NISTIR 7586

Hypothesis Test of Fingerprint-Image Matching Algorithms in Operational ROC Analysis

Jin Chu Wu
Alvin F. Martin
Raghu N. Kacker

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

NISTIR 7586

Hypothesis Test of Fingerprint-Image Matching Algorithms in Operational ROC Analysis

Jin Chu Wu
Alvin F. Martin
Raghu N. Kacker

National Institute of Standards and Technology
Gaithersburg, MD 20899

June 2009

U.S. DEPARTMENT OF COMMERCE
Gary Locke, Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
Dr. Patrick D. Gallagher, Deputy Director

Hypothesis Test of Fingerprint-Image Matching Algorithms in Operational ROC Analysis

Jin Chu Wu[*a], Alvin F. Martin[a] and Raghu N. Kacker[b]
[a]Information Access Division, [b]Mathematical and Computational Sciences Division,
Information Technology Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

To evaluate the performance of fingerprint-image matching algorithms on large datasets, a receiver operating characteristic (ROC) curve is applied. From the operational perspective, the true accept rate (TAR) of the genuine scores at a specified false accept rate (FAR) of the impostor scores is usually employed. And the equal error rate (EER) can also be used. The accuracies of the measurement TAR and EER in terms of standard errors and 95 % confidence intervals can be computed using the nonparametric two-sample bootstrap based on our studies of bootstrap variability on large fingerprint datasets. In this article, the hypothesis testing is performed to determine whether the difference between the performance of one algorithm and a hypothesized value, or the difference between the performances of two algorithms where the correlation is taken into account is statistically significant. In the case that the alternative hypothesis is accepted, the sign of the difference is employed to determine which is better than the other. Examples are provided.

Keywords: Receiver operating characteristic (ROC) curve; Fingerprint; Biometrics; Nonparametric bootstrap; Standard error; Confidence interval; Significance test; Comparison

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

1. Introduction

To evaluate the performances of algorithms for fingerprint technologies in particular, and for biometrics in general, a receiver operating characteristic (ROC) curve is used as an important tool. In the applications of analyzing fingerprint data on large data sets, comparing two different fingerprint images of the same subject generates a genuine score, and matching two fingerprint images of two different subjects creates an impostor score. Both scores may be referred to as similarity scores in this article. Different fingerprint-image matching algorithms may invoke different kinds of scoring systems, such as integers, or real numbers in different ranges. However, they can all be converted to integers for data analysis [1].

These two sets of similarity scores constitute two distributions, respectively. The cumulative probabilities of genuine scores and impostor scores from the highest similarity score to a specified similarity score (i.e., the threshold score) are defined as the true accept rate (TAR) and the false accept rate (FAR), respectively. As is well-known, the two rates, namely, 1 – TAR (i.e., the type I error) and FAR (i.e., the type II error), are traded off of each other. When these two rates are equal, such a rate is defined as the equal error rate (EER). As the threshold moves from the highest similarity score down to the lowest similarity score, an ROC curve is then constructed in the FAR-and-TAR coordinate system. All these are illustrated in Figure 1 (A) and (B) assuming that the distributions of similarity scores are continuous.

While analyzing fingerprint data, as stated above, the distribution functions explored are all discrete probability distribution functions and thus an ROC curve is no longer a smooth curve. As opposed to continuous distribution, some concepts and definitions need to be established and modified accordingly [2, 3]. For instance, first, the ties of genuine scores and/or impostor scores at a threshold can often occur on large fingerprint datasets and thus must be taken into account while computing the estimated TAR at a specified FAR. Second, while computing the cumulative discrete probability at a score, the probability at this score must be taken into account [4]. Third, it seems that generally speaking there does not exist such a similarity score (range) at which the type I error is exactly equal to the type II error.

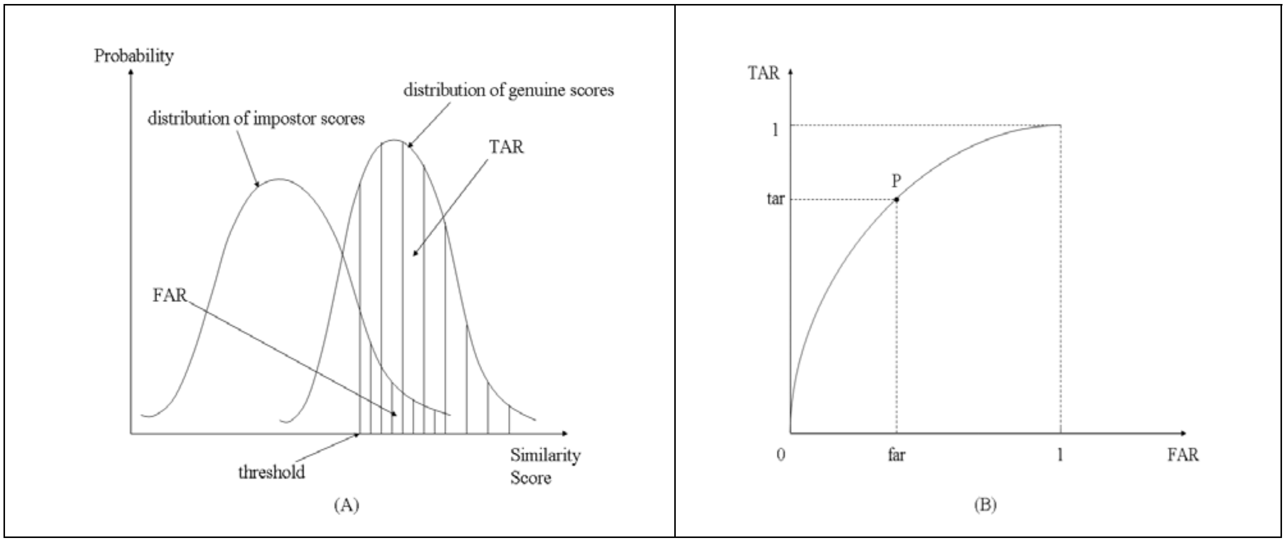

Figure 1. (A): A schematic diagram of distributions of continuous genuine scores and impostor scores, showing three related variables: TAR, FAR, and threshold. (B): A schematic drawing of an ROC curve constructed by moving the threshold from the highest similarity score down to the lowest similarity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

Our previous studies suggest that the nonparametric analysis be pertinent to evaluating fingerprint-image matching algorithms on large-scale datasets [1]. It was revealed that first, there is usually no underlying parametric distribution function for genuine scores and impostor scores; second, the distribution of genuine scores and the distribution of impostor scores are considerably different in general; and third, the distributions vary substantially from algorithm to algorithm in a way that differentiates algorithms in terms of matching accuracy. Hence, the empirical distribution is assumed for each of the observed similarity scores.

