Improving the Accuracy of a Score Fusion Approach Based on Likelihood Ratio in Multimodal Biometric Systems

Emanuela Marasco and Carlo Sansone

Dipartimento di Informatica e Sistemistica,
Università degli Studi di Napoli Federico II
Via Claudio, 21 I-80125 Napoli, Italy
{emanuela.marasco,carlosan}@unina.it

Abstract. Multimodal biometric systems integrate information from multiple sources to improve the performance of a typical unimodal biometric system. Among the possible information fusion approaches, those based on fusion of match scores are the most commonly used. Recently, a framework for the optimal combination of match scores that is based on the likelihood ratio (LR) test has been presented. It is based on the modeling of the distributions of genuine and impostor match scores as a finite Gaussian mixture models. In this paper, we propose two strategies for improving the performance of the LR test. The first one employs a voting strategy to circumvent the need of huge datasets for training, while the second one uses a sequential test to improve the classification accuracy on genuine users.

Experiments on the NIST multimodal database confirmed that the proposed strategies can outperform the standard LR test, especially when there is the need of realizing a multibiometric system that must accept no impostors.

1 Introduction

Nowadays, companies spend a lot of efforts in security concerns, such as access control. One of the means to guarantee such aspects is providing reliable authentication methods to identify an individual requesting services of critical applications. A Biometric System recognizes individuals by their biological or behavioral characteristics [1], such as fingerprints, iris, facial patterns, voice, gait, etc. [2]. There is a growing interest in industries towards biometrics because they offer the highest level of security. In fact, biometric features are typically a portion of the body belonging to the person to be authenticated, instead of something he/she knows (e.g., passwords) or he/she possess (e.g., keys or badges). So, they cannot be stolen or forgotten.

However, systems relying on the evidence of a single modality are vulnerable in real world applications, since: (i) the *distinctiveness* of a single biometric feature is limited [3], (e.g., twins are non distinguishable by using the face as biometric feature); (ii) *noisy sensed data*, captured in unfavorable ambient conditions, can

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 509–518, 2009.

be incorrectly labeled, resulting in the rejection of a genuine user; (iii) there is the *non-universality* problem; it may not be possible to acquire meaningful data from a certain individual, because for such subject the quality of the features used to represent his/her required trait is not enough for a successful enrollment; (iv) finally, a biometric system have to face *spoofing* attacks, that can circumvent it by introducing fake biometric data that artificially reproduce physical traits.

The latest researches indicate that using a combination of biometric modalities, the human identification is more reliable [4]. A typical unimodal biometric system acquires the raw biometric data by an appropriate sensor module, extracts a feature set and compares it to the biometric sample (template) stored in a database, and then outputs a score used to establish the identity.

So, the information presented by multiple traits may be consolidated at various levels of recognition process. At feature extraction level, a new feature set is produced by fusing the features sets of multiple modalities, and this new feature set is used in the matching module. At match score level, the scores produced by multiple matchers are integrated, while at decision level the decisions made by the individual systems are combined. The integration at feature extraction level is expected to perform better, but the feature space of different biometric traits may not be compatible and most commercial systems do not provide access to information at this level. So, researchers found at score level a good compromise between the ease in realizing the fusion and the information content.

In literature three main approaches [5] are available to implement the fusion at score level [2]. First, the so called *Classifier-Based Scheme* [6] uses the output scores of each different matcher to construct a feature vector for training a classifier. This is accurate to correctly discriminate between genuine and impostor classes, regardless of the non-homogeneity of the score, but it typically requires a large training set. Second, the *Transformation-Based Scheme* [7] combines the match scores provided by different matchers: they are first transformed into a common domain (*score normalization*) and then are combined via a simple fusion rule such as *sum*, *min*, *max* or *weighted sum*. This approach is quite complex since it implicates a wide experimental analysis to choose the best normalization scheme and combination weights for the specific dataset of interest.

