Predicting Performance in Large-Scale Identification Systems by Score Resampling

Sergey Tulyakov and Venu Govindaraju
Center for Unified Biometrics and Sensors, University at Buffalo
{tulyakov,govind}@cedar.buffalo.edu

February 29, 2012

Abstract

In this paper we investigate the problem of predicting the closed set identification performance of biometric matchers in large-scale applications given their corresponding performances in small-scale applications. We identify two major effects responsible for the prediction errors in previously proposed methods: the binomial approximation effect and the score mixing effect. We propose to use a score resampling method for prediction, which is not susceptible to the binomial approximation effect. We also reduce score mixing effect by using score selection based on identification trial statistics. The experiments on NIST biometric score dataset show the accuracy of our proposed prediction method.

1 Introduction

With the wider deployment of biometric authentication systems and the increased number of enrolled persons in such systems, the problem of correctly predicting the performance becomes more important. The number of available testing samples is usually

smaller than the number of enrolled persons the biometric system will handle. The accurate performance prediction allows system integrators to optimally select the biometric matchers for the system, as well as to properly set the decision thresholds.

The research in predicting the performance in large-scale biometric systems is still limited and mostly theoretical. Wayman [20] introduced multiple operating scenarios for biometric systems and derived the equations for predicted performance assuming that the densities of genuine and impostor scores are known. Jarosz et al. [9] presented an overview of possible performance estimation methods including extrapolation of large-scale performance given performance on smaller-scale databases, binomial approximation of performance and the application of extreme value theory. Bolle et al. [4] derived the performance in identification systems (CMC curve) assuming that the performance in corresponding biometric verification system (ROC curve) is known. The major assumption used in all these works is that the biometric match scores are independent and identically distributed - genuine scores are randomly drawn from a genuine score distribution and impostor scores are randomly and independently drawn from impostor score distribution. As we will show in this paper this assumption does not generally hold and using it leads to the underestimation of identification performance.

The need to account for match score dependencies was previously noted in [10] and [7]. Grother and Phillips [7] proposed two practical methods to deal with score dependencies - conditioning impostor scores used in the prediction on corresponding genuine scores obtained in the same test identification trial and applying T-normalization to test scores [3]. We will discuss these methods later in this paper and evaluate their performance.

The research on predicting the biometric system performance on a single test sample [19] can be considered as related to our topic since the accurate modeling of matching score dependencies in identification trials is required. The problem of estimating identification system performance was also previously studied in the area of handwrit-

ten word recognition [21, 12].

Since we will use the results of experiments throughout the paper in order to confirm our discussions, we will introduce the problem statement and experimental setup at the beginning, in section 2. Sections 3 and 4 describe two major effects influencing the prediction results - score mixing effect and binomial approximation effect. Sections 5 and 6 analyze two previously proposed methods for predicting large-scale identification system performance - binomial model and T-normalization. In section 7 we present our prediction method - resampling utilizing identification trial statistics. Finally, sections 8 and 9 contain additional discussion and conclusion.

2 Experimental Setup

We have used the biometric matching score set BSSR1 distributed by NIST[1]. This set contains matching scores for a fingerprint matcher and two face matchers 'C' and 'G'. Fingerprint matching scores are given for the left index 'li' finger matches and right index 'ri' finger matches. Since the performance of fingerprint matcher is different for two fingers, we consider these datasets as being two separate identification systems. In summary, we consider the predictions in four possible identification systems corresponding to each of these score subsets: 'C', 'G', 'li' and 'ri'.

Each of these sets contains matching scores for 6000 identification trials, and each trial has scores for either 3000 (for face sets) or 6000 (for fingerprints) enrollees. One score in each trial is genuine, and remaining are impostors related to different enrollees. In order to avoid dealing with different numbers of enrollees we restricted the number of scores in identification trials for fingerprints to 3000. Furthermore, some enrollees and some identification trials had to be discarded due to apparent enrollment errors. Finally, we obtained four datasets of 5982 identification trials with each trial having 2991 matching scores.

We use a bootstrap testing procedure [5]: for 100 iterations, we randomly split the

data in two parts - 2991 identification trials used as separate prediction and testing sets. Since our purpose is to predict the performance in larger identification systems using the performance in smaller systems, for each identification trial in the prediction set we retained only 100 randomly selected impostor scores. So, our task is by using 2991 identification trials with 100 impostor scores in each, try to predict the performance in the test set of 2991 trials and 2990 impostor scores in each trial (one score in each trial is genuine). The results of 100 bootstrap prediction/testing iterations are averaged at the end.

In this work we concentrate on predicting the *closed set identification* performance. The identification trial is considered as successful if a genuine score is higher than all impostor scores of this trial. The correct identification rate, that is a probability of successful identification trials, is a measure of closed set identification performance. Most of the previous works in predicting identification system performance also consider the scenario of *open set identification*, where, in addition to being the top score, the genuine score is required to be higher than some threshold. We chose not to consider open set identification scenario in this paper due to increased complexity of the analysis and our previous observation, that simple thresholding of top score might not deliver the optimal performance [15].

In order to have less confusion we are also not considering the more general k-th rank identification performance measured by CMC curve, though our proposed prediction methods can be easily extended to measure such performance. Our goal is to investigate the general mechanisms of identification system functioning, rather than to consider all possible operating and decision making scenarios applied for identification systems.

3 The Score Mixing Effect

One of the important characteristics of the identification system is the dependence between matching scores assigned to different classes in a single identification trial. For example, in one identification trial all the matching scores might be relatively high, and in the other trial all the scores might be low. Such dependence can be a result of multiple reasons: the quality of the input biometrics, the density of biometric templates around the input template, the particulars of the matching score calculation algorithms.

Only limited research has been carried out so far in investigating score dependencies in identification trials. Li et al. [11] try to connect the measures derived from matching scores with the quality of the image. Wang and Bhanu [19] investigate the possibility of success of the fingerprint match through the properties of fingerprint matching algorithm. Xue and Govindaraju [21] try to predict the performance of the handwritten word recognizer based on the density of the lexicon, but do not consider any other factors, e.g. quality of the word image. The explicit modeling of score dependencies presented in these approaches might be useful, but in our current investigation we are not associating the score dependence with particular characteristics of test template or the matching algorithm. The employed dataset contains only matching scores and does not allow such analysis of matching algorithms.

The following example illustrates the necessity of accounting for matching score dependencies when we try to predict the identification system performance.

3.1 Example of Identification Systems

Consider a following two-class identification system. In each identification trial we have only one genuine and one impostor score. Suppose that genuine and impostor scores are distributed according to score densities shown in Figure 1.

