Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 26 (2005) 679–684
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Fusion of biometric algorithms in the recognition problem

Andrew L. Rukhin [a,b,*], Igor Malioutov [a]

[a] Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[b] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

Received 13 May 2003; received in revised form 20 August 2004
Available online 28 October 2004

Abstract

This note concerns the mathematical aspects of fusion for several biometric algorithms in the recognition or identification problem. It is assumed that a biometric signature is presented to a system which compares it with a database of signatures of known individuals (gallery). On the basis of this comparison, an algorithm produces the similarity scores of this probe to the signatures in the gallery, which are then ranked according to their similarity scores of the probe. The suggested procedures define several versions of aggregated rankings. An example from the Face Recognition Technology (FERET) program with four recognition algorithms is considered.

Keywords: Aggregated algorithm; Gallery; Metrics on permutations; Probe; Permutation matrix; Similarity score

1. Introduction

This note concerns the mathematical aspects of a fusion for algorithms in the recognition or identification problem, where a biometric signature of an unknown person, also known as *probe*, is presented to a system. This probe is compared with a database of, say, N signatures of known individuals called the *gallery*. On the basis of this comparison, an algorithm produces the similarity scores of this probe to the signatures in the gallery, whose elements are then ranked according to their similarity scores of the probe. The top matches with the highest similarity scores are expected to contain the true identity.

A variety of commercially available biometric systems are now in existence; however, in many instances there is no universally accepted optimal algorithm. For this reason it is of interest to investigate possible aggregations of two or several different algorithms. See Xu et al. (1992), Ho et al. (1994), Lam and Suen (1995), Kittler et al. (1998), Jain et al. (2000) for a review of different

* Corresponding author.
E-mail address: rukhin@email.nist.gov (A.L. Rukhin).

0167-8655/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2004.09.021

schemes for combining multiple matchers. A common feature of many recognition algorithms is representation of a biometric signature as a point in a multidimensional vector space. The similarity scores are based on the distance between the gallery and the query (probe) signatures in that space (or their projections onto a subspace of a smaller dimension). Because of inherent commonality of the algorithms, the similarity scores and their resulting orderings of the gallery can be dependent for two different algorithms. For this reason traditional methods of combining different procedures, like classifiers in pattern recognition are not appropriate. Another reason for failures of popular methods like bagging and boosting (e.g. Schapire et al., 1998; Breiman, 2004) is that the gallery size is much larger than the number of algorithms involved. Indeed the majority voting methods used by these techniques (as well as in analysis of multi-candidate elections and social choice theory, Stern, 1993) are based on aggregated combined ranking of a fairly small number of candidates obtained from a large number of voters, judges or classifiers. The axiomatic approach to this fusion leads to the combinations of classical weighted means (or random dictatorship) (Marley, 1993).

As the exact nature of the similarity scores derivation is typically unknown, the use of nonparametric measures of association seems to be appropriate. The utility of such statistics such as rank correlation statistics, like Spearman's rho or Kendall's tau, for measuring the relationship between different face recognition algorithms, was reported by Rukhin et al. (2002). Rukhin and Osmoukhina (in press) employed the so-called copulas to study the dependence between different algorithms. They had shown that for common image recognition algorithms the strongest (positive) correlation between algorithms similarity scores happens for both large and small rankings. Thus, in all observed cases the algorithms behave somewhat similarly, not only by assigning the closest images in the gallery but also by deciding which gallery objects are most dissimilar to the given image exhibiting significant positive tail dependence. This finding is useful for construction of new procedures designed to combine several algorithms and also underlines the difficulty with a direct application of boosting techniques.

Notice that the methods of averaging or combining ranks can be applied to several biometric algorithms, one of which, say, is a face recognition algorithm, and another is a fingerprint (or gait, or ear) recognition device. Jain et al. (1999), and Snelick et al. (2003) discuss several experimental studies of multimodal biometrics, in particular, fusion techniques for face and fingerprint classifiers. They can be useful in a *verification* problem when a person presents a set of biometric signatures and claims that a particular identity belongs to these signatures.

