International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 24, No. 2 (2010) 173–191

DOI: 10.1142/S0218001410007944

ON THE DIFFERENCE BETWEEN OPTIMAL COMBINATION FUNCTIONS FOR VERIFICATION AND IDENTIFICATION SYSTEMS

SERGEY TULYAKOV, CHAOHONG WU
and VENU GOVINDARAJU

Center for Unified Biometrics and Sensors
University at Buffalo, 520 Lee Entrance,
Suite 202, Amherst, NY 14228, USA

We have investigated different scenarios of combining pattern matchers. The combination problem can be viewed as a construction of a postprocessing classifier operating on the matching scores of the combined matchers. The optimal combination algorithm for verification systems corresponds to the likelihood ratio combination function. It can be implemented by the direct reconstruction of this function with genuine and impostor score density approximations. However, the optimal combination algorithm for identification systems is difficult to express analytically. We will show that this difficulty is caused by the dependencies between matching scores assigned to different classes by the same classifier. The experiments on the large sets of scores from handwritten word recognizers operating on postal images and biometric matchers (NIST biometric score set BSSR1) confirm the existence of such dependencies and that the optimal combination functions for verification and identification systems are different.

Keywords: Combination of classifiers; biometric identification systems; likelihood ratio; weighted sum.

1. Introduction

In this paper, we investigate the problem of combining the outputs of multiple classifiers. Combined classifiers might use different features or different matching algorithms, and as the large body of previous research shows, the combined algorithm can have superior performance when compared to any single participating classifier. We will assume that the set of classifiers being combined is fixed. Therefore we do not consider the problem of classifier ensembles with dynamically generated set of classifiers. In fact, in this study we will be only considering the combination of two given classifiers. Thus our problem consists of learning the statistical properties of each classifier's output and finding the proper combination algorithm.

We will also assume that each classifier outputs a numerical matching score for each class which reflects the confidence that the input belongs to that class. We will call such classifiers "matchers" to distinguish them from other types of classifiers which output only a single class label corresponding to the most probable class of the input or class ranks. Xu *et al.*[28] described three types of classifier combinations based on the types of output produced by a single participating classifier. In this paper, we deal with type III combinations which return a measurement level scores corresponding to every class which can be used to rank the classes.

The applications considered in this paper include the combination of multimodal biometric matchers and handwritten word recognizers. In both cases, two matchers process the input and produce an output consisting of two matching scores for each class. In the case of biometric matchers, the input consists of fingerprint and face templates, and the classes are the enrolled persons. The two matching fingerprint and face scores are used to obtain a single combined score for each person, and the person corresponding to the best combined score is output as the system's classification result. In case of handwritten word recognizers, the input is an image of a word, and the classes are the words in a lexicon. Two matchers are used to obtain two matching scores for each lexical entry. The combination algorithms produce a single combined score for each entry, and the lexical entry with best combined score is taken as the classification result.

1.1. *Problem description*

Let M denote the number of combined classifiers and N denote the number of classes. Each classifier $j = 1, \ldots, M$ produces sets of matching scores s_i^j assigned to each of $i = 1, \ldots, N$ classes. Our combination methods will operate on these scores. In both the biometric and word recognition applications, a combination function f of scores is used to combine M matching scores corresponding to each class, and the classification result C is determined by the corresponding combination rule:

$$C = \arg \max_{i=1,\ldots,N} f(s_i^1, \ldots, s_i^M) \tag{1}$$

Note that the upper index of the score corresponds to the classifier which produced the score, and the lower index corresponds to the class for which it was produced. The sum function $f(s^1, \ldots, s^M) = s^1 + \cdots + s^M$ corresponds to the sum rule, the product function $f(s^1, \ldots, s^M) = s^1 \cdots s^M$ corresponds to the product rule and so on. Such combination functions commonly used by researchers are usually fixed (as opposed to being learned from training data).[17] Using these fixed combination rules usually requires a processing step to normalize the matching scores. In general, these *ad hoc* functions are not optimal. We are interested in deriving the optimal combination function f of Eq. (1) using training data and machine learning algorithms. We illustrate with the help of artificial examples the difficulty of this task even when a sufficient number of training samples are available.

The set of matching scores available for the combination algorithm is shown (Fig. 1) as a lattice with rows containing the scores produced by a classifier j and columns containing the scores assigned to a class i. The combination function f

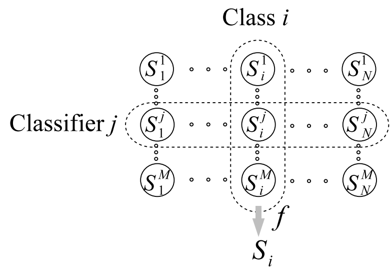


Fig. 1. The set of matching scores available for the combination algorithms in identification systems.

accepts as parameters scores for class i $(s_i^1, \ldots, s_i^M)$ and produces a combined score S_i.

1.2. *Verification and identification modes*

Note, that two modes of operation are usually considered for the matching system. In the first, *verification* mode, the identity of the class i is provided as a hypothesis, and the decision to accept or reject class i as a matching result is based on comparing the combined score S_i to some threshold. In the second, *identification* mode, no hypothesis is provided and combined scores S_i are calculated for all classes i. The final classification decision is made by choosing the class with the greatest score returned by Eq. (1). *Our goal is to find the optimal combination function for the identification system.* The guiding intuition of our research is to compare the problem of finding the optimal combination function f of the identification system of Eq. (1) with the problem of finding the optimal combination function of the same system operating in the verification mode.

Different modes of operation require different measures of performance. Whereas ROC or DET curves are useful for measuring performance in verification systems, the performance in identification systems is usually measured by the correct identification rate or cumulative match curve (CMC). In this paper, we use correct identification (classification) rate, that is the frequency of correctly identifying the class by Eq. (1).

One very important notion that we explore in this paper is the notion of score dependence. Note, that there could be two types of score dependencies in a matching system (Fig. 1). The first type of dependence is that between matchers — between the scores assigned by different matchers to a single class. This is the dependence between scores in a column $(s_i^1, \ldots, s_i^M)$. The second type of dependence is that between scores produced by a single matcher and assigned to different classes. One can view it as a dependence between scores located in a single row $(s_1^j, \ldots, s_N^j)$. The first type of dependence has been the focus of researchers thus far and is adequately addressed in the construction of the combination function f. However, the main focus of our research in this paper is in the second type of dependence.

