve jumps from one side of the FAR-and-TAR coordinate system to the final point (1, 1) (see Figure 6). This is because the peak of the probability distribution of the non-match similarity scores, occurring at the lowest similarity score 0, overwhelmingly occupies 97.56% of the population as shown before (see Figure 2). The ROC curve of Algorithm 1 is generally higher than the one of Algorithm 2 in the region where the non-match similarity scores are more significant. This feature is evidenced by Figure 7, in which the upper parts of the ROC curves are shown on an enlarged scale.

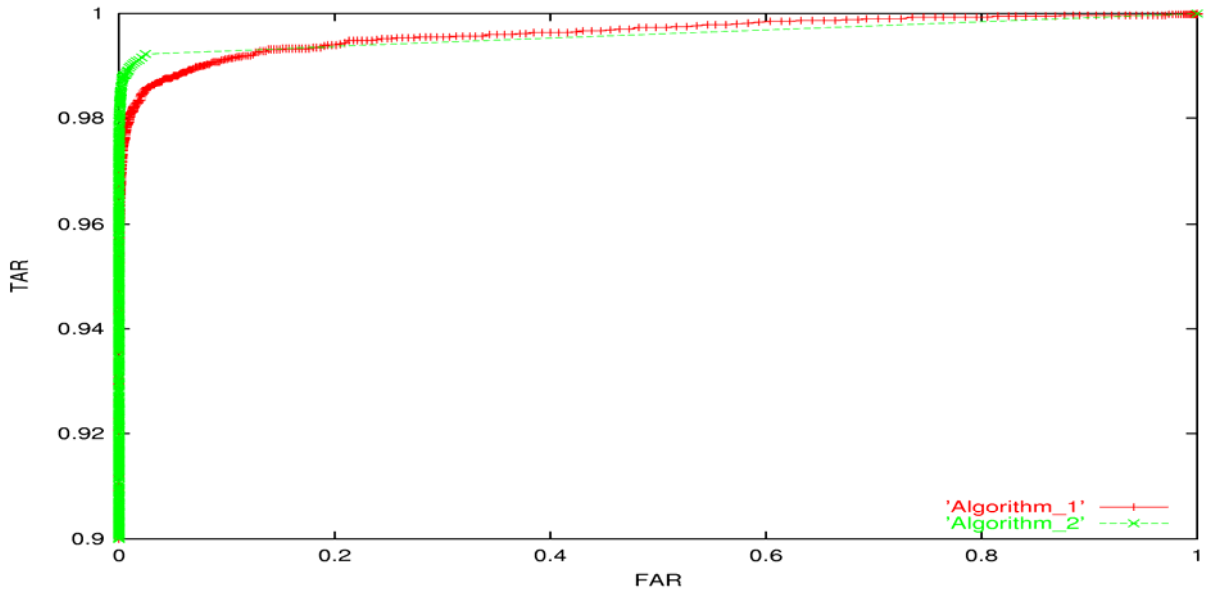

Figure 7 Enlarged parts of ROC curves of Algorithms 1 and 2, where the non-match similarity scores are more significant. In this region, the ROC curve of Algorithm 1 is generally higher than the one of Algorithm 2. The cross points represent the points on which the ROC curves are constructed.

The same qualitative analyses can be applied to the ROC curves of Algorithm 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 (see Figure 3, Figure 4, and Figure 5). The evaluation of the performance of an ROC curve should take account of the whole ROC curve from the beginning point to the end point.

For large datasets, 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

The performance of an ROC curve can be quantitatively measured using the area under the ROC curve. This had been studied in the literature [9-11]. The area under an ROC curve is a very important index, which represents the probability that, in our case, the score obtained for the genuine match is higher than the score assigned for the impostor match given both genuine match and impostor match, i.e., ***Prob*** $(s_G > s_I)$, where s_G is a genuine-match score and s_I is an impostor-match score. 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 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.

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. Therefore, 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. The rectangle in the first trapezoid, where the origin of the FAR-and-TAR coordinate system and the point $(c_F(s_{max}), c_T(s_{max}))$ (i.e., $(p_F(s_{max}), p_T(s_{max}))$) are two corner points, does not exist. Also it is clear that the trapezoid can be reduced to a rectangle, a vertical line, or a point, if the zero-frequency match and/or non-match similarity scores are involved.

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: $(c_F(s + 1), 0)$, $(c_F(s + 1), c_T(s + 1))$, $(c_F(s), c_T(s))$, and $(c_F(s), 0)$, in clockwise direction, assuming $c_F(s_{max} + 1) = c_T(s_{max} + 1) = 0$. This boundary condition corresponds to the above first trapezoid (reduced to a triangle), 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 Equations (7) and (8), it follows that at the score $s \in \{\bar{s}\}$ where the scores are counted consecutively in the descending order from s_{max} to s_{min}, 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. Therefore, the area under an ROC curve can be computed as

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

where the first trapezoid at the beginning of an ROC curve is at $s = s_{max}$. The summation starts from the highest similarity score s_{max} and ends at the lowest similarity score s_{min}, with scores taken consecutively in the descending order, including zero-frequency scores. This is because the ROC curve is built from the cumulative probabilities of the match and non-match similarity scores, respectively, from the highest similarity score s_{max}.