An ROC curve can be measured by computing the area under the ROC curve [1, and references therein]. Regarding this metric, first, the area under an ROC curve is equal to the probability of correctly identifying which is more likely than the other in the two stimuli under investigation and measures the overall ROC curve. Second, if using the trapezoidal rule, this area is equivalent to the Mann-Whitney statistic that is formed by genuine and impostor scores. Hence, the variance of the Mann-Whitney statistic can be utilized as the variance of the area under an ROC curve. Third, because the Mann-Whitney statistic is asymptotically normally distributed regardless of the distributions of genuine and impostor scores thanks to the Central Limit Theorem, 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.

From the operational perspective, at any point on an ROC curve in combination with two distributions of genuine scores and impostor scores, the three variables, which are TAR, FAR, and threshold score, are related to each other. Any one of these three variables can determine the other two variables. In an ROC analysis of fingerprint-image matching algorithms, three scenarios are of interest. They are 1) the measures and accuracies of TAR and threshold score at a specified FAR, 2) the measures and accuracies of TAR and FAR at a given threshold score, and 3) the measures and accuracies of the EER and the corresponding threshold score. Indeed, these three scenarios exhaust all possible cases [2, 3].

The accuracies of all these measures in terms of standard errors (SE) and 95 % confidence intervals (CI) were investigated in our previous work using the nonparametric two-sample bootstrap based on our studies of bootstrap variability on large fingerprint datasets [2-3, 5-7]. The two samples are a set of genuine scores and a set of impostor scores. Their corresponding distributions are interrelated with each other by the fingerprint-image matching algorithm that generates them. All statistics of interest are influenced under the combined impact of these two samples. In this article, for each algorithm, the total numbers of genuine scores and impostor scores are fixed, which are a little over 60 000 and 120 000, respectively [8]. With this amount of similarity scores, the FAR was set to be 0.001 while dealing with issues in Scenario 1 [9].

Given measurement accuracies, the next issue is comparisons. Here are two categories. The first category is the one-algorithm significance test, which is to determine whether the difference between the performance of one fingerprint-image matching algorithm and a hypothesized value is real or by chance. The second category is the two-algorithm significance test, which is to investigate whether the difference between the performances of two algorithms is statistically significant. In this respect, the metric TAR at a given FAR (i.e., Scenario 1) and/or the metric EER (i.e., Scenario 3) are typically employed.

In some applications, it is of interest to determine if the matching accuracies derived from two different samples of data are statistically different. Indeed, this case does belong to the second category, in which the performances of two different algorithms on the same dataset are replaced by the performances of a single algorithm on two different datasets. As will be shown in Section 2.2, all methods remain the same.

In the case where TAR and FAR are determined by a given threshold score (i.e., Scenario 2), there are two measures, i.e., TAR and FAR, for each algorithm. Different algorithms may invoke different threshold scores to generate TAR and FAR. As is well-known, larger TAR is better and smaller FAR is better. Thus it is hard to compare algorithms using these two metrics simultaneously. For instance, when the TAR of Algorithm A is greater than the TAR of Algorithm B, and in the meantime the FAR of Algorithm A is also greater than the FAR of Algorithm B, it is hard to determine which algorithm is better.

Such comparison issues can be dealt with intuitively to some extent using 95 % CIs. For one-algorithm significance test, an accuracy criterion value μ_o of the statistic of interest is set first, and then whether a fingerprint-image matching algorithm passes or fails the test is determined based on whether its performance is better than μ_o or not. Thus, if the 95 % CI of the estimator of the statistic of interest TAR (or EER) for an algorithm is above (or below) μ_o, this algorithm passes the test with at least 95 % confidence level. Otherwise, the algorithm fails the test. It can be proven that such a confidence interval approach is backed by the hypothesis testing. But this approach cannot provide the quantitative information of how much the p-value is, i.e., what the statistical significance of the difference is.

For two-algorithm significance test, for instance, if the 95 % CI of the TAR for a given FAR of one algorithm is exclusively higher than the 95 % CI of the TAR for the same FAR of another algorithm, then it is trivial to reach the conclusion that the performance of the former algorithm is better than the performance of the latter algorithm. This approach cannot offer the quantitative information regarding the statistical significance of the difference either. Further, what if the 95 % CIs of two algorithms overlap? Thus, the issue of determining whether the difference is real or by chance must be dealt with using the statistical hypothesis testing.

One possible approach is to mimic the hypothesis testing with the bootstrap but in the context of two fingerprint-image matching algorithms, each of which generates two distributions, respectively [10]. Suppose that the statistic of interest is TAR while FAR is specified, and that Algorithm 1 generates N_{G1} genuine scores and N_{I1} impostor scores as well as TAR1 thereafter, and Algorithm 2 generates N_{G2} genuine scores and N_{I2} impostor scores as well as TAR2 thereafter. The genuine scores of the two algorithms are combined to constitute a genuine score set ***G*** with $N_{G1} + N_{G2}$ scores, and the impostor scores of the two algorithms are combined to form an impostor score set ***I*** with $N_{I1} + N_{I2}$ scores.

In each of bootstrap Monte Carlo iterations, N_{G1} and N_{G2} genuine scores are randomly selected with replacement (WR) from the genuine score set ***G*** to form two bootstrap sets ***G1*** and ***G2***, respectively, and N_{I1} and N_{I2} impostor scores are randomly selected WR from the impostor score set ***I*** to form two bootstrap sets ***I1*** and ***I2***, respectively. Then, the bootstrap genuine score set ***G1*** and the bootstrap impostor score set ***I1*** generates a bootstrap replication *TAR1*, and the bootstrap genuine score set ***G2*** and the bootstrap impostor score set ***I2*** creates a bootstrap replication *TAR2*. Finally, from the relationship between the observed value of the difference TAR1 – TAR2 and the distribution of *TAR1 – TAR2*, a hypothesis testing could be carried out.