Understanding this dependence is a necessary step in constructing optimal combination function for identification systems.

1.3. *Paper outline*

In the next section, we review some of the previous research in classifier combination field. In Sec. 3, we show that the optimal combination function for verification systems will be also optimal for equivalent identification systems only if an additional condition of independence of matching scores assigned to different classes is satisfied. Section 4 presents two important examples to illustrate that if this independence condition is not satisfied, then the optimal combination functions of the two systems are necessarily different. This is a fundamental finding in the classifier combination field based on our assessment and literature review.

Section 5 provides experiments confirming the difference in the optimal combination functions for verification and identification systems. First, we introduce the handwritten word recognizers (Sec. 5.1) and biometric matchers (Sec. 5.2) which are used in the experiments. The dependence between scores returned by a single matcher for different classes is presented in Sec. 5.3. In Sec. 5.5, we compare the use of likelihood ratio and weighted sum functions for combination.

2. Previous Work

Although research in the classifier combination field has produced several new combination algorithms, a theoretical underpinning for this research area is still missing. Jain *et al.*[15] stated that methods claiming optimality actually make rather strict assumptions on the properties of the classifiers being combined. For example, Kittler *et al.*[17] assumed that the matching scores produced by the classifiers participating in the combination correspond to posterior class probabilities, thus justifying their use of product or sum combination rules depending on some additional assumptions. In the applications we have considered in this paper, the matching scores reflect distance measures between the biometric templates or between the handwritten word image and a lexicon word. These distances can be converted to probabilities,[6,9,14] but this conversion is nontrivial and prone to errors.[3]

Snelick *et al.*[20] investigated the combination of three fingerprint and one face biometric matchers. Five combination methods and five score normalization functions are tested to construct the combination algorithm. Since only a limited number of combination algorithms are tested, there is no guarantee that the method finally chosen is optimal or even close to optimal.

Bayesian and Dempster-Shafer combination methods in Ref. 28 required learning confusion matrices for each classifier participating in the combination. The Behavior-Knowledge Space combination method in Ref. 13 requires learning a decision space of a set of classifiers participating in the combination. Although these approaches can be considered to be optimal in some sense, their utility is restricted to applications with a few classes. However, in our applications of biometrics and handwritten word recognition, the number of classes N is of the order of thousands.

The goal of combination methods proposed in this paper is to perform combinations in identification systems [Eq. (1)] with a large number of classes N. We show that optimal combinations in identification and verification systems need not be the same. Our primary interest is in approximating the optimal combination functions for identification systems. Previous work in classifier combination has actually failed to address this differentiation, and the performance of generic combination methods has been evaluated without paying attention to whether they are designed for identification systems or verification systems. For example, Lee *et al.*[18] explicitly reduced the problem of combining matchers in a biometric identification system to the task of applying a classifier (SVM) trained for an equivalent verification system. We show in this paper, by means of artificial examples (Sec. 4), that such an approach may not produce an optimal combination algorithm for identification systems.

Some researchers[5,12] have tried to predict the performance of an identification system given data about the performance of an equivalent verification system. The necessary condition used in these predictions is the independence of matching scores assigned to different classes. Our experiments show that this condition is unrealistic and the scores are usually dependent. Therefore, the predictions of performance of identification systems based on observations made in verification systems might not be valid. In fact, the score dependence is precisely the reason why verification and identification systems require different classifier combination algorithms.

One way to account for the dependence of scores assigned to different classes is to use not only single scores assigned to one class by combined classifiers, but some additional information derived from scores assigned to other classes. Such information might include the rank of the current score, the difference between this score and the best assigned score, or any other statistic of the score set produced by the same classifier. The combination methods based on ranks, e.g. Behavior-Knowledge Space,[13] might be efficient in using this information, but the original score gets discarded. More complex schemes, such as in Ref. 2, consider the weighting of information (e.g. difference between scores) and can provide better performance than combinations using ranks only. But, as we discussed in Ref. 22, such combinations belong to more complex type of combinations not defined by Eq. (1). We restrict our attention in the current paper to seeking proper combination functions of Eq. (1), though investigating more complex combinations explicitly including the dependence information should be one of future research directions.

We have presented the initial results of our investigation into the properties of optimal combination functions in identification systems in Refs. 25 and 26. In this paper, we provide a deeper discussion on the relationship between dependence of matching scores and the construction of optimal combination function. In particular, two illustrative examples in Sec. 4 give an insight to the difficulty of finding the optimal combination function for identification systems.

3. Likelihood Ratio Based Combination Rule

In both applications of biometrics and handwriting recognition, we encounter verification and identification modes of operation. We have already described in Sec. 1.2 the two modes in the context of biometrics. The verification mode of operation occurs in handwriting recognition applications such as a bank check recognition system where the recognizers have to verify the hypothesis generated by the numeric string in the courtesy field.[11] The identification mode is precisely how handwritten word recognizers operate in postal applications where the task is to recognize the city and street names.

3.1. *Verification systems*

Verification systems separate two classes: genuine and impostor verification attempts. By considering the combination task as a pattern classification problem in the M-dimensional space, Bayesian minimization of the misclassification cost results in the likelihood ratio combination function[25]:

$$f_{lr}(s^1, \dots, s^M) = \frac{p_{\text{gen}}(s^1, \dots, s^M)}{p_{\text{imp}}(s^1, \dots, s^M)} \tag{2}$$

p_{gen} and p_{imp} are M-dimensional densities of score tuples $\{s^1, \dots, s^M\}$ corresponding to the two classes — genuine and impostor verification attempts. We can estimate the densities p_{gen} and p_{imp} from the training data and use the above formula to calculate the combined score and threshold it. Alternatively, generic classification methods such as neural networks or SVMs can be used for direct classification of genuine and impostor classes.