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 ½; and if greater, it counts zero. That is, for discrete scoring, the rule invoked here for comparing non-match similarity scores against match similarity scores, or vice versa, can be expressed as [10]

$$\begin{cases} 1 & \text{if } s_F < s_T \end{cases}$$

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

By including zero-frequency scores, the first term in Equation (12) shows the total number of score pairs in which the non-match similarity score is equal to the match similarity score, weighted by ½ and divided by $N_T N_F$. And the second term in Equation (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 [15]. In other words, the area under an ROC curve can be re-written as

$$\hat{\mathbf{A}} = \frac{1}{N_T N_F} * \sum_{sT=1}^{NT} \sum_{sF=1}^{NF} \mathbf{R}\,(s_T, s_F) \tag{14}$$

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 these accumulates the probabilities of the match similarity scores from the highest similarity score down to the score that is one score higher than the current score,

$$\boldsymbol{Q_T} = \{\, q_T(s) = \sum_{\tau=s+1}^{s\max} p_T(\tau) \mid \forall\, s \in \{s\} \,\} \tag{15}$$

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

$$\boldsymbol{Q_F} = \{\, q_F(s) = \sum_{\tau=s\min}^{s-1} p_F(\tau) \mid \forall\, s \in \{s\} \,\} \tag{16}$$

where another boundary condition $\sum_{\tau=s\min}^{s\min-1} = 0$ is assumed. Thereafter, using Equations (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

Algorithms	Areas (Â)	Standard Errors (SE (Â))

1	0.996228	0.000544
2	0.996002	0.000659
3	0.974103	0.001535
4	0.970838	0.001492

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

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

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

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

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

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

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 with corresponding confidence level and thus the confidence interval for each area under an ROC curve can be accordingly constructed for large-size fingerprint datasets. The area under an ROC curve, i.e., Equation (12), can also be expressed in terms of Equations (3), (6), and (15). The formulas of computing the area and its variance are presented in an explicitly mathematical way so that they can be easily coded using arrays. This will be very helpful for dealing with large datasets used for fingerprint system testing.

The areas under ROC curves generated by four fingerprint-image matching algorithms along with their corresponding standard errors are shown in Table 1. The area of Algorithm 1 is slightly larger than the one of Algorithm 2. So are the areas of Algorithms 3 and 4. But the areas

of Algorithms 1 and 2 are both larger than the areas of Algorithms 3 and 4. However, 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].

The ROC curve of Algorithm 1 leaves the TAR coordinate axis in the FAR-and-TAR coordinate system at 89.05%, which is lower than 93.47% where the ROC curve of Algorithm 2 leaves the TAR coordinate axis. This relation also holds true in the region where the non-match similarity scores are just becoming significant (see Figure 5). However, in the region where the non-match similarity scores are becoming more and more significant, the relation is reversed. While the FAR reaches about 20%, the ROC curve of Algorithm 1 starts to be higher than the ROC curve of Algorithm 2 (see Figure 7). Therefore the area of Algorithm 1 is a little larger than that of Algorithm 2. This example shows that even if the performance of a part of an ROC curve produces a higher TAR value at a specified FAR value, this does not guarantee the performance of an ROC curve as a whole produces a higher cumulative accuracy.

The area under a whole ROC curve measures the ability of fingerprint-image matching algorithms to produce matches over the entire range of match and non-match similarity. In this regard, Algorithm 1 has slightly higher matching power than Algorithm 2. Is this difference by chance or real? By the same token, Table 1 shows that the matching power of Algorithm 3 is quite close to that of Algorithm 4, but both Algorithm 1 and 2 are much better than Algorithms 3 and 4. The same question arises. Is this difference by chance or real? All these questions can be resolved by the statistical significance test of the difference between two areas under the fingerprint 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 similarity scores and the 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,

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

where $\hat{A}_1$ and $\hat{A}_2$ are two areas, 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 under ROC curves. While comparing the performance 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 datasets, or different portions of the same dataset, 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 dataset, the two areas under ROC curves are correlated. They are correlated through matrix elements, and the matrix 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 similarity scores of two algorithms as well as the correlation between two sets of non-match similarity scores, respectively, and thereafter the correlation between two ROC curves.

As shown in the literature [14,15], 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 dataset. For example, if N equals 6,000, 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 similarity score to the match where two fingerprint images are more similar and a lower similarity score to the match where two fingerprint images are less similar. Such a characteristic of fingerprint data may cause higher positive correlation between two sets of match similarity scores of two algorithms as well as higher positive correlation between two sets of non-match similarity scores. On the other side of the coin, this higher correlation may be reduced due to the large magnitude of the size of the fingerprint datasets.