Different algorithms may employ different scoring systems. If two algorithms invoke two different scoring systems, then before applying this approach, the scores must be converted to integer scores. After score conversion, different algorithms may have very different integer score ranges for genuine scores and impostor scores, respectively. After a large number of sampling, *TAR1* and *TAR2* may always be quite close and considerably differ from TAR1 and TAR2, respectively. This was observed in our testing; therefore, this approach may not work for the hypothesis testing on fingerprint-image matching algorithms.

Another possible approach is to examine the relationship between two distributions formed by bootstrap replications of the TAR for two algorithms, respectively. In such an approach, the statistic of interest is switched from the TAR of the algorithm to whatever the statistic is, such as mean, median, etc., which is computed from the bootstrap replications of TARs. Moreover, the number of bootstrap replications was determined to be 2000 in order to reduce the bootstrap variance based on our prior bootstrap variability studies [2]. For such a large number of data points, generally speaking, a difference of means, even though it is small compared to the standard deviation, could be very significant [11]. Thus, this approach is also inappropriate. The same assertion holds true for EER as well.

In our applications, the statistics of interest are both TAR at a given FAR and EER. For these statistics of interest, the relationship between two types of 95 % CIs was examined in various cases in our previous studies [2, 3]. One type of 95 % CI was computed using the definition of quantile; another type of 95 % CI was calculated if the distribution of 2000 bootstrap replications of the statistic is assumed to be normal. It was found that these two types of 95 % CIs were approximately matched up to the fourth decimal place for relatively high-accuracy algorithms and the third decimal place for relatively low-accuracy algorithms. Moreover, the Shapiro-Wilk normality test [12] was conducted on the 2000 bootstrap replications of the statistics of interest for different algorithms, and it was observed that the majority of p-values are greater than 5 %, especially for relatively high-accuracy algorithms.

All these suggest that the statistics of interest in our applications be assumed to be normally distributed regardless of the distributions of genuine and impostor scores. Under such a normality assumption, the straightforward way to test the significance of the difference between one algorithm and a hypothesized value or between two algorithms is the Z-test. In analogy to the area under an ROC curve [1], the Z statistic can be formulated using estimators of statistics of interest of one algorithm or two algorithms along with their variances and the correlation coefficient, and it is subject to the standard normal distribution with zero expectation and a variance of one.

The methods of computing the standard errors of the statistics of interest in our applications were explored in our previous studies [2, 3]. The statistics of interest of two fingerprint-image matching algorithms may or may not be correlated, depending on how the sets of similarity scores are generated. In our applications, the sets of similarity scores were generated in a way that might cause the correlation between the two statistics of interest. Thus, the method of calculating the related correlation coefficients in our case will be provided in this article.

In reference [13], the false non-match rate (FNMR) for a given FAR was employed as a metric to evaluate the fingerprint technologies. FNMR, i.e., the type I error, is equal to 1 – TAR. It is trivial to prove that the standard errors of TAR and FNMR for an algorithm are equal, the correlation coefficients of TAR and FNMR given two algorithms are also the same, and so are the Z scores and the p-values of TAR and FNMR for two algorithms. The only difference is that the upper (or the lower) bound of 95 % CI of FNMR is one minus the lower (or the upper) bound of 95 % CI of TAR [2, 3]. As a result, every method for TAR stated in this article can be applied to FNMR, and every result obtained for TAR holds true for FNMR.

The general formulas of hypothesis testing for one fingerprint-image matching algorithm and two algorithms are presented in Section 2. An algorithm for computing the correlation coefficient in our applications is provided in Section 3. The results of examples involving six fingerprint-image matching algorithms[1] are shown in Section 4. Finally the conclusion and discussion is found in Section 5.

2. Hypothesis testing [4]

As pointed out in Section 1, the hypothesis testing is performed in two categories: one-algorithm hypothesis testing and two-algorithm hypothesis testing; the statistics of interest from the operational perspective are in two scenarios: the metric TAR at a given FAR and the metric EER. There is no reason to believe *a priori* that the performance of one algorithm is likely to be better than a hypothesized value or the performance of the other algorithm. Further, the two-tailed test is generally more conservative than the one-tailed test in the sense that the former is more difficult to reject the null hypothesis for a given significance level [14]. Thus, the two-tailed test is invoked in this article. In the case that the alternative hypothesis is accepted, the

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

sign of the difference between the estimate and a hypothesized value or the two estimates is employed to determine which is better than the other.

2.1 One-algorithm hypothesis testing

Let $\hat{\text{STI}}$ denote the estimator of the statistic of interest of an algorithm and μ_o denote the hypothesized value. Then, the null and alternative hypotheses are

$$\begin{aligned} H_o &: \hat{\text{STI}} = \mu_o \\ H_a &: \hat{\text{STI}} \neq \mu_o \end{aligned} \qquad (1)$$

Assume that the statistic of interest in our case is normally distributed regardless of the distributions of genuine and impostor scores, as pointed out in Section 1. Then, the Z statistic

$$Z = \frac{\hat{\text{STI}} - \mu_o}{\text{SE}\ (\hat{\text{STI}})} \qquad (2)$$

where SE $(\hat{\text{STI}})$ stands for the standard error of the statistic of interest STI, is subject to the standard normal distribution with zero expectation and a variance of one. The standard error can be computed using the nonparametric two-sample bootstrap [2, 3].

While evaluating the performance of an algorithm with respect to an accuracy criterion value, besides p-value, other factors need to be taken into account as well, such as the characteristic of the statistic of interest (the larger the better or the smaller the better) and the sign of the difference between the estimator and the accuracy criterion value. For instance, if the statistic of interest is TAR (the larger the better) and its estimator is less than μ_o, then less-than-5 % p-value indicates that this algorithm fails the test.