Consider two possible scenarios on how the matching scores are generated during identification attempt:

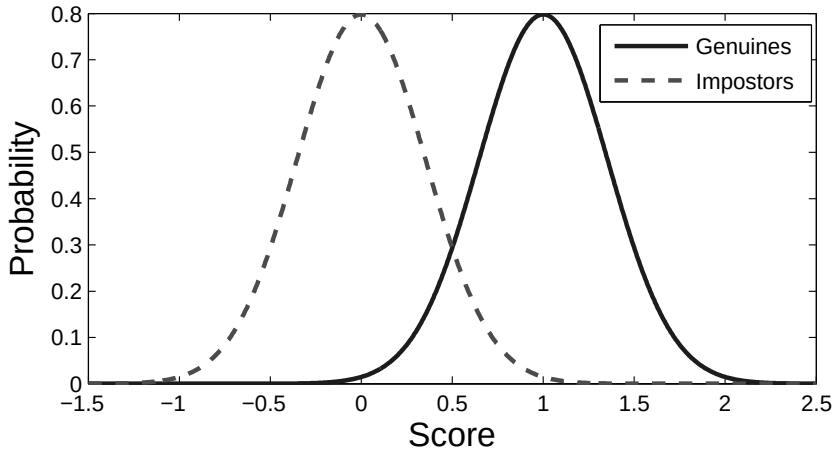

Figure 1: Hypothetical densities of matching(genuine) and non-matching(impostors) scores.

1. Both scores s_{gen} and s_{imp} are sampled independently from genuine and impostor distributions.

2. In every identification attempt : $s_{imp} = s_{gen} - 1$.

If our identification system follows first scenario, there will be identification trials with impostor score higher than the genuine score. Consequently, the correct identification rate for such system will be less than 100%. In the second scenario the identification system always correctly places genuine sample on top and has correct identification rate of 100%. Score distributions of Figure 1 do not reflect this difference. So, if we want to predict identification system performance, we need to learn the dependencies between matching scores produced in a single identification trial. Using genuine and impostor score densities alone might not be sufficient for correct prediction.

3.2 Performance of Systems with Randomized Impostor Scores

In order to confirm the presence of score dependencies in our experimental systems and the necessity to account for this dependence, we conduct the following experiment. Instead of original sets of identification trial scores, we consider identification trials with randomly chosen impostor scores belonging to different trials. In practice, we randomly permute the impostor matching scores from different identification trials. Such randomization converts our original identification systems into identification system having the same distributions of genuine and impostor scores, but impostor scores in identification trials become independent and identically distributed. Comparing with the example of the previous section, we convert the identification system with dependent scores of second scenario into identification system with independent scores of first scenario.

Matchers	True Performance	Randomized Impostors
C	0.811	0.738
G	0.774	0.669
li	0.823	0.777
ri	0.885	0.850

Table 1: Identification system performance using original identification trials ('True Performance') and using random impostors in identification trials ('Randomized Impostors').

Table 1 compares the performances of our original identification systems and corresponding identification systems with randomized impostor scores. For all matchers the difference in performances of corresponding identification systems is rather significant.

In all cases we observe that the performance of original systems is higher and not lower. This might be explained by the positive correlations between genuine and im-

postor scores for all considered matchers. When matching scores are positively correlated, we will have particular identification trials having both high genuine and high impostor scores. By distributing high impostor scores to other trials we might make them unsuccessful. This explains the lowered performance of identification systems with randomized scores.

3.3 Score Mixing Effect

When we try to predict the performance of large scale identification systems, we could be calculating some parameters or functions using matching scores from separate training identification trials. For example, most of the previous work utilizes the density of the impostor scores $n(x)$ or the cumulative distribution function of impostor scores $N(t) = \int_{-\infty}^{t} n(x)dx$ (we are using notation of [7] here). If we use all our training impostor scores to estimate these distributions, then our prediction will be a prediction of the identification system with randomized scores (as in previous section), rather than the prediction of the performance in the original system.

The *score mixing effect* is the result of considering scores from different identification trials simultaneously instead of considering the sets of matching scores from each training identification trial as separate entities for calculating prediction. The presence of score mixing effect becomes apparent as soon as practical experiments on real data are performed (see [7], section 4.2), instead of making purely theoretical predictions [20] or experimenting with synthetic data.

When we try to predict the performance of large scale identification systems, we might have only samples of training identification trials with a small number of impostors. In our experimental setup we predict performance in a systems with 2990 impostors by using training identification trials with only 100 impostors. Given 100 impostors of a single identification trial we have a great difficulty to correctly estimate the distribution of a highest score in a set of 2990 impostors. In order to make any

meaningful predictions, instead of a single trial with 100 impostors, we also have to use scores from other trials. So, it seems inevitable, that we have to mix the scores from different trials and we need to learn how to minimize the score mixing effect.

4 The Binomial Approximation Effect

In order to perform a further analysis, we will temporarily for this section assume that the scores in identification trials are independent and identically distributed according to either genuine or impostor distributions. The systems with randomized scores of previous section will serve as our test systems here.

Assuming the independence of matching scores in identification trials, the closed set identification performance in a system with G enrollees is represented by the following formula [7]:

$$R = \int_{-\infty}^{\infty} N^{G-1}(x) m(x) dx \tag{1}$$

where $N(x)$ is the cumulative distribution function of impostor (non-matching) scores, $m(x)$ is the density of genuine (matching) scores. This formula also assumes that largest score corresponds to identification result ('larger score' = 'better score'), which is true for all four matchers we have for experiments. Note that this formula can be considered as a specific case of more general formula for calculating the probability of genuine score to be in rank k (or CMC curve) [7]. Due to involvement of binomial terms in the formula for CMC, the prediction approach utilizing equation (1) is called binomial approximation prediction method.

The formula (1) can be interpreted as an expectation of function $N^{G-1}(x)$ with respect to genuine samples x, and the traditional approximation of the expectation is given by the mean of function values over the set of genuine samples in the training set:

$$R \approx \frac{1}{L} \sum_{i=1}^{L} N^{G-1}(x_i) \tag{2}$$

where L is the number of training identification trials and is the same as the number of training genuine score samples ($L = 2991$ in our experiments). It is also traditional to approximate the cumulative distribution function $N(x)$ by the empirical distribution function:

$$N(x) \approx \hat{N}(x) = \frac{1}{K} \sum_{j=1}^{K} I(y_j < x) \tag{3}$$

where K is the number of impostor scores y_j used for approximating $N(x)$, I is the identity function (1 if input parameter is true, 0 if false). After substituting (3) into (2) we obtain

$$R \approx \frac{1}{L} \sum_{i=1}^{L} \left(\frac{1}{K} \sum_{j=1}^{K} I(y_j < x_i) \right)^{G-1} \tag{4}$$

This formula can be alternatively derived using combinatorial methods similar to [10], but in our derivation we explicitly state used approximations of the theoretically correct prediction equation (1).

Using our experimental setup and and randomized training score sets of section 3.2, we evaluated the prediction capabilities of binomial approximation method (4) on all four our matchers. Note, that since the scores are randomized, the independence condition is satisfied and binomial approximation method should be theoretically optimal.