Last, the *Density-Based Scheme* [8] considers the match scores as random variables, whose class conditional densities are not *a prior* known. So, this approach requires an explicit estimation of density functions from the training data [2]. A recent method belonging to this category is the score fusion framework based on the Likelihood Ratio test, proposed by Nandakumar et al. in [5]. It models the scores of a biometric matcher by a mixture of Gaussians and perform a statistical test to discriminate between genuine and impostor classes. This framework produces high recognition rates at a chosen operating point (in terms of False Acceptance Rate), without the need of parameter tuning by the system designer once the method for score density estimation has been defined. Optimal performance, in fact, can be achieved when it is possible to perform accurate estimations of the genuine and impostor score densities. The Gaussian Mixture Model (GMM) lets to obtain reliable estimations of the distributions,

even if the amount of data needed for it increases as the number of considered biometrics increases. Moreover, as noted by the authors in [5], the performance of their method can be improved by using a suitable *quality measure* together with each score. Most of the available biometric systems, however, do not provides such measures.

Starting from the last considerations, in this paper we present two novel score fusion strategies based on the likelihood ratio scheme, that can be used when an high security level is needed. We propose both a sequential test and a voting strategy. By using them, on one hand we tried to implicitly use the quality information embedded into the scores. On the other hand, we obtained a system that demonstrated to be more robust than the original one with respect to the lack of data for training.

The rest of the paper is as follows: in Section 2 the likelihood ratio test is reviewed. In Section 3 the proposed strategy are illustrated and motivated, while the experimental results are presented in Section 4. Finally, some conclusions are drawn in Section 5.

2 Background and Open Issues

2.1 The Likelihood Ratio Test

Nandakumar and Chen [5] formulate the problem of Identity Verification in terms of hypothesis testing: let Ψ denote a statistical test for deciding if the hypothesis H:{*the score vector* $\mathbf{s}$ *belongs to the Genuine class*} has been correctly formulated. The choice is based on the value of observed match score and it lies between only two decisions: accepting H or rejecting it. As it is known [9], different tests should be compared with respect to the concepts of *size* and *power*, that are respectively the probability of accepting H when it is false (also called *False Accept Rate* - FAR) and the probability of accepting H when it is true (also called *Genuine Accept Rate* - GAR) [9]. In the context of *prudential decision making* [10], the NP lemma [9] recognizes that, in choosing between a hypothesis H and an alternative, the test based on the Likelihood Ratio test is the best because it maximizes the *power* for a fixed *size* [9]. Let

$$LR(\mathbf{s}) = \frac{f_{gen}(\mathbf{s})}{f_{imp}(\mathbf{s})} \tag{1}$$

be the *Likelihood Ratio* (LR), that is the probability of the observed outcome under H divided by the probability of assuming its alternative. As stated by the Neyman and Pearson theorem [9], the framework proposed by Nandakumar and Jain ensures that the most powerful test is the one, say $\Psi(\mathbf{s})$, that satisfies the equations (1) for some η

$$\Psi(\mathbf{s}) = \begin{cases} 1, & \text{when } LR(\mathbf{s}) \geq \eta \\ \\ 0, & \text{when } LR(\mathbf{s}) < \eta \end{cases} \tag{2}$$

where $\mathbf{s} = [s_1, s_2, ...s_K]$ is an observed set of K match scores that is assigned to the genuine class if $LR(\mathbf{s})$ is greater than a fixed threshold η, with $\eta \geq 0$.

2.2 The Estimation of Match Score Densities

As it is known in biometric literature [2], it is hard to choose a specific parametric form for approximating the density of genuine and impostor match scores, because the match distributions have a large tail, discrete components and not only one mode.

Given a training set, density estimation can be done by employing parametric or non-parametric techniques [11]. The non-parametric techniques do not assume any form of the density function and are completely data-driven; on the contrary, parametric techniques assume that the form of the density function is known (e.g., Gaussian) and estimate its parameters from the training data. The power of this scheme resides in its generality [12]: exactly the same procedure can be used also if the known functions are a mixture of Gaussians. In [5] the authors have proved the effectiveness of the GMM for modeling score distributions and of the likelihood ratio fusion test in achieving high recognition rates when densities estimations are based on GMM [5].