3.2. *Identification systems*

We wish to investigate whether the likelihood ratio function found to be optimal for verification systems will be also optimal for identification systems. Suppose we performed a match of the input sample by all M matchers against all N classes and obtained MN matching scores $\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}$. Assuming equal prior class probabilities, the Bayes decision theory states that in order to minimize the misclassification rate, the sample should be classified as the one with the highest value of the likelihood function $p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_i)$. Thus, for any two classes ω_1 and ω_2 we can classify the input as ω_1 rather than ω_2 if

$$p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_1) > p(\{s_i^j\}_{i=1,\dots,N;j=1,\dots,M}|\omega_2) \tag{3}$$

Let us make an assumption that the scores assigned to each class are sampled independently from scores assigned to other classes; scores assigned to genuine class are sampled from the M-dimensional genuine score density, and scores assigned to impostor classes are sampled from the M-dimensional impostor score density:

$$\begin{aligned} &p(\{s_i^j\}_{i=1,\ldots,N;j=1,\ldots,M}|\omega_i) \\ &\quad = p(\{s_1^1,\ldots,s_1^M\},\ldots,\{s_{\omega_i}^1,\ldots,s_{\omega_i}^M\},\ldots,\{s_N^1,\ldots,s_N^M\}|\omega_i) \\ &\quad = p_{\text{imp}}(s_1^1,\ldots,s_1^M)\ldots p_{\text{gen}}(s_{\omega_i}^1,\ldots,s_{\omega_i}^M)\ldots p_{\text{imp}}(s_N^1,\ldots,s_N^M) \end{aligned} \tag{4}$$

After substituting (4) in (3) and canceling out the common factors we obtain the following inequality for accepting class ω_1 (rather than ω_2):

$$p_{\text{gen}}(s_{\omega_1}^1,\ldots,s_{\omega_1}^M)p_{\text{imp}}(s_{\omega_2}^1,\ldots,s_{\omega_2}^M) > p_{\text{imp}}(s_{\omega_1}^1,\ldots,s_{\omega_1}^M)p_{\text{gen}}(s_{\omega_2}^1,\ldots,s_{\omega_2}^M) \tag{5}$$

or

$$\frac{p_{\text{gen}}(s_{\omega_1}^1,\ldots,s_{\omega_1}^M)}{p_{\text{imp}}(s_{\omega_1}^1,\ldots,s_{\omega_1}^M)} > \frac{p_{\text{gen}}(s_{\omega_2}^1,\ldots,s_{\omega_2}^M)}{p_{\text{imp}}(s_{\omega_2}^1,\ldots,s_{\omega_2}^M)} \tag{6}$$

The terms in each part of the above inequality are exactly the values of the likelihood ratio function f_{lr} calculated for classes ω_1 and ω_2. Thus, the class maximizing the MN-dimensional likelihood function of inequality (3) is the same as the class maximizing the M-dimensional likelihood ratio function of inequality (6). Thus the likelihood ratio combination rule is optimal under the assumption of score independence. Our goal is to show that this assumption does not generally hold for real-life matchers, and, as a result, likelihood ratio combination method might be detrimental for the performance of the matching system.

It must be noted, *that the score independence assumption refers to scores assigned to different classes by the same matcher* ($\{s_i^j\}_{i=1,\ldots,N}$*: rows in Fig.* 1), *but not to the scores assigned to the same class by different matchers*. The latter score dependence has been investigated a number of times in classifier combination research with respect to the concept of classifier diversity (e.g. Ref. 7). We are interested in the former dependence,[23] which has received little attention thus far in the research community.[8,19]

The dependence of the matching scores obtained during a single identification trial is usually not taken into account by practitioners.[5,12,18] Apparently, all matching scores are derived independently from each other: the same matching process is applied repeatedly to all enrolled biometric templates or all lexicon words, and the matching score for one class is not influenced by the presence of other classes or the matching scores assigned to other classes. So it might seem that the matching scores are independent, but this is rarely true in practice. The main reason for the assumption to not hold is that all the matching scores produced during an identification trial are derived using the same input signal. For example, a fingerprint matcher, whose matching scores are derived from the number of matched minutia in enrolled and input fingerprint, will produce low scores for all enrolled fingerprints if the input fingerprint has only a few minutiae. Similarly, if the quality of the sensor is poor, all enrollees may receive a low score.

In our experiments, we measured the correlations between genuine scores and impostor scores produced in the same identification trials (Sec. 5.3), and obtained significant correlation values, especially for the considered word recognizers. The existence of such dependence between genuine and impostor scores increases the chances of the diminished performance of the likelihood ratio combination rule and its nonoptimality.

4. Illustrative Examples

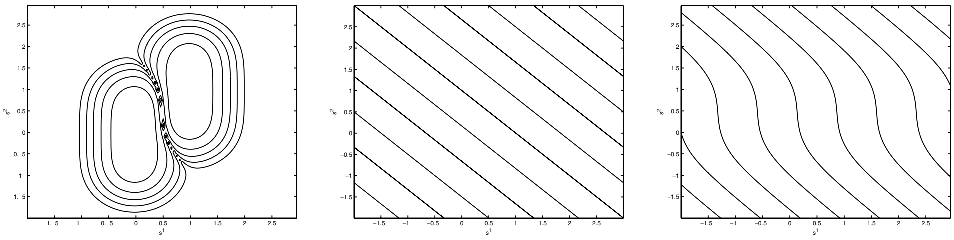
In order to further prove our claim, we present two examples that show that optimal combination functions for verification and identification systems are not necessarily the same. The examples also show that the derivation of the optimal combination function for identification systems is actually a nontrivial task.

4.1. *Example 1*

Let X_{gen}, X_{imp} and Y be independent two-dimensional random variables, and suppose that genuine scores in our identification system are sampled as a sum of X_{gen} and Y: $\mathbf{s}_{\text{gen}} = \mathbf{x}_{\text{gen}} + \mathbf{y}$, and impostor scores are sampled as a sum of X_{imp} and Y: $\mathbf{s}_{\text{imp}} = \mathbf{x}_{\text{imp}} + \mathbf{y}$, $\mathbf{x}_{\text{gen}} \sim X_{\text{gen}}$, $\mathbf{x}_{\text{imp}} \sim X_{\text{imp}}$ and $\mathbf{y} \sim Y$. Bold symbols here denote two-dimensional vector in the space (s^1, s^2). The variable Y provides the dependence between scores in identification trials. We assume that its value $\mathbf{y}$ is the same for all scores in any one identification trial.