Figure 2 shows the predicted performance of matcher 'C' using binomial approximation method (4) for different numbers of training impostor scores K used for approximating $N(x)$. The experiments on other three matchers showed similar dependence of prediction on the number of used impostor samples, and we are omitting their graphs from the paper.

As we expected the predicted performance indeed converges to the true performance of the system with randomized scores with the increase in the number of used impostor scores. But this convergence is rather slow and requires a large number of training impostor samples. When the number of used impostors is small we see a significant overestimation of the identification system performance.

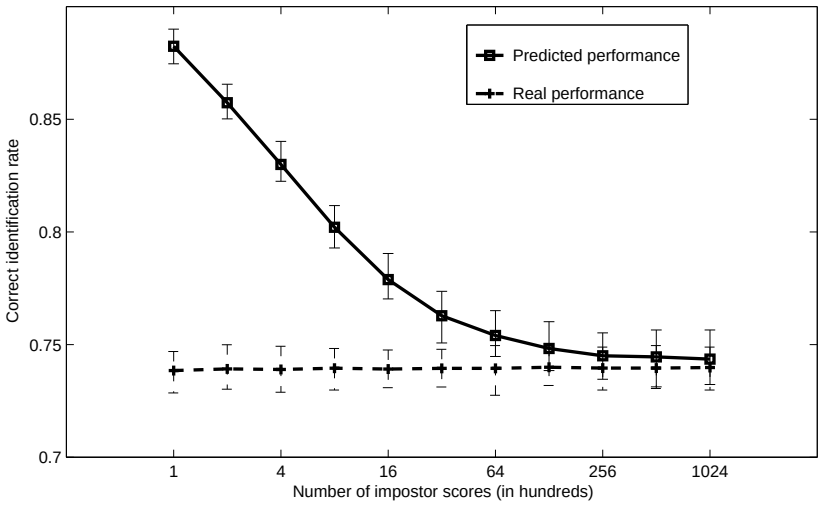

Figure 2: Dependence of predicted performance on the number of impostor scores used in binomial approximation for matcher 'C' with randomized scores.

In order to understand why such overestimation occurs, let us consider the case of $K = 100$ impostor scores used to predict performance in our system with $G-1 = 2990$ impostors. From equation (3) the values of function $\hat{N}(x)$ will be multiples of $\frac{1}{K}$. If, according to equation (4), we consider powers $\hat{N}(x)^{G-1}$, we will see that the values of these powers will be negligible with the exception of case when $\hat{N}(x) = 1$. For example, if $\hat{N}(x) = \frac{K-1}{K} = \frac{99}{100}$, then $\hat{N}(x)^{G-1} = .99^{2990} \approx 8.9 * 10^{-14}$. Effectively, in this case the application of binomial approximation (4) will simply count the number of genuine scores which are bigger than all impostors (for which $\hat{N}(x) = 1$), and the calculated performance will be close to the performance of identification system with $K = 100$ impostors instead of desired performance of a system with $G - 1 = 2990$ impostors.

Note that the overestimation of performance by binomial approximation occurs not only when $K < G - 1$, but also for bigger numbers of training impostor samples K. Doddington et al. [6] proposed to use the following rule of thumb when evaluating the

performance of biometric systems: to be 90% confident that the true error rate is within $\pm 30\%$ of the observed error rate, there must be at least 30 errors. The imprecision in predicting identification system performance is mostly explained by the errors in approximating the impostor distribution $N(x) \rightarrow \hat{N}(x)$ in the area of high values of x. In this area we might have approximated $\hat{N}(x) = 1$, which implies that for a given x we did not find any training impostor value higher than it. But the rule of thumb suggests that we need at least 30 errors (or impostors higher than x) to correctly estimate $N(x)$.

So, for the precise estimation of $N(x)$ in the area of x where we would normally get only 1 error (impostor) in our predicted system with $G - 1$ impostors, we would need to have around 30 errors (impostors). This means we would need around $30(G - 1)$ impostors to make predictions for a system with $G - 1$ impostors using binomial approximation, and the results of Figure 2 seem to confirm this reasoning. Hence, we can restate the rule of thumb of [6] with respect to predicting identification system performance by binomial approximation: the number of impostor training samples should be at least 30 times bigger than the size of identification system for which the prediction is made - $K/G > 30$.

5 The Combination Of Score Mixing and Binomial Approximation Effects

In the last section we considered identification systems with randomized scores, and thus bypassed the existence of score mixing effect. What happens if we try to predict the performance of original identification systems and both effects, score mixing effect and binomial approximation effect, influence our predictions?

The first effect underestimates identification system performance, and the second effect overestimates it. It might turn out, that we will accidentally predict correctly the performance in larger identification systems with binomial approximation and mixed

scores. Note, that the true performance of system 'C' given in Table 1 is $.811$, and from Figure 2 the performance of identification system with randomized scores is around the same number when the number of impostors used in binomial approximation is 600. So if we simply considered binomial approximation (4), taken $K = 600$ and chose random impostors y_j, our predicted performance would have coincided with the true performance.

We suspect that the influence of both effects contributed to the good prediction results reported in [10]. Though in that paper the training sets of impostors are retained, each impostor set is used with all training genuine samples. Thus the score mixing effect should be present in this approach. Also, the binomial formula for calculating prediction (7) of [10] involves term $(i/K)^{G-1}$ where $K = 100$, and, as in the analysis of previous section, we expect the binomial approximation effect to be significant. In our experiments we were not able to obtain good prediction results using approach of [10], and thus we do not report its performance.

One of the approaches considered in [7] to deal with the dependence of scores in identification trials is to condition the cumulative distribution function $N(x)$ of impostor scores on the values of genuine scores obtained in the same identification trials. Let us denote $n(y|x)$ as a density of impostors scores with the condition that impostor scores belong to identification trials having genuine score x and tet $N_x(t) = \int_{-\infty}^{t} n(y|x)dy$ denote the corresponding conditional distribution function of impostor scores. Then, assuming that impostor scores in each identification trial are independent and identically distributed according to $n(y|x)$, we can derive the following closed set identification performance prediction similar to (1):

$$R = \int_{-\infty}^{\infty} N_x^{G-1}(x)m(x)dx \tag{5}$$

In order to approximate $N_x(x)$, authors of [7] split the training identification trials into B bins of equal size according to their genuine scores. Then they approximated $N_x(x)$ using only training impostor samples from the identification trials of one bin. By

increasing the number of bins B they were trying to control the dependence between matching scores, but they disregarded the effect of binomial approximation which is dominant for larger number of bins and correspondingly smaller number of impostor scores used for approximations.

Here we repeat those experiments, but instead of splitting training identification trials into bins, for each training genuine sample x we are using impostors from K_n training identification trials with values of genuine scores closest to x. In this way, we are more precise in estimating $N_x(x)$ when the value of x might have been near some bin's boundary.