The example considered in Section 4 comes from the Face Recognition Technology (FERET) program (Phillips et al., 2000) in which four recognition algorithms each produced rankings from galleries in three 1996 FERET datasets of facial images.

The authors are grateful to P. Grother and J. Phillips for these datasets.

2. Averaging of ranks via minimum distance

It is suggested to think of the action of an algorithm (its ranking) as a permutation π of N objects in the gallery. Thus $\pi(i)$ is the rank given to the gallery element i; in particular, if $\pi(i) = 1$, then the item i is the closest image in the gallery to the given probe, i.e., its similarity score is the largest.

If the goal is to combine K independent algorithms whose actions π_k, $k = 1, \ldots, K$, can be considered as permutations of a gallery of size N, then the combined (average) ranking of observed rankings $\pi_1, \ldots, \pi_K$ can be defined by the analogy with classical means. Namely, let $d(\pi, \sigma)$ be a distance between two permutations π and σ. The list of the most popular metrics (see Diaconis, 1988) includes Hamming's metric d_H, Spearman's L_2, Footrule L_1, Kendall's distance, Ulam's distance and Cayley's distance. The Spearman L_2 metric,

$$d_S^2(\pi, \sigma) = \sum_{i=1}^{N} [\pi(i) - \sigma(i)]^2,$$

and Footrule L_1 metric,

$$d_F(\pi,\sigma)=\sum_{i=1}^{N}|\pi(i)-\sigma(i)|,$$

(besides the metric d_H used here) are the most convenient in calculations. The "average permutation", $\hat{\pi}$, of $\pi_1,\ldots,\pi_K$ can be defined as the minimizer (in π) of

$$\sum_{j=1}^{K} d(\pi_j,\pi)\left(\text{or of}\sum_{j=1}^{K} d_S^2(\pi,\sigma)\right).$$

Then $\hat{\pi}$ can be taken as the action of the combined algorithm.

However, this approach does not take into account different precisions of different algorithms, Indeed, equal weights are implicitly given to all π_i, and the dependence structure of algorithms, which are to be combined, is neglected. A possible model for the combination of dependent algorithms employs a distance $d((\pi_1,\ldots,\pi_K),(\sigma_1,\ldots,\sigma_K))$ on the direct product of K copies of the permutation group. Then the combined (average) ranking $\hat{\pi}$ of observed rankings $\pi_1,\ldots,\pi_K$ is the minimizer (in π) of $d((\pi_1,\ldots,\pi_K),(\pi,\ldots,\pi))$. The simplest metric is the sum $\sum_{j=1}^{K} d(\pi_j,\pi)$ as above.

To define a more appropriate distance, we associate with a permutation π the $N\times N$ permutation matrix P with elements $p_{i\ell}=1$, if $\ell=\pi(i)$; $=0$, otherwise. A distance between two permutations π and σ can be introduced as the matrix norm of the difference between the corresponding permutation matrices.

For a matrix P, one of the most useful matrix norms is

$$\|P\|^2=\mathrm{tr}(PP^{\mathrm{T}})=\sum_{i,\ell}p_{i\ell}^2.$$

Here $\mathrm{tr}(A)$ denotes the trace of the matrix A.

For two permutation matrices P and S corresponding to permutations π and σ, the resulting distance $d(\pi,\sigma)=\|P-S\|$ essentially coincides with Hamming's metric,

$$d_H(\pi,\sigma)=N-\mathrm{card}\{i:\pi(i)=\sigma(i)\}.$$

For a positive definite symmetric matrix $\boldsymbol{C}$ (which is designed to capture dependence between π's) a convenient distance $d((\pi_1,\ldots,\pi_K),(\sigma_1,\ldots,\sigma_K))$ is defined as

$$d_C((\pi_1,\ldots,\pi_K),(\sigma_1,\ldots,\sigma_K))=\mathrm{tr}((\Psi-\Sigma)\boldsymbol{C}(\Psi-\Sigma)^{\mathrm{T}}),$$

with $\Psi=P_1\oplus\cdots\oplus P_K$ the direct sum of permutation matrices corresponding to $\pi_1,\ldots,\pi_K$, and Σ similarly defined for $\sigma_1,\ldots,\sigma_K$.