Let X_{gen} and X_{imp} have gaussian densities $p_{X_{\text{gen}}}(s^1, s^2)$ and $p_{X_{\text{imp}}}(s^1, s^2)$ with unit covariance matrices. For any value of $\mathbf{y}$, conditional densities of genuine and impostor scores $p_{X_{\text{gen}}+Y|Y=\mathbf{y}}(s^1, s^2)$ and $p_{X_{\text{imp}}+Y|Y=\mathbf{y}}(s^1, s^2)$ are also Gaussian and independent. Since these Gaussians have the same covariance matrices, the optimal decision surfaces separating these two classes coincide with the contours of $s^1 + s^2 = c$.[21] The optimal combination rule for such conditional distributions coincides with the likelihood ratio combination function $f(s^1, s^2) = s^1 + s^2$, and this rule will be optimal for every identification trial and its associated value $\mathbf{y}$. The rule itself does not depend on the value of $\mathbf{y}$, so we can use it for every identification trial, and this is our optimal combination rule for the identification system. Figure 2(b) shows the contours of the optimal combination function in this identification system.

On the other hand, this rule might not be optimal for the verification system defined by the above score distributions. For example, if Y is uniformly distributed on the interval $0 \times [-1, 1]$, then the distributions of genuine and impostor scores $X_{\text{gen}} + Y$ and $X_{\text{imp}} + Y$ will be as shown in Fig. 2(a) and the optimal combination rule separating them will be as shown in Fig. 2(c). By changing the distribution of Y and thus the character of dependence between genuine and impostor scores, we will also be changing the optimal combination rule for the verification system. At the same time, the optimal combination rule for identification system will stay the same — $f(s^1, s^2) = s^1 + s^2$.



(a) (b) (c)

Fig. 2. (a) Two-dimensional distributions of genuine and impostor scores, for example, (b) contours of optimal combination rule in identification system, (c) contours of the likelihood ratio combination function.

If all that we know is the overall score distributions [Fig. 2(a)], then we do not have enough information to find the optimal combination function for the identification system case. If the scores are generated by the initial construction, linear combination function in Fig. 2(b) is the optimal combination function. If the score vectors having distributions in Fig. 2(a) are independent on their own, then the likelihood ratio combination in Fig. 2(c) is optimal for the identification system. Thus, there could be different optimal combination functions for identification systems with scores distributed as in Fig. 2(a), and the difference is determined by the nature of the score dependencies in identification trials.

Figures 2(b) and 2(c) show the possible optimal combination functions for identification and verification systems. This example illustrates that when searching for the optimal combination function one must take into account the mode (verification or identification) of the system.

4.2. *Example 2*

In this example, we are combining the scores of two matchers in an identification system with the number of classes, N, equal to 2. Thus matcher j, $j = 0$ or 1, outputs two scores s_1^j and s_2^j, with one of these scores being genuine, s_{gen}^j, and the other score being impostor, s_{imp}^j. Suppose, that the scores of matchers are sampled from bivariate normal distribution: $\{s_{\text{gen}}^j, s_{\text{imp}}^j\} \sim N(\{1,0\}, \Sigma_j)$, with

$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}$$

Thus the two scores of matcher 1 are independent and the two scores of matcher 2 are dependent if $\lambda \neq 0$. The marginal distributions of genuine scores of both matchers are normal N(1,1), and the marginal distributions of impostor scores of both matchers are normal N(0,1). Further, we will assume that the scores related to two matchers are independent; so the joint distribution of two genuine scores is normal $\{s_{\text{gen}}^1, s_{\text{gen}}^2\} \sim N(\{1,1\}, I)$ and the joint distribution of two impostor scores is normal $\{s_{\text{imp}}^1, s_{\text{imp}}^2\} \sim N(\{0,0\}, I)$, I is unit matrix.

If this system operates in verification mode, then the optimal score combination function, likelihood ratio, has same contours as $s^1 + s^2 = c$, and therefore, we can take $f_{lr}(s^1, s^2) = s^1 + s^2$ as our optimal combination function for the verification system. The distributions of genuine, $\{s^1_{\text{gen}}, s^2_{\text{gen}}\}$, and impostor, $\{s^1_{\text{imp}}, s^2_{\text{imp}}\}$, score pairs do not depend on λ, and optimal combination function for verification system, $f_{lr}(s^1, s^2)$, is the same for any choice of λ. But, as we show next, for the identification system the situation is different: optimal combination function and the performance of the combined system will depend on λ.

First, we can measure the identification system performance of the single second matcher for different values of λ. Table 1 presents some performance numbers derived by numerically integrating joint density of scores $p(s^2_{\text{gen}}, s^2_{\text{imp}}) = N(\{1, 0\}, \Sigma_2)$ over the area $s^2_{\text{gen}} > s^2_{\text{imp}}$. The identification system performance increases with the increase of λ. Intuitively this can be explained as following: if we have a positive correlation between genuine and impostor scores, for a high impostor score we have bigger probability that genuine will also be high, and the identification attempt will still succeed; similarly for low genuine scores we have bigger probability of even lower impostor scores. For negatively correlated scores ($\lambda < 0$) we observe a decrease in performance.

In order to calculate the identification system performance of combination function f, we numerically integrate $p(s^1_{\text{gen}}, s^2_{\text{gen}}, s^1_{\text{imp}}, s^2_{\text{imp}}) = p(s^1_{\text{gen}}, s^1_{\text{imp}})p(s^2_{\text{gen}}, s^2_{\text{imp}})$ over the region $f(s^1_{\text{gen}}, s^2_{\text{gen}}) > f(s^1_{\text{imp}}, s^2_{\text{imp}})$. The performance of the likelihood ratio combination function $f = f_{\text{lr}}$ is given in the fourth row in Table 1. Its performance reflects the change in performance of matcher 2: the better matcher 2 performs, the better is the performance of likelihood ratio combination. But notice that for large values of λ (e.g. $\lambda = 0.7$) the performance of likelihood ratio gets worse than the performance of the single matcher 2. The decrease in performance clearly indicates that likelihood ratio might not be an optimal combination function for identification systems. It is also possible to perform simple experiments by considering weighted sum combination functions $f_w(s^1, s^2) = ws^1 + (1 - w)s^2$, with bigger weight assigned to matcher 2 (with better performance in identification mode); by the proper choice of w it is easy to achieve better performance than using f_{lr}.