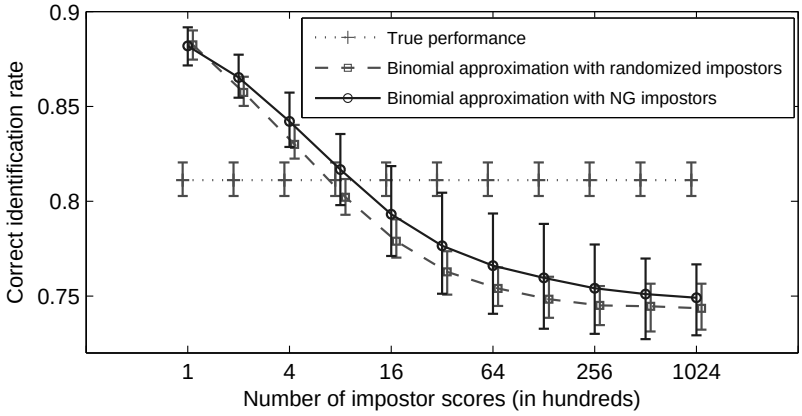

Figure 3: Dependence of predicted performance on the number of impostor scores used in binomial approximation for matcher 'C' with randomized scores and for impostor scores chosen by nearest genuine principle.

Figure 3 contains the results of these experiments on set 'C' (other sets have similar behavior). We called the method presented in this section as 'binomial approximation with NG (nearest genuine) impostors' and compared it with the binomial approximation method with randomized scores from previous section. For the same numbers of impostor scores used in binomial approximations ($K = 100K_n$), the selection of

impostor scores using nearest genuine criteria has higher predicted performance than random choice of impostors. This means that the influence of score mixing effect is reduced and the method does improve the prediction. On the other hand, the observed prediction improvements are not significant, and we can see that this method, similar to binomial approximation with randomized scores, is greatly influenced by the two previously described effects, score mixing and binomial approximation.

6 T-normalization

Another technique, which was proposed in [7] to account for score dependencies in identification trials, is to perform T(test)-normalization of matching scores before applying binomial approximation prediction:

$$x_{ij} \to \frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i} \tag{6}$$

where x_{ij} is the jth score from ith training identification trial, $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the sample mean and the sample variance of the scores in ith training identification trial. Note, that though [7] use the term Z(zero)-normalization, it seems that they actually perform T-normalization by Eq. (6) (Z-normalization has similar formula with μ and σ derived using either all available scores or scores related to a particular enrolled template).

Suppose we have some score density $p(x)$ with mean of 0 and the variance of 1. Also, suppose that for each identification trial i we are given two random parameters μ_i and σ_i, and the scores in the identification trial are independently sampled according to

$$p_i(x) = p_{\mu_i,\sigma_i}(x) = \frac{1}{\sigma_i} p(\frac{x - \mu_i}{\sigma_i}) \tag{7}$$

It is easy to show that in this case the mean of scores in the identification trial i is μ_i and the variance is σ_i. By calculating sample mean and variance estimates, $\hat{\mu}_i$ and $\hat{\sigma}_i$, and by applying T-normalization (6) to the identification trial scores, the transformed

scores will be approximately (due to approximations $\mu_i \approx \hat{\mu}_i$ and $\sigma_i \approx \hat{\sigma}_i$) distributed according to $p(x)$.

Equation (7) represents a possible model of how the dependencies between matching scores in identification trials originate. We can call it the *linear score dependency model*. Previously, Navratil and Ramaswamy [13] described the T-normalization using the property of *local gaussianity*, which assumes that function $p_i(x)$ is close to normal density with mean μ_i and variance σ_i. In our description we are not making any assumptions on the form of $p_i(x)$ except that it is generated for each identification trial by Eq. (7) using some common density p. There is also no assumptions on distributions of μ_i and σ_i (which are randomly chosen for each identification trial).

According to linear score dependency model the range of scores in each identification trial is shifted by μ_i and stretched by σ_i. Note, that there are two types of scores in identification trials - genuine and impostors, and it is quite possible that they might have different dependence models. But the number of genuine scores in identification trials is limited (usually only one genuine score), and it is not possible to learn the dependency model for genuine scores. Therefore, we will assume that the same model is applied for both types of scores; the sample estimates $\hat{\mu}_i$ and $\hat{\sigma}_i$ can be computed using both genuine and impostor samples, but in this work we use only impostor score samples.

T-normalization is a linear transformation for each identification trial, and it does not change the order of matching scores. So, if identification trial was successful, it will remain successful after T-normalization. Thus, instead of making performance prediction in an identification system with linear score dependency model (7), we can make predictions in an identification system with T-normalized scores. More specifically, assuming that genuine and impostor scores in each identification trial are the

result of linear score dependency model and have distributions

$$m_i(x) = m_{\mu_i,\sigma_i}(x) = \frac{1}{\sigma_i} m(\frac{x-\mu_i}{\sigma_i})$$
$$n_i(x) = n_{\mu_i,\sigma_i}(x) = \frac{1}{\sigma_i} n(\frac{x-\mu_i}{\sigma_i}) \quad (8)$$

after T-normalization genuine and impostor scores will be independently and identically distributed according to $m(x)$ and $n(x)$, and the closed set identification performance of original system will be similar to the performance of identification system with i.i.d. scores with densities $m(x)$ and $n(x)$.

Since the total number of impostor scores in our experimental setup is sufficient to make binomial approximation performance prediction of closed set identification system with independent scores, we made such predictions on T-normalized scores for all four identification systems. Table 2 shows the results of this prediction.

Matchers	True	T-norm & BA
C	0.811	0.818
G	0.774	0.602
li	0.823	0.838
ri	0.892	0.902

Table 2: True performances of identification systems ('True') and prediction using T-normalized scores and binomial approximation on a full set('T-norm & BA').

The use of T-normalization seems to give almost perfect prediction results for 3 systems, but failed for predicting the performance of identification system 'G'. This failure means that the linear score dependence model does not represent the set of matching scores in system 'G', and we have to search for some other model of score dependence. Additionally, even if other systems do achieve good performance prediction after T-normalization, it is not necessary that linear score dependence model exactly describes the dependencies of scores in identification trials and the actual de-

pendencies might be more complex.

7 Resampling Methods

In this work we introduce the resampling method for predicting large-scale identification system performance. The method is rather simple: we simulate the work of the identification system by choosing the genuine and impostor scores from the training set. Specifically, for each training genuine sample, we choose $G - 1 = 2990$ training impostor samples. If the genuine score is the highest, then the identification trial is successful and the performance of the simulated system is calculated as a proportion of successful identification trials.

It is clear that this method requires bigger number of training impostor scores than the number G of enrolled persons in simulated system. But, since we analyzed that the number of impostor scores for binomial approximation should be at least 30 times more than G, we can expect that the approximation abilities of resampling method will be on par with the abilities of binomial approximation. In order to confirm the approximation abilities of proposed method, we compared its performance with binomial approximation method on an identification system with randomized scores (section 3.2) using the full training set of $L = 2991$ genuine samples and $K = 2991 * 100$ impostor scores. Whereas in binomial approximation method (4) for each genuine score we used all $K = 2991 * 100$ impostors, in resampling method we were randomly choosing $G - 1 = 2990$ impostors for each genuine.