For four fingerprint-image matching algorithms investigated in this article, the six correlation coefficients between two sets of 6,000 match similarity scores range from 0.56 to 0.67. The size of non-match similarity score data is about 36,000,000. It is impractical to compute the Kendall's τ for this size of datasets, since its complexity is $O(N^2)$. Thus, the stochastic approach is invoked. 360,000 uniformly distributed random-sample data without replacement out of about 36,000,000 data are taken for each 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 shown in the reference [9], the six resultant correlation coefficients between two areas under ROC curves are from 0.17 through 0.24.

As shown in Table 1, as far as the value of area is concerned, Algorithm 1 is very close to Algorithm 2, and Algorithm 3 is quite close to Algorithm 4. However, the areas of Algorithms 1 and 2 are greater than the areas of Algorithms 3 and 4, respectively. To simplify the presentation, first, the two-tailed test is conducted. For these four fingerprint-image matching algorithms, the pairwise two-tailed p-values of two areas under ROC curves 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.

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

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

The two-tailed p-value between Algorithm1 and Algorithm 2 is 0.7714, which is much greater than 5%. According to the approach as shown in the article [9], the resultant correlation coefficient between two areas under ROC curves cannot be greater than the largest one of the two correlation coefficients that are for the match similarity scores and the non-match similarity scores, respectively. For Algorithm 1 and Algorithm 2, the largest one is 0.60, which is the Kendall's τ between two sets of 6,000 match similarity scores. Conservatively, even if using 0.60 for the correlation coefficient, the two-tailed p-value between Algorithm 1 and Algorithm 2 is 0.6797, which is also much greater than 5%. This indicates that the difference between two areas under ROC curves for Algorithms 1 and 2, respectively, 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 at the 67.97% in conservative way.

The two-tailed p-value between Algorithm 3 and Algorithm 4 is 0.0862 that is greater than 5% by 3.62%. By the same token, the conservative two-tailed p-value for Algorithms 3 and 4 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 datasets are large, the Z statistic hypothesis test of using the areas under ROC curves along with their variances can be implemented. This provides a sound statistical ground for testing the significance of the differences between two areas under ROC curves that are constructed by using two different fingerprint-image matching algorithms. Hence, the performances of two algorithms can be evaluated concisely.

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 datasets, have no definite underlying distribution functions. These distributions vary very much from algorithm to algorithm. The Kolmogorov-Smirnov Test had been used to determine whether there was any relationship between the two probability distribution functions of match similarity scores as well as the two probability distribution functions of non-match similarity scores. This test indicated that no relationship existed. As a consequence, the nonparametric approach must be employed in the analysis of the fingerprint similarity matcher scores.

Although the size of fingerprint datasets is much larger than the size of the datasets that are dealt with in the medical practice, a precise ROC curve can still be realistically constructed without any assumption to the score threshold, 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 is equivalent to the Mann-Whitney statistic directly formed from the match and non-match similarity scores.

The area under the 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. Therefore, to evaluate the fingerprint-image matching algorithm, the performance of an ROC curve as a whole rather than the performance of an ROC curve at a specific point or within a chosen region can be taken into account. The examples presented in this article have shown that even if the performance of a part of an ROC curve in one region of the curve produces higher TAR values, this does not guarantee that the performance of a whole

ROC curve is better. What ultimately matters under some operational conditions is the ROC curve as a whole, not a part of it.

Thanks to the relation between the area under an ROC curve computed using the trapezoidal rule and the Mann-Whitney statistic directly formed by the match and non-match similarity scores on which the ROC curve is built, the variance of the area under an ROC curve can be obtained by calculating the variance of the Mann-Whitney statistic. In addition, 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 conservative way depending on how to handle the correlation coefficient.

The fingerprint datasets are large-size datasets, and even on the same dataset 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 computed by using two areas under ROC curves along with their variances offers a systematic way to detect the statistical significance of differences between two underlying ROC curves, namely, differences between two performances of fingerprint-image matching algorithms. The method investigated in this article provides the information on which algorithm produces better results. This method also provides information about whether the difference is real or just by chance. Further, The method quantifies how much of the difference is real or how much is due to chance.

The approach of analyzing the ROC curves using the area under the ROC curve 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 dataset. It can also be applied, for instance, to evaluating the relations 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.

Even though the size of the fingerprint dataset is very large, the approach discussed in this article can be implemented without any difficulty. The match similarity scores and the non-match

similarity scores as the original input, the discrete probability distribution functions of the match and non-match similarity scores, the ROC curve, and the area under the ROC curve along with its variance can all be easily coded using the explicit mathematical formulas presented in this article. Finally, the corresponding Z statistic and the p-values can be obtained. Furthermore, the computing power required is not large.

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, Academic Radiology 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, Medical 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, John Wiley and Sons, 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 Psych 12 (1975) 387-415.
13. G.E. Noether, Elements of nonparametric statistics, John Wiley and Sons, New York, 1967, pp. 31-32.
14. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical recipes in C++: the art of scientific computing, second ed., Cambridge University Press, New York, 2002, pp. 647-648.
15. B.L. van der Waerden, Mathematical Statistics, Springer-Verlag, Berlin, 1969, p. 274 and pp. 333-335.