The optimization problem, which one has to solve for this metric, consists of finding the permutation matrix Π minimizing the trace of the block matrix formed by submatrices $(P_j-\Pi)\boldsymbol{C}_{jm}(P_m-\Pi)^T$, with $\boldsymbol{C}_{jm}$, $j,\ m=1,\ldots,K$ denoting $N\times N$ submatrices of the partitioned matrix $\boldsymbol{C}$. In other terms, one has to minimize

$$\sum_{j=1}^{K}\mathrm{tr}((P_j-\Pi)\boldsymbol{C}_{jj}(P_j-\Pi)^{\mathrm{T}})=\mathrm{tr}\left(\Pi\sum_j\boldsymbol{C}_{jj}\Pi^{\mathrm{T}}\right)-2\mathrm{tr}\left(\Pi\sum_j\boldsymbol{C}_{jj}P_j^{\mathrm{T}}\right)+\mathrm{tr}\left(\sum_j P_j\boldsymbol{C}_{jj}P_j^{\mathrm{T}}\right).\quad(1)$$

Matrix differentiation (Rogers, 1980) shows that the minimum is attained at the matrix,

$$\Pi_0=\left[\sum_j P_j\boldsymbol{C}_{jj}\right]\left[\sum_j\boldsymbol{C}_{jj}\right]^{-1}.$$

The matrix Π_0^{T} is stochastic, i.e., with $e=(1,\ldots,1)^{\mathrm{T}}$, $e^{\mathrm{T}}\Pi_0=e^{\mathrm{T}}$, but typically it is not a permutation matrix, and the problem of finding the closest permutation matrix, say, determined by a permutation π_0, remains. In this problem with $\Pi_0=\{\hat{p}_{i\ell}\}$ we seek the permutation $\hat{\pi}$ which maximizes $\sum_i\hat{p}_{i\pi(i)}$,

$$\hat{\pi}=\arg\max_{\pi}\sum_i\hat{p}_{i\pi(i)}.\quad(2)$$

An efficient numerical algorithm to determine π_0 is based on the so-called Hungarian method for the assignment problem. See for example Bazaraa et al., 1990, Section 10.7.

In this setting one has to use an appropriate matrix $\boldsymbol{C}$, which must be estimated on the basis of the training data; $\boldsymbol{C}^{-1}$ is the covariance matrix of all

permutation matrices $P_1,\dots,P_K$ in the training sample.

3. Linear aggregation

Since we have to estimate matrix C and numerical evaluation of (2) for large N can be difficult, one may look for a simpler aggregated algorithm.

Such an algorithm can be defined by the matrix P, which is a convex combination of the permutation matrices $P_1,\dots,P_K$, $P=\sum_{j=1}^K w_jP_j$. The problem is that of assigning non-negative weights (probabilities) $w_1,\dots,w_K$, such that $w_1+\cdots+w_K=1$, to matrices $P_1,\dots,P_K$. The fairness of all (dependent) algorithms can be interpreted as $EP_i=\mu$ with the same "central" matrix μ. In other terms, we assume that in average, for a given probe, all algorithms measure the same quantity, the main difference between them is their accuracy. The optimal weights $w_1^0,\dots,w_K^0$, minimize $E\|\sum_j w_j(P_j-\mu)\|^2$.