Let $\mathbf{s} = [s_1, s_2, ...s_K]$ denote the score vector of K different biometric matchers, where s_j is the random variable representing the match score provided by the j^{th} matcher, with $j = 1, 2, ..., K$. Let $f_{gen}(\mathbf{s})$ and $f_{imp}(\mathbf{s})$ denote the conditional joint density of the score vector $\mathbf{s}$ given respectively the genuine and impostor class. The estimates of $f_{gen}(\mathbf{s})$ and $f_{imp}(\mathbf{s})$ are obtained as a mixture of Gaussians:

$$\hat{f}_{gen}(s) = \sum_{j=1}^{M_{gen}} p_{gen,j} \Phi^K(s; \mu_{gen,j}, \Sigma_{gen,j}) \tag{3}$$

$$\hat{f}_{imp}(s) = \sum_{j=1}^{M_{imp}} p_{imp,j} \Phi^K(s; \mu_{imp,j}, \Sigma_{imp,j}) \tag{4}$$

where $\Phi^K(s; \mu; \Sigma) = (2\pi)^{-K/2} |\Sigma|^{-1/2} exp(-\frac{1}{2}(s-\mu)^T \Sigma^{-1}(s-\mu))$ denotes the Gaussian density with mean μ and covariance matrix Σ, and M_{gen} (M_{imp}) represents the number of mixture components. Mixture parameters can be approximated by employing the fitting procedure of Figuereido and Jain [13], that uses EM algorithm and Minimum Message Length (MML) criterion. It also estimates the optimal number of Gaussians and is able to treat discrete values by modeling them as a mixture with a very small variance represented as a regularization factor added to the diagonal of the covariance matrix.

Fusion based on GMM estimations achieves high performance [5], but there is an important drawback. In practice, one has to determine reliable models for estimations of genuine and impostor match score densities from the available score to be used for training. In absence of a large database, it is hard to obtain an accurate model, and this limitation is particularly true for multibiometric systems, as the number of considered biometrics increases.

3 The Proposed Approaches

As said in the Introduction, the quality of the acquired biometric data affects the efficiency of a matching process [14]. When the samples presented to a matcher are of poor quality, it cannot reliably distinguish between genuine and impostor users. For example, some true minutiae may not be detected in noisy fingerprint images, and missing minutiae may lead to errors. Moreover, as stated in the previous Section, when several biometrics are available, a not huge dataset could be not sufficient for having a proper density estimate by means of the GMM. So, we propose the following two approaches for improving the performance of the standard LR test:

1. An analysis of how the exclusion of some biometric modalities affects the GMM estimate: this approach (hereinafter denoted as *voting LR*) can be associated to the attempt of implicitly individuating degraded quality samples, when the quality measures are not available. In practice, given a K-dimensional score vector, we estimate the K conditional class joint densities of K-1 scores, by using a GMM technique. Then, we fixed for each of the K estimates a threshold η on the training set that gives rise to a FAR equal to 0%. When we have to judge a new sample, K LR tests are made on the K densities and if at least one of the LR tests recognizes the sample as genuine it is declared as genuine by the system. The ratio of this procedure lies in the fact that we want to detect if a particular score, say s_i, coming from a genuine sample, could be affected by a low quality. In this case, it can be expected that all the score vectors including s_i it will result in a low LR value, giving rise to a false rejection. Only the K-1 dimensional score vector that do not include s_i could have a LR value able to overcome the threshold. So, if at least one test is passed, the sample with a single modality affected by low quality can be correctly recognized. The choice of fixing η on the training set so as to obtain a FAR equal to 0%, is motivated by the need of having a system characterized by a FAR as low as possible. Since this approach uses only K-1 dimensional score vectors, it should be also more robust to the lack of training data.
2. A sequential likelihood ratio test (hereinafter denoted as *Sequential LR*) that introduces the option of suspending the judgment if the hypothesis is accepted or rejected with a not sufficient degree of confidence. This is a sort of sequential probability test (as stated in [15] by Wald) that use additional data for taking the final decision, when it is not possible to make a decision with a sufficient reliability by only using the initial observation. In this case $LR(\mathbf{s})$ is first compared with two different thresholds, say A_k and B_k:

$$\Psi(\mathbf{s}) = \begin{cases} 1, & \text{when } LR(\mathbf{s}) > A_k \\ Suspension & \text{when } B_k \leq LR(\mathbf{s}) \leq A_k \\ 0, & \text{when } LR(\mathbf{s}) < B_k \end{cases} \tag{5}$$

The thresholds A_k and B_k should be chosen so as to draw an uncertainty region around the value of the threshold η given by the standard LR test. In practice, a fraction ν of this threshold can be chosen, so as $B_k = (1 - \nu) \cdot \eta$ and $A_k = (1 + \nu) \cdot \eta$. If $LR(s) > A_k$, the decision is in favour of the genuine class, while if $LR(s) < B_k$, the decision is in favour of the impostor class. In the case of *suspension*, i.e., when $B_k \leq LR(s) \leq A_k$, the test procedure does not make any decision but activates a further step. The suspension of the judgment is motivated by the fact that samples that are quite near to the threshold could be misclassified due to the presence of one biometric trait acquired with a low quality. So, as a second step we propose to adopt the same approach presented in the previous case. In other words, K tests are made on score vectors of K-1 dimensions and the hypothesis is refused only if it is refused by all the K voting components.

4 Experimental Results

4.1 Dataset

The performances of our approaches are evaluated on a public domain database, namely, NIST-BSSR1 (Biometric Scores Set - Release 1). The BSSR1 is a *true* multimodal database i.e., the face and the fingerprint images coming from the same person at the same time. We performed experiments by employing the first partition made up of face and fingerprint scores belonging to a set of 517 people. For each individual, it is available a score coming from the comparison of two right index fingerprint, a score obtained by comparing impressions of two left index fingerprint, and two scores (from two different matchers, say C and G) that are the outputs of the matching between two frontal faces. So, in this case the match score for each modality indicates a *distance*. Then, our dataset consists in an unbalanced population composed by 517 genuine and 266,772 (517*516) impostor users.

4.2 Evaluation Procedure

We have performed a first experiment in which the training set is composed by half of the genuine and half of the impostor randomly selected from the dataset. The rest of the data are used as test set. The second experiment was directed to analyze how the reduction of the available scores for training affects the accuracy of the densities model. So, we performed another test in which the training set is halved with respect to the previous case, while the size of the test set remains unchanged. Both of these training-test partitioning have been randomly repeated 10 times and we report the average performance over the 10 runs.

4.3 Results

Tables 1 and 2 report the result of the two proposed approaches compared with the standard LR test. Moreover, we also report the K-1 dimensional score vector

Table 1. Test set results with a training set of equal size

	LR	LR on K-1 Matchers (LfInd,RxInd,FaceG)	Voting LR	Sequential LR $\nu = 0.2$	Sequential LR $\nu = 0.25$	Sequential LR $\nu = 0.30$
FAR	0.0%	0.0%	0.0%	0.0%	0.0%	0.000003%
GAR	95.60%	93.26%	97.77%	98.22%	98.22%	98.30%

Table 2. Test set results with a training set of halved size

	LR	LR on K-1 Matchers (LfInd,RxInd,FaceG)	Voting LR	Sequential LR $\nu = 0.2$	Sequential LR $\nu = 0.25$	Sequential LR $\nu = 0.30$
FAR	0.0%	0.0003%	0.0%	0.000009%	0.000011%	0.000011%
GAR	81.24%	95.35%	98.30%	88.09%	88.09%	88.01%