It turns out that it is possible to exactly derive the optimal combination function for the identification system in our example. Suppose that in one identification trial, we obtained the following scores from both matchers: $\{s^1_1, s^1_2\}$ from matcher 1 and

Table 1. Identification system performance (the frequency of top score being genuine) of single matchers and their combinations for Example 2.

λ	0	0.3	0.5	0.7	−0.5
Matcher 1	76.01%	76.01%	76.01%	76.01%	76.01%
Matcher 2	76.01%	80.08%	84.11%	90.13%	71.81%
Likelihood ratio	84.13%	86.09%	87.58%	89.25%	81.44%
Optimal combination	$s_1 + s_2$	$s_1 + \frac{10}{7}s_2$	$s_1 + 2s_2$	$s_1 + \frac{10}{3}s_2$	$s_1 + \frac{2}{3}s_2$
Optimal performance	84.13%	86.47%	88.96%	92.95%	81.93%

$\{s_1^2, s_2^2\}$ from matcher 2. The combination function f produces combined scores $S_1 = f(s_1^1, s_1^2)$ and $S_2 = f(s_2^1, s_2^2)$. By comparing S_1 and S_2, we decide which of two classes, 1 or 2, is genuine or impostor. In order to minimize the classification error, we have to use optimal Bayesian classification: classify the sample as class 1 instead of class 2, if and only if

$$p(\{s_1^1, s_2^1\}, \{s_1^2, s_2^2\}|\text{class 1 is genuine})$$
$$> p(\{s_1^1, s_2^1\}, \{s_1^2, s_2^2\}|\text{class 2 is genuine}). \tag{7}$$

So, the optimal combination function f should be such that $f(s_1^1, s_1^2) > f(s_2^1, s_2^2)$ if and only if Eq. (7) holds. After utilizing the independence of matchers ($p(\{s_1^1, s_2^1\}, \{s_1^2, s_2^2\}|\cdots) = p(\{s_1^1, s_2^1\}|\cdots)p(\{s_1^2, s_2^2\}|\cdots)$), we substitute the given normal densities of score pairs produced by each matcher:

$$p(\{s_1^j, s_2^j\}|\text{class 1 is genuine}) = N\left(\begin{pmatrix} s_1^j \\ s_2^j \end{pmatrix}; \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \Sigma_j\right) \tag{8}$$

$$p(\{s_1^j, s_2^j\}|\text{class 2 is genuine}) = N\left(\begin{pmatrix} s_2^j \\ s_1^j \end{pmatrix}; \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \Sigma_j\right) \tag{9}$$

After substitution, we can transform Eq. (7) into the following inequality:

$$s_1^1 + s_1^2 * \frac{1}{1-\lambda} > s_2^1 + s_2^2 * \frac{1}{1-\lambda} \tag{10}$$

Therefore, we can take the following function as the optimal combination function for the identification system in our example:

$$f_{\mathrm{id}}(s^1, s^2) = s^1 + s^2 * \frac{1}{1-\lambda} \tag{11}$$

The combination based on f_{id} coincides with the combination based on f_{lr} only when $\lambda = 0$. In other cases, f_{id} performs better than f_{lr} in the identification operating mode. The last two rows in Table 1 contain samples of optimal combination function f_{id} for the identification mode and the corresponding correct identification rates. In all cases, f_{id} performs better than any single matching participating in combination.

4.3. *Discussion*

The examples presented in this section underline the complexity of the task of finding an optimal combination function for identification systems. Given sufficient number of training genuine and impostor score samples, we might be able to approximate the genuine and impostor score densities (e.g. Fig. 2(a)). Given such density approximations, we can deduce the optimal combination function for verification systems (Fig. 2(c)). But we would still not have a good method to derive the optimal combination function for identification systems (Fig. 2(b)). It is possible that the likelihood ratio combination rule of Fig. 2(c) is the optimal combination function in identification system. It will certainly be the case if the identification trial scores are independent.

Suppose that score dependence had to be accounted for between different classes and an optimal combination function had to be derived for identification systems as in the second example. The first problem that we would have faced is to properly represent the joint density of a set of scores produced by a single matcher, e.g. $p(s_1^1, s_2^1, \ldots, s_N^1)$. If the number of classes, N, is large, we might want to reduce the number of variables by considering score set statistics. The second problem would be the construction of the combination function f_{id} given the reconstructed joint densities. Note, that in order to derive a combination function from Eq. (7), scores are to be separated related to different classes, so that the combination function would operate only on the scores related to a single class: $f(s_1^1, s_1^2) > f(s_2^1, s_2^2)$. We were successful in doing so in Example 2 only due to a particular form of score densities (Gaussian). In the general case, such derivation might be difficult to achieve.

As an additional consequence of our examples, we can assert that any score normalization based on reconstructed genuine and impostor score densities does not provide an answer to finding optimal combination function in identification systems. The only normalizations which might be beneficial for combinations in identification systems (assuming that we have a trainable combination algorithm able to take care of simple score translation normalizations) will be based on considering sets of identification trial scores similar to T-normalization.[4] Our paper[23] contains a deeper discussion on such normalizations.

A final corollary of the examples is that the training of optimal combination function for identification systems requires simultaneous consideration of the genuine and impostor scores from the same identification trials. In particular, we cannot simply take a set of all impostor scores and mix them. By doing so, the training will take place on the genuine and impostor densities, giving a combination algorithm trained for verification rather than for identification systems.

5. Experiments

In this section, we present the experimental results to support the claims made in this paper. First, in Secs. 5.1 and 5.2, we introduce the considered identification systems, handwritten word recognition and biometric person identification, and define the testing procedures. The next section presents the analysis of the dependence between scores assigned to different classes in all considered matchers. Section 5.5 presents the results on likelihood ratio and the weighted sum combination methods.