This optimization problem reduces to the minimization of

$$\sum_{1\leqslant j,m\leqslant K} w_jw_mE\mathrm{tr}(P_jP_m^{\mathrm{T}})-2K\sum_{1\leqslant j\leqslant K} w_jE\mathrm{tr}(P_j\mu^{\mathrm{T}}).$$

Note that for all m

$$E\mathrm{tr}(P_mP_m^{\mathrm{T}})=E\sum_{r,q}\delta_{r\pi(q)}=N,$$

and for $m\neq j$

$$E\mathrm{tr}(P_jP_m^{\mathrm{T}})=E\mathrm{card}\{\ell:\pi_m(\ell)=\pi_j(\ell)\}.$$

These "covariances" can be estimated from the available training data which can also be used to estimate μ by the grand mean $\hat{\mu}$ of all matrices in the training set, Then $d_j=E\mathrm{tr}(P_j\mu^{\mathrm{T}})=\sum_i\mu_{\pi_j(i)i}$ can be estimated by $\mathrm{tr}(P_j\hat{\mu}^{\mathrm{T}})$.

Let Σ denote the positive definite matrix formed by the elements $E\mathrm{tr}(P_mP_j^{\mathrm{T}})$, $m,\ j=1,\dots,K$. This matrix can be estimated by, say, $\widehat{\Sigma}$. With the vectors $w=(w_1,\dots,w_K)^{\mathrm{T}}$, and $d=(d_1,\dots,d_K)^{\mathrm{T}}$, our problem is that of finding

$$\min_{w^{\mathrm{T}}e=1}\left[w^{\mathrm{T}}\Sigma w-2w^{\mathrm{T}}d\right].$$

Basic linear algebra gives the form of the solution,

$$w^0=\Sigma^{-1}d+\frac{(1-e^{\mathrm{T}}\Sigma^{-1}d)}{e^{\mathrm{T}}\Sigma^{-1}e}\Sigma^{-1}e,$$

provided that Σ is nonsingular.

Thus, to implement the linear fusion, use the training data to get the estimated optimal weights

$$\hat{w}=\widehat{\Sigma}^{-1}\hat{d}+\frac{(1-e^{\mathrm{T}}\widehat{\Sigma}^{-1}\hat{d})}{e^{\mathrm{T}}\widehat{\Sigma}^{-1}e}\widehat{\Sigma}^{-1}e. \tag{3}$$

After these weights have been determined from the available data and found to be nonnegative, define a new combined ranking $\hat{\pi}_0$ on the basis of newly observed rankings $\pi_1,\dots,\pi_K$ as follows. Let the N-dimensional vector $Z=(Z_1,\dots,Z_N)$ be formed by coordinates $Z_i=\sum_{j=1}^K\hat{w}_j\pi_j(i)$, representing a combined score of element i. Put $\pi_0(i)=\ell$ if and only if Z_i is the ℓ-th smallest of $Z_1,\dots,Z_N$. In other terms, π_0 is merely the rank corresponding to Z. In particular, according to π_0 the closest image in the gallery is m_0 such that

$$\sum_{j=1}^K\hat{w}_j\pi_j(m_0)=\min_m\sum_{j=1}^K\hat{w}_j\pi_j(m).$$

This ranking π_0 is characterized by the property

$$\sum_{i=1}^N\left(\sum_{j=1}^K\hat{w}_j\pi_j(i)-\pi_0(i)\right)^2=\min_\pi\sum_{i=1}^N\left(\sum_{j=1}^K\hat{w}_j\pi_j(i)-\pi(i)\right)^2,$$

i.e., π_0 is the permutation that is the closest in the L_2 norm to $\sum_{j=1}^K\hat{w}_j\pi_j$. (See Theorem 2.2, p. 29 in Marden, 1995.)

Clearly some of the weights $\hat{w}$ can be negative. In this situation these weights must be replaced by 0, and the remaining positive weights are to be renormalized by dividing by their sum. This method can be easily extended to the situation when only partial rankings are available, i.e., when only the several top ranks are given. In this case one has to consider metrics on the coset space of all permutations with respect to the set of all permutations leaving the first several ranks fixed. Critchlow (1985) discusses mathematical properties of these metrics.