2.2 Two-algorithm hypothesis testing

Let $\hat{\text{STI}}_1$ and $\hat{\text{STI}}_2$ denote the estimators of the statistic of interest of two algorithms, respectively. As mentioned in Section 1, $\hat{\text{STI}}_1$ and $\hat{\text{STI}}_2$ may stand for the estimators of the statistic of interest of a single algorithm performing on two different datasets, respectively. Then, the null and alternative hypotheses are

$$\begin{aligned} H_o &: \hat{\mathrm{STI}}_1 = \hat{\mathrm{STI}}_2 \\ H_a &: \hat{\mathrm{STI}}_1 \neq \hat{\mathrm{STI}}_2 \end{aligned} \qquad (3)$$

Assume that the statistic of interest in our case is normally distributed regardless of the distributions of genuine and impostor scores, as pointed out in Section 1. Then, the Z statistic can be expressed as

$$Z = \frac{\hat{\mathrm{STI}}_1 - \hat{\mathrm{STI}}_2}{\sqrt{\mathrm{SE}^2\,(\hat{\mathrm{STI}}_1) + \mathrm{SE}^2\,(\hat{\mathrm{STI}}_2) - 2\,\mathrm{r}\,\mathrm{SE}\,(\hat{\mathrm{STI}}_1)\,\mathrm{SE}\,(\hat{\mathrm{STI}}_2)}} \qquad (4)$$

where SE $(\hat{\mathrm{STI}}_1)$ and SE $(\hat{\mathrm{STI}}_2)$ are two standard errors of $\mathrm{ST1}_1$ and STI_2, respectively, and r stands for the correlation coefficient between $\mathrm{ST1}_1$ and STI_2. It is subject to the standard normal distribution with zero expectation and a variance of one.

These standard errors can be computed using the nonparametric two-sample bootstrap [2, 3]. If the two statistics of interest are correlated and the correlation coefficient r in Eq. (4) is not taken into account, it can leave the denominator of Eq. (4) larger and Z smaller; thereby reduce the chance of detecting a difference between the performances of two algorithms.

3. An algorithm for computing the correlation coefficient

The two statistics of interest of any two algorithms may or may not be correlated, depending on how the sets of similarity scores are created. If two sets of genuine scores and/or two sets of impostor scores that construct the two ROC curves, respectively, do not co-vary, then the two statistics of interest of two algorithms are not correlated.

In our tests, different fingerprint-image matching algorithms generated different sets of similarity scores, respectively, using the same set of fingerprint images. In other words, for the two sets of similarity scores created by two algorithms, respectively, any two similarity scores with the same ordinal number of entry were generated using the same two images. Thus, they co-vary. By comparing two matches of fingerprint images, i.e., two data entries in each set of similarity scores, all 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 positive correlation between two sets of similarity scores of two algorithms. Subsequently, this correlation of similarity scores may eventually result in the correlation between the statistics of interest of two algorithms. On the other hand, this correlation may be reduced due to the large magnitude of the size of the fingerprint datasets.

The genuine score sets of fingerprint-image matching Algorithms A and B are denoted as

$$\mathbf{G^i} = \{\, m^i_j \mid i \in \{ A, B \} \text{ and } j = 1, \ldots, N_G \} , \tag{5}$$

and the impostor score sets of Algorithms A and B are expressed as

$$\mathbf{I^i} = \{\, n^i_j \mid i \in \{ A, B \} \text{ and } j = 1, \ldots, N_I \} , \tag{6}$$

where N_G and N_I are the total numbers of genuine scores and impostor scores, respectively. It is assumed that Algorithms A and B generate the same amount of genuine scores as well as impostor scores. As stated above, the two j-th genuine scores m^i_j where $i \in \{ A, B \}$ are generated using the same two images but employing different algorithms. So do the two j-th impostor scores n^i_j where $i \in \{ A, B \}$.

The statistics of interest are both TAR at a given FAR and EER. An algorithm for computing the correlation coefficient of the statistic of interest ST of Algorithms A and B in our applications is as follows.

Algorithm

```
for $i = 1$ to M do
    Synchronized_WR_Random_Sampling ($N_G$, $\mathbf{G^A}$, $\mathbf{\Theta^A_i}$, $\mathbf{G^B}$, $\mathbf{\Theta^B_i}$)
    Synchronized_WR_Random_Sampling ($N_I$, $\mathbf{I^A}$, $\mathbf{\Xi^A_i}$, $\mathbf{I^B}$, $\mathbf{\Xi^B_i}$)
    the new genuine score set $\mathbf{\Theta^A_i}$ and the new impostor score set $\mathbf{\Xi^A_i}$ => statistic $\hat{ST}^A_i$
    the new genuine score set $\mathbf{\Theta^B_i}$ and the new impostor score set $\mathbf{\Xi^B_i}$ => statistic $\hat{ST}^B_i$
end for
$\{ \hat{ST}^A_i \mid i = 1,\ldots,M \}$ and $\{ \hat{ST}^B_i \mid i = 1,\ldots,M \}$ => the correlation coefficient $r^{AB}_{ST}$
end
function Synchronized_WR_Random_Sampling (N, $\mathbf{S^A}$, $\mathbf{\Gamma^A}$, $\mathbf{S^B}$, $\mathbf{\Gamma^B}$)
```

```
1.2: for $j = 1$ to N do
1.3:     randomly select WR an index $k \in \{ 1, \ldots, N \}$
1.4:     $\gamma^{A}_{j} = s^{A}_{k}$
1.5:     $\gamma^{B}_{j} = s^{B}_{k}$
1.6: end for
1.7: end function
```

where s^{A}_{k}, γ^{A}_{j}, s^{B}_{k}, and γ^{B}_{j} are members of the sets $\mathbf{S}^{\mathbf{A}}$, $\mathbf{\Gamma}^{\mathbf{A}}$, $\mathbf{S}^{\mathbf{B}}$, and $\mathbf{\Gamma}^{\mathbf{B}}$, respectively. Based on our previous bootstrap variability studies [2], the number of iterations M is set to be 2000 in our applications.