5.1. *Handwritten word recognizers*

We consider the application of handwritten word recognizers in the automatic processing of United Kingdom mail. The destination information of the mail piece contains the name of the postal town or county. After automatic segmentation of the mail piece image, the goal of the handwritten word recognizer is to match the hypothesized town or county word image against a lexicon of possible names, which contains 1681 entries.

We use two handwritten word recognizers for this application: Character Model Recognizer (CMR)[10] and Word Model Recognizer (WMR).[16] Both recognizers employ similar approaches to word recognition: they oversegment the word images, match the combinations of segments to characters and derive a final matching score for each lexicon word as a function of the character matching scores.

Our data consists of three sets of word images of approximately the same quality. The data was initially provided as these three subsets and therefore, we did not regroup them. The images were manually truthed and only those images containing any of the 1681 lexicon words were retained. The word recognizers were run on these images and their match scores for all 1681 lexicon words were saved. Note, that both recognizers reject some lexicon entries if, for example, the lexicon word is too short or too long for the presented image. We assume that in real systems, such rejects will be dealt with separately (it is possible that the lexicon word corresponding to image truth will be rejected), but for our combination experiments we keep only the scores of those lexicon words which are not rejected by either of the recognizers. Thus for each image I_k we have a variable number N_k of score pairs $(s_i^{\mathrm{cmr}}, s_i^{\mathrm{wmr}})$, $i = 1, \ldots, N_k$ corresponding to nonrejected lexicon words. One of these pairs corresponds to the true word of the image which we refer to as "genuine" scores, and the other "impostor" score pairs correspond to nontruth words.

After discarding images with nonlexicon words, and images where the truth word was rejected by either recognizer, we are left with three sets of 2654, 1723 and 1770 images and related sets of score pairs. We will refer to the attempt of recognizing a word image as an identification trial. Thus each identification trial has a set of score pairs $(s_i^{\mathrm{cmr}}, s_i^{\mathrm{wmr}})$, $i = 1, \ldots, N_k$ with one genuine score pair and $N_k - 1$ impostor pairs. The scores of each recognizer were also linearly normalized so that each score is in the interval [0,1] and the bigger score implies a better match.

Since our data was already separated into three subsets, we used this structure for producing the training and testing sets. Each experiment was repeated three times. Each time one subset was used as a training set, and the other two sets were used as test sets. The final results are derived as averages of these three training/testing phases.

5.2. *Biometric person matchers*

We used 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 left index "li" finger matches and right index "ri" finger matches. For experiments, we used four combinations involving both fingerprint and face score subsets: "li&C", "li&G", "ri&C" and "ri&G".

Though the BSSR1 score set has a subset of scores obtained from the same physical individuals, this subset is rather small — 517 identification trials with 517 enrolled persons. In our previous experiments[22] we used this subset, but the number of failed identification attempts for most experiments was less than ten and it is difficult to compare algorithms with those few negatives. Therefore, we used larger subsets of fingerprint and face matching scores of BSSR1 by creating virtual persons. The fingerprint scores of a virtual person come from a physical person and the face scores come from a different individual. The scores are not reused, and thus we are limited to a maximum of 6000 identification trials and a maximum of 3000 classes (or enrolled persons). Some enrollees and some identification trials are also required to be discarded since the corresponding matching scores were invalid probably due to enrollment errors. Finally, we split the data into two parts — 2991 identification trials with 2997 enrolled persons, with each part used as training and testing sets in two phases.

5.3. *Dependence of matching scores assigned to different classes*

We have made a key observation in Sec. 3 that the likelihood ratio combination rule might be optimal for identification systems if the matching scores assigned to the different classes by the same classifier are statistically independent. In order to test the score independence assumption we calculated the correlations between the matching scores assigned to the different classes: between genuine and impostor scores, and between two impostor scores obtained in the same identification trial. The results are presented in Fig. 2 for all matchers on our datasets.

The calculation of correlation values was performed using a subset of 2654 identification trials for word recognizers and one subset of 2991 identification trials for biometric matchers. In each trial, 50 random impostor scores were selected for calculating correlations. As a result, the calculation of $\mathrm{cor}(s_{\mathrm{gen}}, s_{\mathrm{imp},i})$ involves averaging of 2654×50 terms and the calculation of $\mathrm{cor}(s_{\mathrm{imp},i}, s_{\mathrm{imp},j})$ involves averaging of $2654 \times (50 \times 49/2)$ terms for word recognizers, and correspondingly 2991×50 and $2991 \times (50 \times 49/2)$ terms for biometric matchers. Nonzero correlation values confirm our hypothesis that the score independence assumption does not hold.

In addition, Ref. 25 contains calculations of correlation values between genuine scores and some functions of the impostor scores in the identification trials, for example, the correlation between the genuine score and the maximum of impostor scores obtained in the same trial. Those correlations were greater than what is

Table 2. Correlations between scores assigned to different classes during same identification trials.

Matchers	$\mathrm{cor}(s_{\mathrm{gen}}, s_{\mathrm{imp},i})$	$\mathrm{cor}(s_{\mathrm{imp},i}, s_{\mathrm{imp},j})$
CMR	0.043941	0.102119
WMR	0.364168	0.409941
li	0.106033	0.125387
ri	0.138155	0.149010
C	0.039175	0.094667
G	0.067829	0.125417

currently presented in Fig. 2. This indicates that the dependence between scores might be complex. Modeling these dependences constitutes a part of our work on classifier combinations utilizing identification models.[23]

5.4. *Description of combination algorithms*

We have explored two combination methods in this paper: likelihood ratio and the weighted sum. For the likelihood ratio combination, we reconstructed the densities using the Parzen window method with Gaussian kernels. The window widths are found by maximum likelihood leave-one-out cross validation method on a training set. Note that the reconstructed densities $p_{\text{gen}}(s^1, s^2)$ and $p_{\text{imp}}(s^1, s^2)$ of the likelihood ratio combination function 2 are two-dimensional. Given a large number of training samples, using two-dimensional kernels in the Parzen method results in a good approximation of the densities.[24]

We have compared the performance of the likelihood ratio rule with the weighted sum combination rule, which is one of the most frequently used rules in classifier combination tasks. The weighted sum rule is expressed by the combination function $f(s^1, \ldots, s^M) = w_1 s^1 + \cdots + w_M s^M$. The weights w_j are usually[20] chosen heuristically so that the better performing matchers have a bigger weight. The optimal weights can also be estimated for linear combinations of classifiers subject to the minimization of classification error.[27]

In our experiments, we have trained the weights so that the number of successful identification trials on the training set is maximized. The previously proposed methods of training resulting in the minimization of classification error[27] are not directly applicable due to much bigger number of classes in our case. Since we have only two matchers in all our configurations, it was possible to utilize a brute-force approach: we calculate the correct identification rate of the combination function $f(s^1, s^2) = ws^1 + (1 - w)s^2$ for different values of $w \in [0, 1]$, and find w corresponding to the highest recognition rate. Despite being brute-force, due to simplicity of weighted sum method, this approach was faster to train than likelihood ratio.