The results of these experiments are shown in Figure 4. Both methods show similar approximating performance. The spread of error bars is also similar and slightly bigger than the spread of error bars on test set. Note, that evaluation on the test set ('True performance') works in essentially the same way as resampling method. The only difference is that evaluation on the test set uses $2991 * 2990$ test impostor scores with 2990 non-repeating impostors for each genuine score, but resampling method uses only

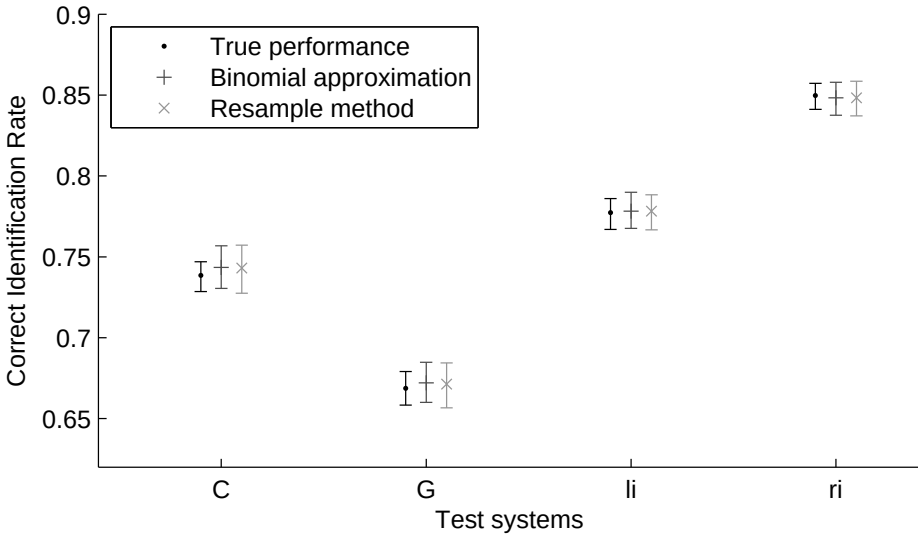

Figure 4: Performance prediction of identification system with randomized scores by binomial approximation and score resampling methods.

$2991 * 100$ impostors and has to repeatedly use impostors with each impostor used approximately 30 times. The reuse of training impostor samples explains the bigger spread of error bars for resampling method.

7.1 Resampling Using Genuine Score Neighbors

The key advantage of the resampling method and the reason for its use is that it allows us to more precisely control the score mixing effect when performing prediction. The binomial method requires mixing more than $30G$ impostors by formula (4) for each training genuine score, but resampling method uses only $G - 1$ impostors for each genuine. The binomial approximation effect did not allow us to correctly predict performance by approximating cumulative distribution functions $N_x(x)$ conditioned on genuine scores x in section 5. The resampling method is not susceptible to the binomial approximation effect and allows us to more precisely evaluate the benefits of

utilizing genuine score conditioning.

In this section we modify the experiments of section 5 using resampling method.

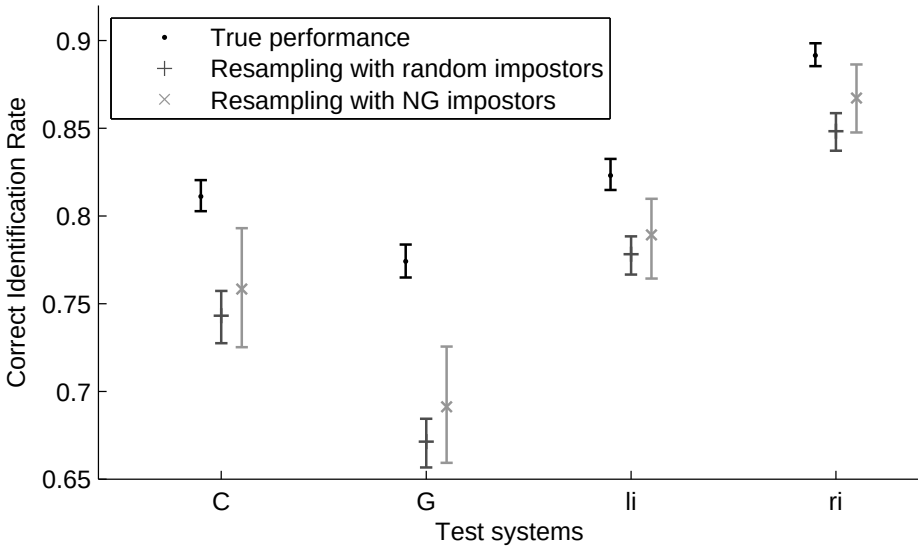

Figure 5: Performance prediction in original identification systems by score resampling methods with randomized sampling and nearest genuine sampling.

Figure 5 compares the performance of resampling method utilizing the nearest genuine sampling method with the resampling method using random impostors and the true performances of our systems. Clearly, using nearest genuine identification trial reduces the score mixing effect, but this reduction is still not sufficient for precise performance prediction. Similar reduction was observed for binomial approximation method (Figure 3), but due to binomial approximation effect we were not able to judge objectively the strength of using nearest genuine principle.

7.2 Score Resampling Using Identification Trial Statistics

In order to control the mixing effect in the resampling method we want to mix scores only from similar training identification trials. Selecting identification trials using closest genuine scores of the previous section is just one possible way of specifying the similarity between identification trials. We expand this method by using statistics of identification trial score sets to determine the similarity between trials.

Let $T_i = \{x_{ij}\}_j$ denote the set of matching scores from the ith training identification trial and let $t(T_i)$ denote some statistic of this set. For example, $t(T_i)$ could be the sample mean $\hat{\mu}_i$ or the sample variance $\hat{\sigma}_i$ statistics used for T-normalization in section 6. Define the distance between identification trials T_i and T_k with respect to statistic function t as a distance between corresponding statistics of two sets:

$$dist_t(T_i, T_k) = |t(T_i) - t(T_k)| \tag{9}$$

Denote G_t as the number of impostor scores in training identification trials ($G_t = 100$ in our experiments). Then the resampling method with identification trial statistic t for predicting identification system performance is formulated as follows:

1. For training identification trial T_i and corresponding genuine score x_i, find $K_n = \lceil (G-1)/G_t \rceil$ training identification trials T_k closest to T_i with respect to distance $dist_t(T_i, T_k)$

2. Choose random $G-1$ impostors from selected identification trials; simulated trial is successful if x_i is bigger than all chosen impostors

3. Repeat 1-2 for all available training identification trials T_i and calculate the predicted system performance as the proportion of successful simulated identification trials

The proposed resampling algorithm is rather simple and does not require any parameter training. However, it does require proper selection of used identification trial

statistic t. In the rest of this section we will investigate the use of different statistics. Note, that $K_n = \lceil (G-1)/G_t \rceil = 30$ in our experiments, so for each genuine score we are looking for 30 training identification trials out of total 2991. Thus, the resampling method seem to be quite selective and might be able to significantly reduce the score mixing effect.