As shown from Step 1 to 6, this algorithm runs M iterations. As indicated in Steps 2 and 3, in the i-th iteration, the synchronized WR random sampling is carried out on two genuine score sets $\mathbf{G}^{\mathbf{A}}$ and $\mathbf{G}^{\mathbf{B}}$ to generate two new genuine score sets $\mathbf{\Theta}^{\mathbf{A}}_{\mathbf{i}}$ and $\mathbf{\Theta}^{\mathbf{B}}_{\mathbf{i}}$, respectively, as well as on two impostor score sets $\mathbf{I}^{\mathbf{A}}$ and $\mathbf{I}^{\mathbf{B}}$ to create two new impostor score sets $\mathbf{\Xi}^{\mathbf{A}}_{\mathbf{i}}$ and $\mathbf{\Xi}^{\mathbf{B}}_{\mathbf{i}}$, respectively.

As shown from Step 1.1 to 1.7, the function, Synchronized_WR_Random_Sampling, runs N iterations, where N is the total number of similarity scores (genuine or impostor scores). As indicated in Step 1.3, in the j-th iteration, an index k is randomly drawn WR from the integer set $\{ 1, \ldots, N \}$. Then as indicated in Steps 1.4 and 1.5, the k-th score of the input score set $\mathbf{S}^{\mathbf{A}}$ is assigned to the j-th score of the new score set $\mathbf{\Gamma}^{\mathbf{A}}$, and the k-th score (i.e., synchronized) of the input score set $\mathbf{S}^{\mathbf{B}}$ is also assigned to the j-th score of the new score set $\mathbf{\Gamma}^{\mathbf{B}}$. With such synchronized random sampling, the co-varying similarity scores (i.e., with the same ordinal number of data entry) between Algorithms A and B are selected simultaneously, and the correlation in the similarity scores between two algorithms is preserved if there is any.

As shown in Steps 4 and 5, after the synchronized WR random sampling, the i-th estimated statistic $S\hat{T}^{A}_{i}$ of Algorithm A is computed from the new genuine score set $\mathbf{\Theta}^{\mathbf{A}}_{\mathbf{i}}$ and the new impostor score set $\mathbf{\Xi}^{\mathbf{A}}_{\mathbf{i}}$, and the i-th estimated statistic $S\hat{T}^{B}_{i}$ of Algorithm B is calculated from the new genuine score set $\mathbf{\Theta}^{\mathbf{B}}_{\mathbf{i}}$ and the new impostor score set $\mathbf{\Xi}^{\mathbf{B}}_{\mathbf{i}}$. After M iterations, the correlation coefficient r^{AB}_{ST} of the statistic of interest ST of Algorithms A and B can be calculated from the two sets of estimated statistics of interest, as indicated in Step 7.

This algorithm involves a synchronized random sampling. Thus, it is a stochastic process. In Practice, if the p-value is not considerably different from the critical values, such as 5 %, 1 %, etc., then in order to reduce the computational fluctuation this algorithm needs to run for several times (e.g., ten, etc.). The average correlation coefficient out of these correlation coefficients is taken to be the resultant correlation coefficient for significance test. In the test shown in this article, this algorithm was always run 10 times in order to obtain the resultant correlation coefficient.

4. Results

Six fingerprint-image matching algorithms were taken to be examples.[2] Among them, Algorithms 1 through 3 were of relatively high accuracy and Algorithms 4 through 6 were of relatively low accuracy. These algorithms used different types of scoring systems. Algorithms 1 through 3 and Algorithm 6 all employed integers but in different ranges of integers. Algorithm 4 invoked less-than-1 real numbers and Algorithm 5 used less-than-100 real numbers. All similarity scores used by vendors were nonnegative.

Algorithms 1 through 3 were taken as examples for both one-algorithm hypothesis testing and two-algorithm hypothesis testing while the statistic of interest was assumed to be TAR at a specified FAR. Algorithms 4 through 6 were used only for two-algorithm significance test while the statistic of interest was set to be EER. The method applied to TAR can be applied to EER, and vice versa. The only difference is that for TAR the larger the better, but for EER the smaller the better.

The estimated $\hat{TAR}$ (f) at a given FAR and the estimated $\hat{EER}$ along with their accuracies in terms of SE and 95 % CI can be computed using the nonparametric two-sample bootstrap with 2000 bootstrap replications [2, 3]. The estimates of TARs, SEs, and 95 % CIs for relatively high-accuracy Algorithms 1 through 3 are shown in Table 1, and the estimates of EERs, SEs, and 95 % CIs for relatively low-accuracy Algorithms 4 through 6 are presented in Table 2.

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

Algorithm	TÂR (f)	SÊ	95 % Confidence interval
1	0.994322	0.000324	(0.993662, 0.994918)
2	0.993255	0.000325	(0.992622, 0.993922)
3	0.989263	0.000470	(0.988307, 0.990159)

Table 1. The estimates of TARs, SEs, and 95 % CIs for relatively high-accuracy Algorithms 1 through 3, while FAR was specified at 0.001.

Algorithm	EÊR	SÊ	95 % Confidence interval
4	0.012409	0.000378	(0.011638, 0.013148)
5	0.012903	0.000360	(0.012205, 0.013609)
6	0.013634	0.000338	(0.012940, 0.014287)

Table 2. The estimates of EERs, SEs, and 95 % CIs for relatively low-accuracy Algorithms 4 through 6.