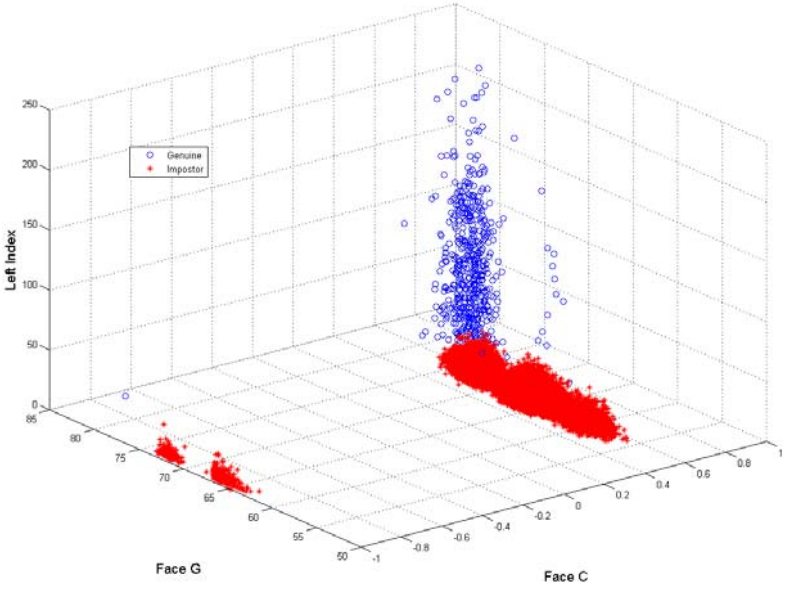

Fig. 1. Score distribution of Left Index, Face C and Face G from NIST-BSSR1

that allowed us to obtain the best results when used alone (in particular this score vector was composed by the outputs of the two fingerprint matchers and of the *Face G* matcher). Three values of ν have been considered, namely 0.2, 0.25 and 0.30.

Our system was designed for reducing to zero the number of accepted impostors. So, in order to have a fair comparison, the chosen operating point for each run of the standard LR test was obtained by fixing the FAR equal to 0% on the test set. The obtained threshold η is also used in the first step of the *sequential LR* approach.

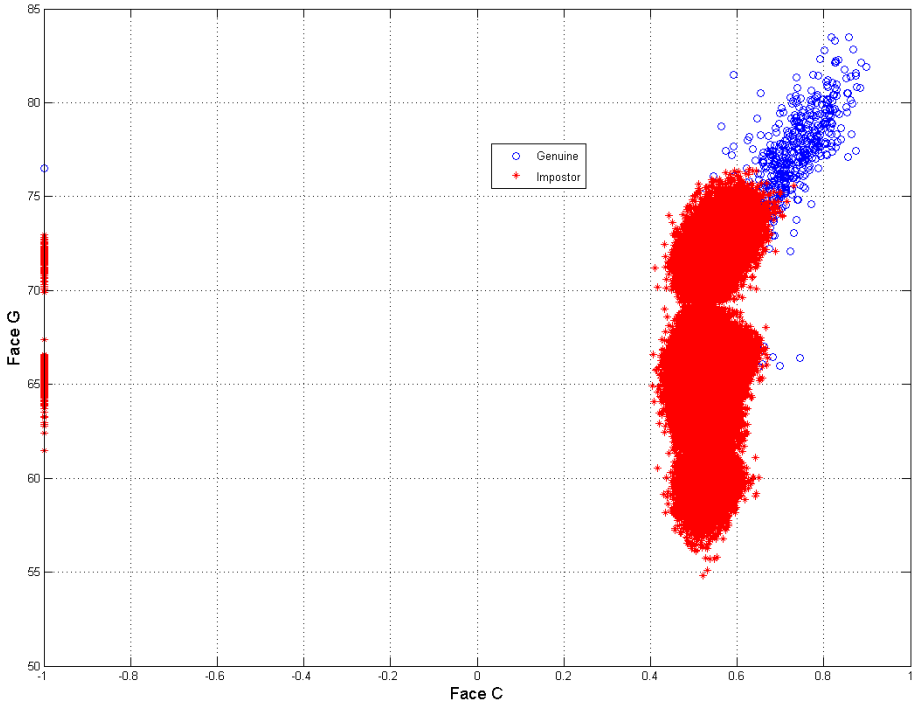

Fig. 2. Score distribution of Face C and Face G from NIST-BSSR1

From the previous tables it is evident that the *sequential LR* always improves the GAR obtainable with a standard LR, since its second stage is able to reduce misclassification of genuine samples with respect to the *pure* likelihood ratio, for those samples classified with a low degree of confidence.