7.3 Resampling and T-normalization

The performance prediction method based on T-normalization (Eq. 6) used two iden-

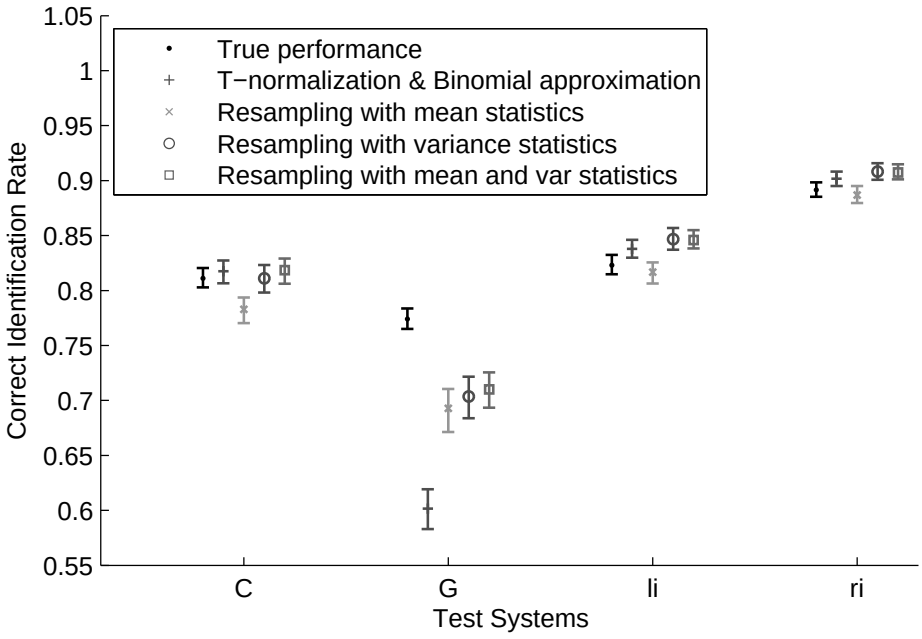

Figure 6: Performance prediction in original identification systems by T-normalization followed by binomial approximation and score resampling methods based on mean and mean-variance statistics.

Figure 6 presents the results of these experiments. Note, that when we use both mean and variance statistics, the statistics of identification trials are two dimensional vectors $t(T_i) = (\hat{\mu}_i, \hat{\sigma}_i)$, and instead of simple absolute difference for calculating distance in equation 9 we use euclidean distance.

T-normalization method has quite good prediction performance for matchers 'C', 'li' and 'ri', and resampling method using mean and variance statistics is also close to the true system performance. The interesting feature here is that variance statistics apparently reduces score mixing effect more than mean statistics. For matcher 'G' on which T-normalization method failed, we see better prediction results by resampling methods using either or both of these statistics, but the prediction is still far from true performance.

If we compare these results with the prediction results of Figure 5, we notice that any of the resampling methods with mean or variance statistics reduce score mixing effect better than resampling using nearest genuine neighbors. Generally, we can view a genuine score from identification trial as a statistic of the trial. But considering a single score as a statistic might not be a reliable way to model dependencies of scores in identification trials, and the poor prediction results of nearest genuine score resampling method confirm this.

7.4 Resampling Using n-th Order Statistics

The other type of frequently used statistics is the n-th order statistics

$$t_i^n = t^n(T_i) = \{\text{the value of n-th highest element in } T_i\} \tag{10}$$

In our experiments we use a set of impostor scores in each identification trial T_i to calculate n-th order statistics t_i^n and use them in resampling method for prediction. Figure 7 shows the results of experiments.

Overall, using the 2-nd order statistics or the second best impostor score gave the best prediction results. The prediction precision seems to decrease with the increased

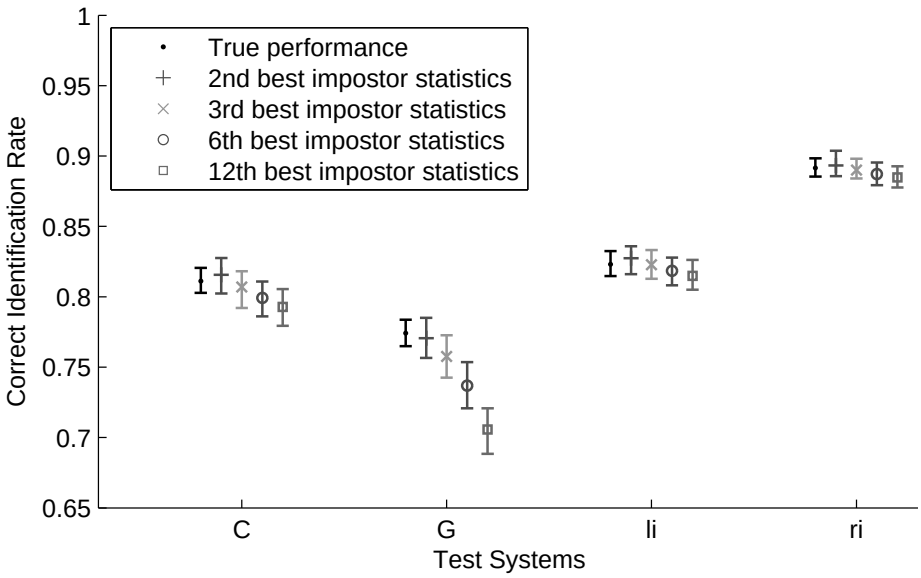

Figure 7: Performance prediction in original identification systems by resampling methods utilizing different nth order statistics of identification trial score sets.

order of used statistics. Note, that the last, 12-th order statistics correspond to $G_t/8$-th order statistics. We also tried higher order statistics, like $G_t/4$ and $G_t/2$, but the prediction accuracy was consistently worse and the predicted performance approached the performance of system with randomized scores (section 3.2).

There seems to be no theoretical proof on why second best impostor statistic should allow making more precise prediction than any other statistic. We can reason that it simply better reflects behaviors of the high score tails of impostor distributions in separate identification trials. If we needed to estimate some other features of impostor distributions, we might have better results using some other statistics.

7.5 Justification of Using Identification Trial Statistics

Consider the following identification trial score dependency model. Suppose that for each identification trial T_i we are given a set of parameters $\boldsymbol{\theta}_i = \{\theta_i^1, \ldots, \theta_i^k\}$, and the scores in the identification trial are independently sampled according to

$$p_i(x) = p(x|\boldsymbol{\theta}_i) \tag{11}$$

As we already pointed out, all the matching scores in a single identification trial are generated using the same test biometric sample, and parameters θ_i could represent test sample quality, the amount of information contained in the test sample (e.g. number of minutia in test fingerprint), the closeness of other samples to the test sample and similar characteristics of test sample.