Table 1
Size of FERET datasets

	D1	D2	D3
Gallery size	1196	552	644
Probe size	234	323	399

4. Example: FERET data

In order to evaluate the proposed fusion methods, four face-recognition algorithms were selected for aggregation (I: MIT, March 95; II: USC, March 97; III: MIT, Sept 96; IV: UMD, March 97). In accordance with the Face Recognition Technology (FERET) protocol, these algorithms were ran on three 1996 FERET datasets of facial images, dupII (D1), dupI training (D2), and dupI testing (D3) (Table 1), yielding similarity scores between gallery and probe images. These scores were used for training and evaluating the new classifiers; all methods were trained and tested on different datasets of similarity scores.

The primary measures of performance used for evaluation were the recognition rate, or the percent of probe images classified at rank 1 by the methods, and the mean rank assigned to the true images. Moreover, the relative recognition abilities were differentiated by the Cumulative Match Characteristic (CMC) Curve, which is a plot of the rank against the cumulative match score (the percent of images identified below the rank).

On different pairs of training and testing datasets the overall recognition rate of the method fell short of this algorithm by 15% in the worst case and surpassed it by 2% in the best case (Table 2). The mean ranks of the two algorithms were generally within 5 ranks of each other. In terms of CMC curves, the method of weighted averaging of ranks (3) outperformed all but the best of constituent algorithms, the algorithm II, which was better in the range of ranks from 1 to 30 (Fig. 1). It looks like this phenomenon is general for linear weighting, namely for small ranks the best algorithm outperforms (3) for any weights giving this particular algorithm a weight smaller than 1. As a matter of fact, the weighted averaging method outperformed all of the four algorithms in the interval of ranks from 30 to 100 in the D2 dataset (Fig. 2). For this method there was about an 85% chance of the true image being ranked 50 or below, which significantly narrowed down the number of possible candidates, from more than a 1000 images to only 50.

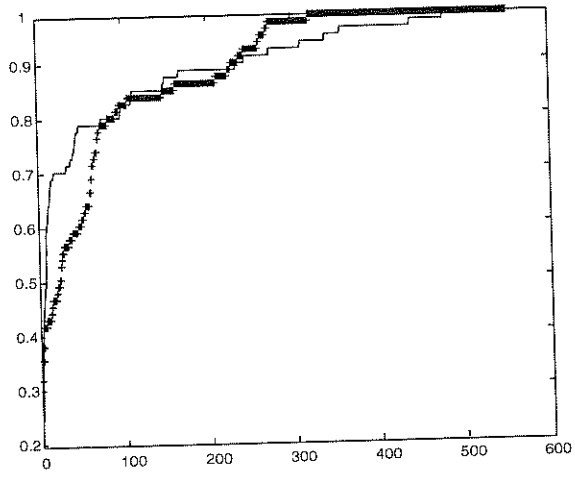

Fig. 1. Graphs of the cumulative match curves for algorithm II (marked by +) and the linear aggregation (marked by −).

The experiment showed that the weights derived from training for the different algorithms were all close (the last column of Table 2), which suggested that equal weights might be given to the different rankings. Although a simple averaging of ranks is a viable alternative to weighted averaging in terms of its computational efficiency, in our examples it was consistently inferior to the method (3) and the benefit of training seems apparent.

Table 2
Percent of images at rank 1

Dataset	Training	(3)	I	II	III	IV	Weights
D2	D3	48.6	26.0	59.8	47.1	37.1	(0.22, 0.32, 0.22, 0.24)
D3	D2	67.2	48.4	65.7	72.4	61.4	(0.20, 0.29, 0.25, 0.26)
D1	D3	36.3	17.1	52.1	26.1	20.9	(0.24, 0.27, 0.24, 0.25)