4.1 One-algorithm hypothesis testing

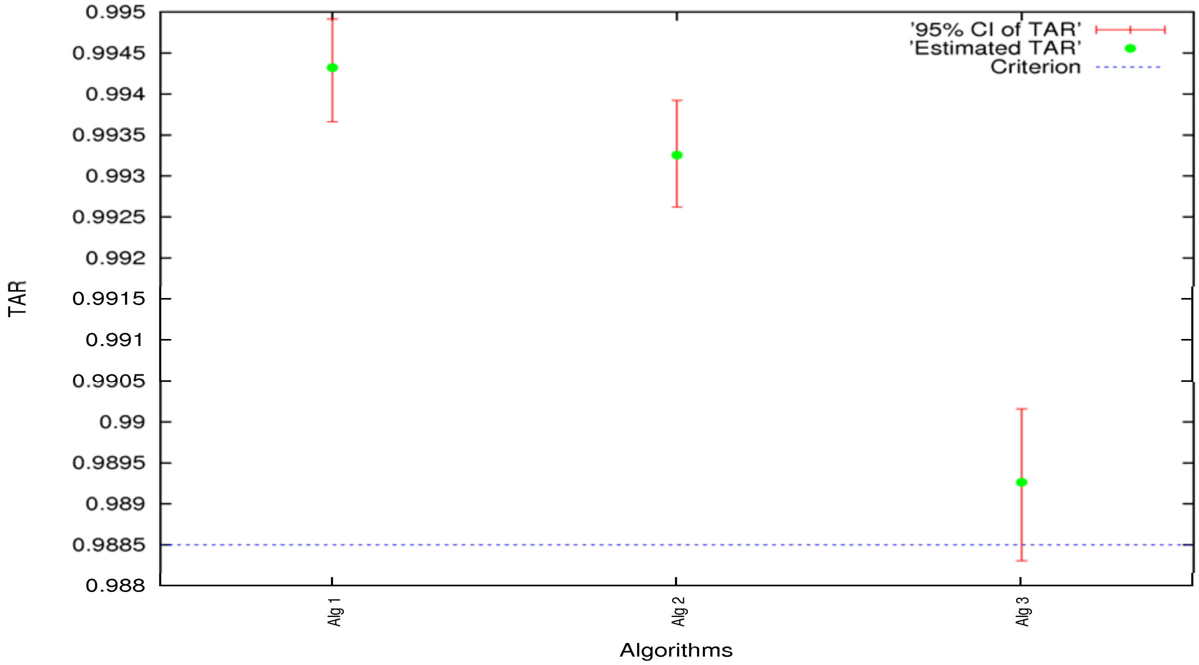

Figure 2. The estimates of TARs and the corresponding 95 % CIs for relatively high-accuracy Algorithms 1 through 3, while FAR was specified at 0.001, along with the hypothesized value μ_o to be set at 0.988500.

For one-algorithm hypothesis testing, the estimate of the statistic of interest of an algorithm is compared against a hypothesized value to see whether the difference is real or by chance. Suppose that the metric TAR at a given FAR is considered in the testing and the hypothesized value μ_o is set to be 0.988500.

In Figure 2 are drawn the estimates of TARs at FAR 0.001 and the corresponding 95 % CIs for relatively high-accuracy Algorithms 1 through 3. The hypothesized value μ_o = 0.988500 is also shown in Figure 2. The two 95 % CIs of Algorithms 1 and 2 are way above 0.988500. Thus, the performances of Algorithms 1 and 2 measured by the metric TAR are better than the accuracy criterion value 0.988500. This observation is supported by applying Eq. (2). Using the corresponding estimated TARs and SEs for these two algorithms from Table 1, it was found that the two two-tailed p-values were all equal to 0.0000 as presented in Table 3. This indicates that the alternative hypothesis $H_a : \hat{\mathrm{STI}} \neq \mu_o$ is very strongly accepted. Further, with the positive sign of the difference between the estimated TAR and the hypothesized value 0.988500, it is concluded that the TÂR (f) of Algorithm 1 and Algorithm 2 are all very significantly greater than the accuracy criterion value 0.988500. In other words, Algorithms 1 and 2 pass the test.

Algorithms	p-value
1	0.0000
2	0.0000
3	0.1049

Table 3. The two-tailed p-values of the statistic of interest TAR with respect to the hypothesized value μ_o = 0.988500 for relatively high-accuracy Algorithms 1 through 3.

In Figure 2, the horizontal line of the hypothesized value μ_o = 0.988500 intersects the 95 % CI of Algorithm 3. After using Eq. (2) by substituting the estimates of TAR and SE for Algorithm 3 from Table 1, it was found that the two-tailed p-value was 0.1049 as presented in Table 3, which is greater than 5 %. This suggests that the null hypothesis $H_o : \hat{\mathrm{STI}} = \mu_o$ be accepted. That is to say, the difference between the estimator TÂR (f) = 0.989263 of Algorithm 3 and the hypothesized value 0.988500 is not real but by chance at the significance level 10 %. Therefore, Algorithm 3 fails the test, if the performance is required to be better than the accuracy criterion value μ_o.

4.2 Two-algorithm hypothesis testing

The hypothesis testing for two algorithms is not as straightforward as the one for a single algorithm. It cannot be judged merely using the confidence interval approach. In order to determine whether the difference between the performances of two fingerprint-image matching algorithms is significant, the two-algorithm hypothesis testing must be carried out.

4.2.1 TAR for a given FAR

The estimates of TARs at FAR 0.001 and the corresponding 95 % CIs for relatively high-accuracy Algorithms 1 through 3 are listed in Table 1 and drawn in Figure 2. The 95 % CI of Algorithm 1 slightly overlaps the one of Algorithm 2. But both of them are above the 95 % CI of Algorithm 3. What is the significance of the differences among the performances of these three algorithms?

The correlation coefficient of the statistic of interest TAR between two matching algorithms can be computed using the algorithm as presented in Section 3. For relatively high-accuracy Algorithms 1 to 3, the average correlation coefficients out of ten runs are listed in Table 4. The positive correlation coefficients for TARs are near 0.5. This indicates that all high-accuracy fingerprint-image matching algorithms have the same tendency to assign higher (lower) similarity scores to the matching results of more (less) similar images.

For relatively low-accuracy Algorithms 4 to 6, the average correlation coefficients of the statistic of interest TAR out of ten runs are 0.223933, 0.240295, and 0.266922, respectively. The positive correlation coefficients for relatively low-accuracy algorithms are not as high as those for the high-accuracy algorithms. It is expected that the tendency of assigning higher (lower) similarity scores to the matching results of more (less) similar images for relatively low-accuracy algorithms is not as strong as the tendency for high-accuracy algorithms. Thus, these results provide evidence that the synchronized algorithm for computing the correlation coefficient is quite reasonable.

Algorithms	1	2	3
1	1.000000	0.496089	0.454423
2		1.000000	0.493979
3			1.000000

Table 4. The average correlation coefficients of the statistic of interest TAR out of ten runs among relatively high-accuracy Algorithms 1 through 3.

Algorithms	1	2	3
1	1.0000	0.0011	0.0000
2		1.0000	0.0000
3			1.0000

Table 5. The two-tailed p-values of two statistics of interest TARs for relatively high-accuracy Algorithms 1 through 3, where the correlation coefficient was taken into account.