Note, that for the weighted sum method, as well as for likelihood ratio, we have separate training and testing subsets; the performance of this rule on test sets is slightly lower than the performance on training sets.

5.5. *Combination results*

The results of the combination using the likelihood ratio and the weighted sum are shown in Table 3. The numbers in the table refer to the correct identification rates, that is, the percentage of trials in which the genuine score receives the best score compared to all impostor scores of the same identification trial. For comparison, we also present the performance of single matchers used in combination.

Although for biometric combinations, the likelihood ratio combination method provided similar or better performance than the weighted sum rule, it performed very

Table 3. Correct identification rates for single matchers and their combinations by likelihood ratio and weighted sum.

Matchers	1st matcher is correct	2nd matcher is correct	Likelihood ratio	Weighted sum
CMR&WMR	54.76%	77.18%	69.84%	81.58%
li&C	81.41%	81.18%	97.24%	97.23%
li&G	81.41%	77.48%	95.90%	95.47%
ri&C	88.53%	81.18%	98.23%	98.09%
ri&G	88.53%	77.48%	97.14%	96.82%

poorly in the combination of handwritten word recognizers. In fact, it resulted in a performance lower than the performance of a single word recognizer. This would clearly imply that the likelihood ratio combination method might not be an optimal combination method for identification systems.

In order to verify that likelihood ratio combination for word recognizers was implemented correctly, we measured its performance in verification operating mode. Figure 3 presents ROC curves for likelihood ratio, as well as for weighted sum combination. As we expected, the likelihood ratio outperforms weighted sum and has superior performance with respect to single matchers.

Example 2 in Sec. 4.2 explains why the combination based on likelihood ratio function performs worse than a single matcher WMR. As in example, WMR

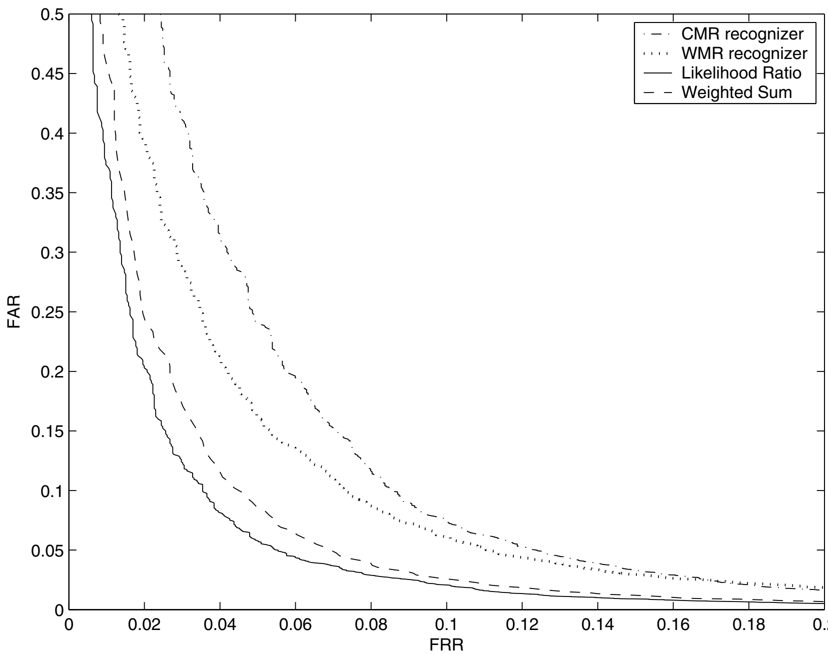


Fig. 3. The verification system performance of word recognizers and their combinations by likelihood ratio and weighted sum.

produces strongly dependent matching scores and has better identification system performance, than another combined matcher, CMR. The likelihood ratio fails to take the score dependence of WMR into account.

Though the likelihood ratio method seems to perform well for identification system with biometric matchers, this might not be the optimal combination method. In our previous work[26] we have presented some combination methods for identification systems (involving considered NIST BSSR1 datasets) which are able to outperform both likelihood ratio and weighted sum. But we still do not know if the proposed methods are optimal for identification system combinations.

6. Summary

In this paper, we show that for different operating scenarios of multiclassifier systems, namely verification and identification, we need to construct different combination algorithms to achieve optimal performance. This is due to the frequent dependence among the scores produced by each matcher during a single identification trial. The optimal combination algorithm for verification systems corresponds to the likelihood ratio combination function. It can be implemented by the direct reconstruction of this function with genuine and impostor score density approximations. Alternatively, many generic pattern classification algorithms can be used to separate the genuine and impostor scores in the M-dimensional score space, where M is the number of combined matchers.

The optimal combination algorithm for the identification systems is more difficult to realize. With the help of artificial examples we have shown that it is difficult to express the optimal combination function analytically. The experiments with existing score sets confirm the nonoptimality of likelihood ratio combination method for identification system. Though the weighted sum combination method can be trained for best identification system performance, due to the limited representation ability, it also might not achieve the optimal performance.