For each identification trial T_i, we can extract a set of score set statistics $\boldsymbol{t}_i$. Thus, statistics are the random variables of parameters $\boldsymbol{\theta}_i$, $\boldsymbol{t}_i = t(T_i(\boldsymbol{\theta}_i))$, and if statistics are stable, then $\boldsymbol{t}_i \approx \phi(\boldsymbol{\theta}_i) = E(t(T_i(\boldsymbol{\theta}_i)))$. Assume that statistics $\boldsymbol{t}_i$ are chosen so that function $\phi(\boldsymbol{\theta}_i)$ is continuous and invertible. In this case we obtain that identification trials having close statistics $\boldsymbol{t}_i$, will have close parameters $\boldsymbol{\theta}_i$ and, as a consequence, close score densities $p_i(x) = p(x|\boldsymbol{\theta}_i)$. Therefore, we can simply use training identification trials with close statistics to estimate score densities $p_i(x)$ without explicit estimation of parameters $\boldsymbol{\theta}_i$.

In practice, since we have no knowledge on what parameters $\boldsymbol{\theta}_i$ might be, we cannot guarantee the good properties of function $\phi(\boldsymbol{\theta}_i)$. Intuitively, if we consider the statistics vector $\boldsymbol{t}_i$ to consist of many diverse statistics of identification trial score set, then we are more sure that close statistics are the result of sampling close conditional densities $p_i(x)$.

7.6 Close Sampling Effect

Suppose that we are given a large number L of training impostor trials and we use sufficiently large number of statistics in the resampling method. When we search for the closest training identification trials, we might find the identification trials very similar to the one we consider at the moment. Indeed, a sufficiently large number of available training trials will result in the existence of very similar trials, and if the chosen statistics well reflects the set of training identification trial scores, these very similar trials will have very similar statistics and will be found during resampling method search.

In the extreme case, all found closest training trials will be exactly the same as the particular training trial whose genuine score we consider at the moment; the simulated trial will be successful if and only if that particular training identification trial was successful. The predicted performance for a larger system will be exactly the same as the performance of smaller system and the prediction algorithm will fail.

Though the extreme case seems to be improbable, some overestimation of predicted system performance might be observed in our experiments. For example, if we use the best impostor score instead of second and other n-th order statistics in resampling method, we will find that predicted performance will be almost the same as the performance in smaller training system with $G_t = 100$ impostors. The reason for this is quite clear - the best impostor in the simulated trial will be among the best impostors in closest training trials, and all of them are close to the best impostor of currently considered training trial. We might call the effect of overestimating identification system performance due to too close neighboring trials as *close sampling effect*. It seems that it is quite difficult to say whether the effect has influence on particular prediction results. Still, we need to control the appearance of this effect by making sure that used statistics in resampling method does not coincide with the property of the system we are trying to predict.

8 Discussions

8.1 Identifciation models

Accounting for the dependencies between matching scores assigned to different classes during single identification trial seems to be the key for correct prediction of identification system performance. The existence of this dependence has been mostly ignored in biometrics research so far. The problem lies in the difficulty of modeling this dependence and deriving any algorithms using it; therefore a simplifying assumption on score independence is usually made.

In [15] we proposed to use in addition to currently considered score a best score from the identification trial besides current (second best score) for making acceptance decisions in verification systems. In [17] we used second best score in combinations of biometric matchers. In order to differentiate the models for score dependencies in identification trials from previously explored score dependencies in cohort and background models, we introduced the term *identification model*. We further formalized the notion of identification models utilizing identification trial score set statistics in [18].

Resampling methods utilizing identification trial statistics can be viewed as an extension of the identification model research in the area of predicting the performance of identification systems. The usefulness of chosen statistics in identification model is judged by the prediction precision, whereas in previous research the usefulness of statistics is determined by its ability to improve performance of either decision making or of combination algorithm. The current research well complements the previous studies - if some statistics is useful for prediction, it must contain information about score dependencies in identification trials and consequently can be successfully utilized in decision making or classifier combination.

Note, that the *second best impostor* statistics used in experiments of current paper is slightly different from the *second best score* statistics utilized in our previous research,

where *second best score* is calculated using all matching scores including genuine. The difference is that in current experiments we precisely know which scores are impostor and which score is genuine in the identification trials. Previous research modeled the situations where such knowledge is not available, for example, if we use some scores in identification trial for combination we are not aware which score is genuine - the final goal of combination algorithm is to find it. Nevertheless, both statistics are closely related, and current research confirms the use of second best score statistics advocated before.

8.2 Extreme Value Theory

The important part of the identification system performance prediction research is modeling the distributions of scores in the tails, especially, the tail of impostor distribution corresponding to high scores. Extreme value theory is a field of statistics investigating the behavior of distribution tails and we can expect the improvement in the prediction if we use its techniques.

One of the results of extreme value theory states that the distribution of values of random variable X satisfying the condition of being in the tail, $X > u$ for sufficiently large u, is close to the generalized Pareto distribution (GPD):

$$F_u(x) = P(X - u > x | X > u) \approx \begin{cases} 1 - (1 - kx/a)^{1/k} & k \neq 0 \\ 1 - \exp(-x/a) & k = 0 \end{cases} \quad (12)$$

The parameters k and a can be learned from training data by different methods [8] for a particular choice of u. Equation (12) provides only an asymptotic approximation of the extreme value distribution of X when u approaches the supremum of all possible values of X. The derivation of sufficient conditions on the minimum number of samples of X, confidence intervals of u, k and a, is a main topic for ongoing research in extreme value theory. Note, that most existing research in extreme value theory is rather theoretical; the ability to predict the performance in identification systems might

be used as an objective practical measure to evaluate the performance of extreme value theory methods.

The main assumption for the application of the extreme value theory is the independence and identical distribution of the samples X. Since there is a dependence between matching scores in identification trials, we expect that extreme value theory will have same problem as binomial approximation for performance prediction in identification systems - we would need to mix sets of scores from different identification trials to make good approximations, and consequently will introduce score mixing effect into prediction.

One possible solution is to use identification trial score set statistics in order to select close training identification trial. Though the results presented in [9] seem to imply that extreme value theory provides better approximations than binomial model, it is not clear if using it along with score set statistics will deliver better prediction than resampling method. Another solution might be to try to parameterize the fitting of GPD to the tails of impostor distributions for different identification trials. Thus, instead of common parameters u, k and a, we would need to find separate u_i, k_i and a_i for each training identification trial T_i. Statistics of identification trials t_i can serve for such parameterization. Alternatively, we might consider joint density modeling of statistics and extreme values of X by means of multivariate extreme value theory [14].