After applying Eq. (4) using the estimates of TARs and SEs from Table 1 and the corresponding correlation coefficients from Table 4, the two-tailed p-values of two statistics of interest TARs for relatively high-accuracy Algorithms 1 through 3 can be computed and are shown in Table 5. The two-tailed p-value between Algorithms 1 and 2 is 0.0011, and other two are 0.0000.

These two-tailed p-values are all much less than 5 %. It suggests that the alternative hypothesis H_a : $\hat{\mathrm{STI}}_1 \neq \hat{\mathrm{STI}}_2$ be strongly accepted. In other words, the differences of performances among Algorithms 1 through 3 are very significant, even though the 95 % CI of Algorithm 1 does slightly overlap the one of Algorithm 2. It follows from the sign of the difference between the two corresponding estimated TARs that the performance of Algorithm 1 is better than the performance of Algorithm 2; and the performances of both of them are better than the performance of Algorithm 3.

4.2.2 EER

The estimates of EERs and the corresponding 95 % CIs for relatively low-accuracy Algorithms 4 through 6 are provided in Table 2 and drawn in Figure 3. The 95 % CIs of the three algorithms mutually overlap. In such a circumstance, how can the statistical significance of the differences among the performances of these three algorithms be determined?

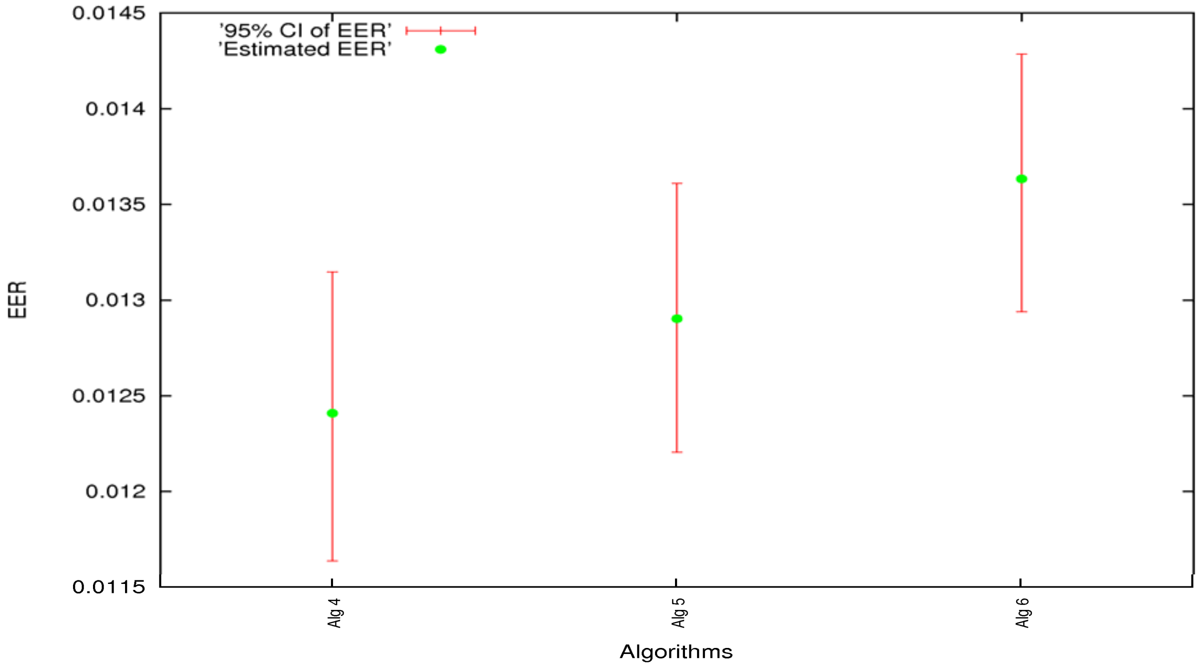

Figure 3. The estimates of EERs and the corresponding 95 % CIs for relatively low-accuracy Algorithms 4 through 6.

The correlation coefficient of the statistic of interest EER can be calculated using the algorithm as shown in Section 3. For relatively low-accuracy Algorithms 4 through 6, the average correlation coefficients out of ten runs are presented in Table 6. For high-accuracy Algorithms 1 to 3, the average correlation coefficients of the statistic of interest EER out of ten runs are 0.513037, 0.529609, and 0.567842, respectively. They are all larger than those for relatively low-accuracy Algorithms 4 through 6. This is expected as discussed in Subsection 4.2.1, and also supports the synchronized algorithm for computing the correlation coefficient.

Algorithms	4	5	6
4	1.000000	0.360888	0.398198
5		1.000000	0.453439
6			1.000000

Table 6. The average correlation coefficients of the statistic of interest EER out of ten runs among relatively low-accuracy Algorithms 4 through 6.

Using Eq. (4) by substituting the estimates of EERs and SEs from Table 2 and the corresponding correlation coefficients from Table 6, the two-tailed p-values of two statistics of interest EERs for relatively low-accuracy Algorithms 4 through 6 can be calculated and are presented in Table 7.

Algorithms	4	5	6
4	1.0000	0.2370	0.0019
5		1.0000	0.0457
6			1.0000

Table 7. The two-tailed p-values of two statistics of interest EERs for relatively low-accuracy Algorithms 4 through 6, where the correlation coefficient was taken into account.

The two-tailed p-value between Algorithms 4 and 5 is 0.2370, which is much greater than 5 %. It suggests that the null hypothesis H_o : $\hat{\mathrm{STI}}_1 = \hat{\mathrm{STI}}_2$ be accepted. In other words, the difference between the performances of Algorithms 4 and 5 is by chance, i.e., not statistically significant, even though the estimated $\hat{\mathrm{EER}}$ 0.012409 of Algorithm 4 is lower than the estimated $\hat{\mathrm{EER}}$ 0.012903 of Algorithm 5. To some extent, this conclusion is supported by the fact that the 95 % CI of Algorithm 4 heavily overlaps the one of Algorithm 5, as illustrated in Figure 3.