References

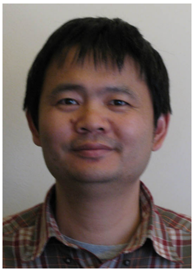
1. Nist biometric scores set. http://www.nist.gov/biometricscores/.
2. K. Al-Ghoneim and B. V. K. Vijaya Kumar, Unified decision combination framework, *Patt. Recogn.* **31**(12) (1998) 2077–2089.
3. H. Altincay and M. Demirekler, Undesirable effects of output normalization in multiple classifier systems, *Patt. Recogn. Lett.* **24**(9–10) (2003) 1163–1170.
4. R. Auckenthaler, M. Carey and H. Lloyd-Thomas, Score normalization for text-independent speaker verification systems, *Dig. Sign. Process.* **10**(1–3) (2000) 42–54.
5. R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha and A. W. Senior, The relation between the ROC curve and the CMC, *Fourth IEEE Workshop on Automatic Identification Advanced Technologies, 2005* (2005), pp. 15–20.
6. D. Bouchaffra, V. Govindaraju and S. N. Srihari, A methodology for mapping scores to probabilities, *IEEE Trans. Patt. Anal. Mach. Intell.* **21**(9) (1999) 923–927.

7. D. Chen, K. Sirlantzis, D. Hua and X. Ma, On the relation between dependence and diversity in multiple classifier systems, *Int. Conf. Information Technology: Coding and Computing, 2005. ITCC 2005*, Vol. 1 (2005), pp. 134−139.
8. J. M. Colombi, J. S. Reider and J. P. Campbell, Allowing good impostors to test, *Conf. Record of the Thirty-First Asilomar Conf. Signals, Systems & Computers, 1997*, Vol. 1 (1997), pp. 296−300.
9. R. P. W. Duin and D. M. J. Tax, Classifier conditional posterior probabilities, *Lecture Notes in Computer Science* **1451** (1998) 611−619.
10. J. T. Favata, Character model word recognition, *Fifth Int. Workshop on Frontiers in Handwriting Recognition* (Essex, England, 1996), pp. 437−440.
11. G. Kim and V. Govindaraju, Bank check recognition using cross validation between legal and courtesy amounts, *Int. J. Patt. Recogn. Artif. Intell.* **11**(4) (1997) 657−674.
12. P. Grother and P. J. Phillips, Models of large population recognition performance, *Proc. 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Vol. 2 (2004), pp. II−68−II−75.
13. Y. S. Huang and C. Y. Suen, A method of combining multiple experts for the recognition of unconstrained handwritten numerals, *IEEE Trans. Patt. Anal. Mach. Intell.* **17**(1) (1995) 90−94.
14. K. Ianakiev and V. Govindaraju, Deriving pseudo-probabilities of correctness given scores, *Pattern Recognition and String Matching*, eds. D. Chen and X. Cheng (Kluwer Publishers, 2002).
15. A. K. Jain, R. P. W. Duin and J. Mao, Statistical pattern recognition: A review, *IEEE Trans. Patt. Anal. Mach. Intell.* **22**(1) (2000) 4−37.
16. G. Kim and V. Govindaraju, A lexicon driven approach to handwritten word recognition for real-time applications, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(4) (1997) 366−379.
17. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, On combining classifiers, *IEEE Trans. Patt. Anal. Mach. Intell.* **20**(3)(1998), pp. 226−239.
18. Y. Lee, K. Lee, H. Jee, Y. Gil, W. Choi, D. Ahn and S. Pan, Fusion for multimodal biometric identification, *Audio- and Video-Based Biometric Person Authentication*, (2005), pp. 1071−1079.
19. A. E. Rosenberg and S. Parthasarathy, Speaker background models for connected digit password speaker verification, *Conf. Proc. 1996 IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96*, Vol. 1 (1996), pp. 81−84.
20. R. Snelick, U. Uludag, A. Mink, M. Indovina and A. Jain, Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems, *IEEE Trans. Patt. Anal. Mach. Intell.* **27**(3) (2005) 450−455.
21. S. Theodoridis and K. Koutroumbas, *Pattern Recognition* (Academic Press, 1999).
22. S. Tulyakov and V. Govindaraju, Classifier combination types for biometric applications, in *2006 IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR 2006), Workshop on Biometrics* (New York, USA, 2006).
23. S. Tulyakov and V. Govindaraju, Identification model for classifier combinations, *Biometrics Consortium Conf.* (Baltimore, MD, 2006).
24. S. Tulyakov and V. Govindaraju, Utilizing independence of multimodal biometric matchers, *Int. Workshop on Multimedia Content Representation, Classification and Security* (Istanbul, Turkey, 2006).
25. S. Tulyakov, V. Govindaraju and C. Wu, Optimal classifier combination rules for verification and identification systems, *7th Int. Workshop Multiple Classifier Systems* (Prague, Czech Republic, 2007).

26. S. Tulyakov, C. Wu and V. Govindaraju, Iterative methods for searching optimal classifier combination function, in *First IEEE Int. Conf. Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007* (2007), pp. 1−5.
27. N. Ueda, Optimal linear combination of neural networks for improving classification performance, *IEEE Trans. Patt. Anal. Mach. Intell.* **22**(2) (2000) 207−215.
28. L. Xu, A. Krzyzak and C. Y. Suen, Methods for combining multiple classifiers and their applications to handwriting recognition, *IEEE Trans. Syst. Man Cybern.* **23**(3) (1992) 418−435.

Sergey Tulyakov received the Ph.D. degree in computer science and engineering from the University at Buffalo in 2006. Currently, he is a research scientist at the Center for Unified Biometrics and Sensors, University at Buffalo.

His research interests include handwriting recognition, fingerprint and face biometric person authentication, and the combination of pattern classifiers.



Chaohong Wu received his Ph.D. (2007) and Masters degree (2001) from the Department of Computer Science and Engineering (CSE), University at Buffalo, State University of New York (SUNY).

His research interests are biometric authentication, pattern recognition and computer vison/image processing.

Venu Govindaraju received the B.Tech. (Hons) degree from the Indian Institute of Technology (IIT), Kharagpur, India, in 1986, and the Ph.D. degree from the University at Buffalo, State University of New York (SUNY), in 1992. He is a UB Distinguished Professor of computer science and engineering at the University at Buffalo, State University of New York (SUNY). He has authored more than 300 scientific papers, including over 60 journal papers and 30 book chapters. He has been the PI/co-PI of projects funded by the government and the industry for about 50 million dollars in the last 15 years. He is a fellow of the IEEE and the International Association of Pattern Recognition (IAPR).