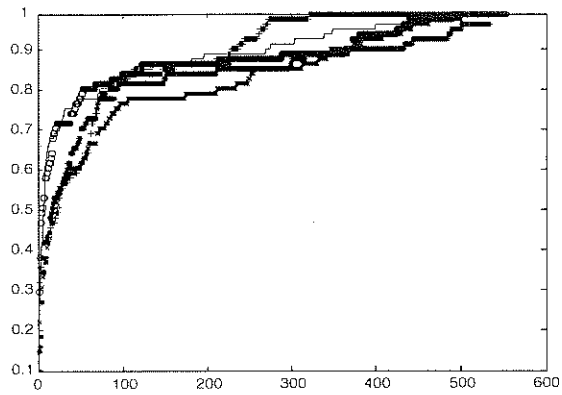

Fig. 2. Graphs of the cumulative match curves for algorithms I–IV (marked by $*$, $+$, $\circ$, $\times$) and the linear aggregation (3) (marked by $-$).

We never encountered negative weights obtained from (3). Moreover, the matrix $\widehat{\Sigma}$ must have positive elements, which suggests to use as weights the coordinates of the normalized eigenvector (with positive elements) corresponding to the largest (positive) eigenvalue. These weights turned out to be close to those found in (3). For example, when D3 is the training set, the corresponding vector is (0.17, 0.32, 0.26, 0.25).

References

Bazaraa, M.S., Jarvis, J.J., Sherali, H.D., 1990. Linear Programming and Network Flows. Wiley, New York.

Breiman, L., 2004. Population theory for boosting ensembles. Ann. Statist. 32, 1–11.

Critchlow, D.E., 1985. Metric Methods for Analyzing Partially Ranked Data. Springer, New York.

Diaconis, P., 1988. Group Representations in Probability and Statistics. Institute of Mathematical Statistics, Hayward, CA.

Ho, T.K., Hull, J.J., Srihari, S.N., 1994. Decision combination in multiple classifiers system. IEEE Trans. Pattern Anal. Mach. Intell. 16, 66–75.

Jain, A.K., Bolle, R., Pankanti, S., 1999. Personal Identification in Networked Society. Kluwer, Dordrecht.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell. 22, 4–37.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Anal. Mach. Intell. 20, 66–75.

Lam, L., Suen, C.Y., 1995. Optimal combinations of pattern classifiers. Pattern Recog. Lett. 16, 945–954.

Marden, J.I., 1995. Analyzing and Modeling Rank Data. Chapman&Hall, London.

Marley, A.A.M., 1993. Aggregation theorems and the combination of probabilistic rank orders. In: Fligner, M.A., Verducci, J.S. (Eds.), Probability Models and Statistical Analyses for Ranking Data, Lecture Notes in Statistics, vol. 80. Springer, New York.

Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J., 2000. The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Trans. Pattern Anal. Mach. Intell. 22, 1090–1104.

Rogers, G.S., 1980. Matrix Derivatives. Marcel Dekker, New York.

Rukhin, A., Grother, P., Phillips, J., Leigh, S., Newton, E., 2002. Algorithm evaluation and comparison. In: Proceedings of ICRP 2002 conference, vol. 2, Quebec City, QC, Canada.

Rukhin, A., Osmoukhina, A., in press. Nonparametric measures of dependence for biometric data studies. J. Stat. Plan. Inf. 27. http://www.math.umbc.edu/rukhin/papers/index.html.

Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. Ann. Statist. 26, 1651–1686.

Snelick, R., Indovina, M., Yen, J., Mink, A., 2003. Multimodal biometrics: issues in design and testing. In: Proceedings of the 5th international Conference on Multimodal Interfaces. Vancouver, BC, Canada.

Stern, H., 1993. Probability models on rankings and the electoral process. In: Fligner, M.A., Verducci, J.S. (Eds.), Probability Models and Statistical Analyses for Ranking Data, Lecture Notes in Statistics, vol. 80. Springer, New York.

Xu, L.A., Krzyzak, A., Suen, C.Y., 1992. Methods of combining multiple classifiers and their applications. IEEE Trans. Syst. Man Cybernet. 22, 418–435.