The two-tailed p-value between Algorithms 5 and 6 is 0.0457. Without considering the correlation coefficient, it increases to 0.1392. As pointed out in Subsection 2.2, neglecting the correlation coefficient can reduce the chance of detecting a difference between the performances of two algorithms. Since 0.0457 is slightly less than 5 %, the alternative hypothesis H_a : $\hat{\mathrm{STI}}_1 \neq$

$\hat{\mathrm{STI}}_2$ is accepted with reasonably strong evidence, despite that the 95 % CI of Algorithm 5 quite overlaps the 95 % CI of Algorithm 6. Further due to the sign of the difference between the two estimated EERs, the performance of Algorithm 5 is reasonably better than the performance of Algorithm 6.

The two-tailed p-value between Algorithms 4 and 6 is 0.0019, which is less than 5 %. It suggests that the alternative hypothesis $H_a : \hat{\mathrm{STI}}_1 \neq \hat{\mathrm{STI}}_2$ be strongly accepted, although the 95 % CIs of these two algorithms slightly overlap. Moreover, because of the sign of the difference between the two estimated EERs, the performance of Algorithm 4 is considerably better than the performance of Algorithm 6.

In addition, the p-value 0.0019 between Algorithms 4 and 6 is much smaller than the p-value 0.0457 between Algorithms 5 and 6. It indicates that the difference between the performances of Algorithms 4 and 6 is more statistically significant than the difference between the performances of Algorithms 5 and 6. To some extent, this conclusion can be supported by the relationship among the 95 % CIs of Algorithms 4 to 6 as illustrated in Figure 3.

5. Conclusion and discussion

In operational ROC analysis of fingerprint-image matching algorithms, it is very important to determine whether the difference between the performance of one algorithm and an accuracy criterion value, or the difference between the performances of two algorithms where the correlation is taken into account is statistically significant. In this regard, no study was found to date. For such comparison issues, the two statistics of interest TAR at a specified FAR and EER are typically employed.

These two statistics of interest can be assumed to be normally distributed regardless of the distributions of genuine scores and impostor scores. This assumption is supported by the matches in various cases between two types of 95 % CIs. One is computed using the definition of quantile, and the other is calculated if the distribution of 2000 bootstrap replications of the statistic of interest is assumed to be normal. It is also partly supported by the Shapiro-Wilk normality test.

Under the normality assumption, the Z-test can be applied. The Z statistic is formulated using the estimated TAR at a specified FAR or EER of one algorithm or two algorithms along with their variances and correlation coefficient, and it is subject to the standard normal distribution with zero expectation and a variance of one. All the standard errors can be computed using the nonparametric two-sample bootstrap with 2000 bootstrap replications based on our variability study of bootstraps.

In this article, an algorithm is provided to calculate the correlation coefficient between two statistics of interest of two fingerprint-image matching algorithms, under the assumption that for these two algorithms any two scores with the same ordinal number of entry in the two sets of similarity scores are generated using the same two images. Otherwise the user needs to provide the correlation coefficient, if they are correlated.

This algorithm is a stochastic process, since it involves a synchronized sampling. In Practice, if the p-value is not considerably different from the critical values, such as 5 %, 1 %, etc., then in order to reduce the computational fluctuation this algorithm needs to run for several times (ten in our case). The average correlation coefficient out of these correlation coefficients is taken to be the resultant correlation coefficient for significance test.

While comparing the performance of a fingerprint-image matching algorithm in terms of TAR at a specified FAR or EER against a criterion value, the confidence interval approach can be adopted. This approach is backed by the one-algorithm hypothesis testing. However, it cannot provide the information of what the statistical significance of the difference is. While comparing the performances of two algorithms, the hypothesis testing can resolve the relationship between two 95 % CIs. As presented in Subsection 4.2.1, although the 95 % CIs of Algorithms 1 and 2 did slightly overlap, the hypothesis testing showed that the difference of performances between these two algorithms was very statistically significant. And also as discussed in Subsection 4.2.2, all three 95 % CIs were mutually overlapped to a certain degree, but the hypothesis testing showed that the statistical significances of the differences in performances among the three algorithms were quite different accordingly. The issue of determining whether the difference between the performances of two algorithms is real or by chance must be dealt with using the significance test.

Conventionally, if the two-tailed p-value is greater than or equal to 5 %, the null hypothesis is accepted; if it is less than 5 %, the alternative hypothesis is accepted. In the literature [10], it suggested: If p-value is less than 0.10, borderline evidence is against H_o; if p-value is less than 0.05, reasonably strong evidence is against H_o; if p-value is less than 0.025, strong evidence is against H_o; if p-value is less than 0.01, very strong evidence is against H_o.

References

1. J.C. Wu, C.L. Wilson, Nonparametric analysis of fingerprint data on large data sets, Pattern Recognition 40 (9) (2007) 2574-2584.
2. J.C. Wu, Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap, NISTIR 7449, National Institute of Standards and Technology, September 2007.
3. J.C. Wu, Operational measures and accuracies of ROC Curve on large fingerprint data Sets, NISTIR 7495, National Institute of Standards and Technology, May 2008.
4. B. Ostle, L.C. Malone, Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed., Iowa State University Press, Ames, 1988.
5. B. Efron, Bootstrap methods: Another look at the Jackknife. Ann. Statistics, 7:1-26, 1979.
6. B. Efron, Better bootstrap confidence intervals, J. Amer. Statist. Assoc. 82 (397) (1987) 171-185.
7. A.C. Davison, D.V. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press, Cambridge, 2003.
8. J.C. Wu, C.L. Wilson, An empirical study of sample size in ROC-curve analysis of fingerprint data, in Biometric Technology for Human Identification III, Proceedings of SPIE Vol. 6202, 620207 (2006).

9. J.C. Wu, M.D. Garris, Nonparametric statistical data analysis of fingerprint minutiae exchange with two-finger fusion, in Biometric Technology for Human Identification IV, Proceedings of SPIE Vol. 6539, 65390N (2007).
10. B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.
11. 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.
12. R: A Language and Environment for Statistical Computing, The R Development Core Team, Version 2.8.0, 2008, at http://www.r-project.org/.
13. R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, A.K. Jain, Performance evaluation of fingerprint verification systems, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28 (1) 2006 3-18.
14. G.E.P. Box, J.S. Hunter, W.G. Hunter, Statistics for experimenters: design, innovation, and discovery, second ed., John Wiley & Sons, Inc., New York, 2005.