8.3 Performance Prediction in Open Set Identification Systems

Whereas the closed set identification problem assumes that the genuine user is enrolled and the match is performed against 1 genuine and $G-1$ impostor templates, the open set identification problem assumes that genuine user might not be enrolled and the correct solution of the identification system will be to reject current identification attempt. Clearly, the analysis of open set identification system should include the assumption on the prior probability of the user to be enrolled. It is not clear if the proper analysis of

open set identification systems has been presented before; recent works discussing open set identification (e.g. [7]) do not use such prior probability. In contrast to traditional ROC curve used for evaluating verification systems and describing the trade-off between two types of errors, false accept and false reject, open set identification systems have three types of error [2]: the error of incorrectly choosing first matching choice, the error of accepting incorrect first choice, and the error rejecting correct first choice. The trade-off between three types of errors might be described by a two-dimensional surface in the three dimensional space, and we are not aware of any research using such performance measures.

Instead of considering the full system with three error types, we can consider the reduced open set identification problem assuming that the test user is always enrolled in the database and the system has the ability to reject the first match choice. Such system indeed will be quite useful since first match choice might be an impostor and rejecting such choice is the correct decision. Similar approach is also taken explicitly in [2] and in our works [15, 16]. In such case we have two types of error - accepting the incorrect first choice and rejecting the correct first choice or identification.

Traditional decision to accept or reject the first choice in open identification systems is to compare the first matching score to some threshold θ [7]. In [15] we showed that such decision is not optimal and we get better results if instead of only single first score s^1 we also use second best score s^2 and base our decision on thresholding some learned function of these two scores: $f(s^1, s^2) > \theta$. We further explored this idea in [16] and showed that the improvement is theoretically present even if scores in identification trials are independent (and impostor scores are identically distributed). The rate of improvement seems to decrease slightly with the increase of the number of impostors.

This discussion implies that the estimation of open set identification system performance is not an easy task. Although, we can follow the traditional derivations [20, 7, 9] specifying that the false match rate in a system with N impostors can be determined

by the function of false match rate of verification system: $FMR_{1:N} = 1 - (1 - FMR_{1:1})^N$, and the false non-match rate stays the same: $FNMR_{1:N} = FNMR_{1:1}$, such measures are not adequate for proper performance description due to broad assumptions: 1. independence of matching scores in identification trials, 2. the decision based on thresholding single top score, 3. the whole system performance can be described by two numbers (note that open set identification systems have three types of error, so these false match and false non-match rates might not be sufficient). Therefore we restricted the topic of current paper to close set identification, and left the investigation of open set identification systems for the future.

The results presented in current paper suggest that the predictions of open set identification system performance might also have to deal with score mixing effect, and we might have to use score set statistics for selecting close identification trials for testing. Note also, that use of the second best score for making decisions is similar to using this score as the statistics of identification trials. Therefore, it is not clear how much benefit using identification set statistics might have on open set identification system already utilizing such scores for decisions.

9 Conclusion

In this paper we investigated the problem of predicting the performance of large-scale closed set identification systems. First, we showed the existing dependency in matching scores assigned to different classes during identification trials. This dependency has major effect on the previously proposed algorithms for estimating system performance. Second, we showed that binomial approximation prediction method introduces its own effect on performance prediction. Third, we discussed the T-normalization and its relationship to the prediction problem. Fourth , we proposed the new prediction method based on resampling of available training scores using identification trial statistics. The utilization of identification trial statistics allows to reduce score mixing effect, and

delivers good prediction results. Finally, we discussed the results of the paper with respect to other research directions: identification models for decisions and matcher combinations, extreme value theory and open set identification system performance prediction.

References

[1] Nist biometric scores set. http://www.nist.gov/biometricscores/.

[2] A.M. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, A. A4 Malegaonkar. Verification effectiveness in open-set speaker identification. *Vision, Image and Signal Processing, IEE Proceedings* -, 153(5):618–624, 2006.

[3] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[4] R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, and A.W. Senior. The relation between the ROC curve and the CMC. In *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, pages 15–20, 2005.

[5] Ruud M. Bolle, Nalini K. Ratha, and Sharath Pankanti. Error analysis of pattern recognition systems–the subsets bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33, 2004.

[6] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254, 2000.

[7] P. Grother and P.J. Phillips. Models of large population recognition performance. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of*

the 2004 IEEE Computer Society Conference on, volume 2, pages II–68–II–75 Vol.2, 2004.

[8] J. R. M. Hosking and J. R. Wallis. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3):339–349, 1987.

[9] Herve Jarosz, Jean-Christophe Fondeur, and Xavier Dupre. Large-scale identification system design. In James Wayman, Anil Jain, Davide Maltoni, and Dario Maio, editors, *Biometric Systems Technology, Design and Performance Evaluation*. Springer London, 2005.

[10] A.Y. Johnson, J. Sun, and A.F. Bobick. Using similarity scores from a small gallery to estimate recognition performance for larger galleries. In J. Sun, editor, *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*, pages 100–103, 2003.

[11] Weiliang Li, Xiang Gao, and T.E. Boult. Predicting biometric system failure. In Xiang Gao, editor, *Computational Intelligence for Homeland Security and Personal Safety, 2005. CIHSPS 2005. Proceedings of the 2005 IEEE International Conference on*, pages 57–64, 2005.

[12] U.-V. Marti and H. Bunke. On the influence of vocabulary size and language models in unconstrained handwritten text recognition. In H. Bunke, editor, *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 260–265, 2001.

[13] Jiri Navratil and Ganesh N. Ramaswamy. The awe and mystery of T-norm. In *8th European Conference on Speech Communication and Technology (EUROSPEECH-2003)*, pages 2009–2012, Geneva, Switzerland, 2003.

[14] Ser-Huang Poon, Michael Rockinger, and Jonathan Tawn. Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *Rev. Financ. Stud.*, 17(2):581–610, 2004.

[15] S. Tulyakov and V. Govindaraju. Combining matching scores in identification model. In *8th International Conference on Document Analysis and Recognition (ICDAR 2005)*, Seoul, Korea, 2005.

[16] S. Tulyakov and V. Govindaraju. Identification model with independent matching scores. In *Biometrics Consortium Conference*, Crystal City, VA, 2005.

[17] S. Tulyakov and V. Govindaraju. Identification model for classifier combinations. In *Biometrics Consortium Conference*, Baltimore, MD, 2006.

[18] Sergey Tulyakov, Zhi Zhang, and Venu Govindaraju. Comparison of combination methods utilizing T-normalization and second best score model. In *CVPR 2008 Workshop on Biometrics*, 2008.

[19] Rong Wang and Bir Bhanu. Predicting fingerprint biometrics performance from a small gallery. *Pattern Recognition Letters*, 28(1):40–48, 2007.

[20] J.L. Wayman. Error rate equations for the general biometric system. *Robotics & Automation Magazine, IEEE*, 6(1):35–48, 1999.

[21] Hanhong Xue and V. Govindaraju. On the dependence of handwritten word recognizers on lexicons. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(12):1553–1564, 2002.