Another interesting results is that the *voting LR* approach seems to be more robust with respect to the lack of training data. When only 25% of the data are used for training, in fact, it is able to significantly improve the GAR with respect to the standard LR approach. In this case, sequential LR is instead only able to slightly improve the LR performance in terms of GAR, but it also introduces few false accepted samples. On the contrary, when sufficient data for densities estimation are available, sequential LR achieves the best performance. All summarizing, it is worth noting that in both experiments the proposed approaches overperformed the standard LR test when a system at FAR=0% have to be realized.

Finally, is is interesting to consider the score distributions reported in Figures 1 and 2, where the joint distributions of *Left Index*, *Face C* and *Face G* and of *Face C* and *Face G* only are respectively shown. As it is evident (see also the considerations made by [16] on this problem), the use of only two

modalities significantly reduces the possibility of distinguish between genuine users and impostors. This is why we did not propose to further iterate the sequential test by considering, for example, also the joint densities of all the possible score pairs.

5 Conclusions and Future Work

In this paper we have proposed two approaches for combining multiple biometric matchers, starting from a density-based approach that use a likelihood ratio (LR) test, in order to set-up a biometric system that minimizes the number of false accepted users. The first approach is based on a voting strategy, while the second one on a sequential probability test. As a result, we obtained that if the density estimate of the standard LR method is accurate, the sequential test can reduce the misclassified samples belong to the uncertainty region, giving rise to very good results in terms of GAR. On the contrary, if the estimate of the density function of the standard LR is not so accurate, it is convenient to implement a voting system for classifying all the samples.

As future work we planned to extend our study to other multimodal biometric datasets.

References

1. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transaction on Circuits and Systems for Video 14(1), 4–20 (2004)
2. Ross, A., Jain, A.: Handbook in MultiBiometrics. Springer, Heidelberg (2008)
3. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
4. Jain, A., Ross, A.: Multibiometric systems. Comm. ACM 47(1), 34–40 (2004)
5. Nandakumar, K., Chen, Y., Dass, S., Jain, A.: Likelihood ratio-based biometric score fusion. IEEE Transaction on Pattern Analysis and Machine Intelligence 30(2), 342–347 (2008)
6. Ma, Y., Cukic, B., Singh, H.: A classification approach to multi-biometric score fusion. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 484–493. Springer, Heidelberg (2005)
7. Nandakumar, K., Jain, A., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition 38(12), 2270–2285 (2005)
8. Dass, S., Nandakumar, K., Jain, A.: A principled approach to score level fusion in multimodal biometric systems. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 1049–1058. Springer, Heidelberg (2005)
9. Graves, S.: On the neyman-pearson theory of testing. The British Journal for the Philosophy of Science 29(1), 1–23 (1978)
10. Lehmann, E., Romano, J.: Testing of Statistical Hypotheses. Springer, Heidelberg (2005)
11. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
12. Duda, R., Hart, P., Stork, D.: Pattern Classification, 2nd edn. Wiley-Interscience, Hoboken (2001)

13. Figuereido, M., Jain, A.: Unsupervised learning of finite mixture models. IEEE Transaction on Patterns Analysis and Machine Intelligence 24(3), 381–396 (2002)
14. Nandakumar, K., Chen, Y., Jain, A., Dass, S.: Quality-based score level fusion in multimodal biometric systems. Pattern Recognition 4, 473–476 (2006)
15. Wald, A.: Sequential tests of statistical hypotheses. The annals of Mathematical Statistics 16(2), 117–186 (1945)
16. Tronci, R., Giacinto, G., Roli, F.: Combination of experts by classifiers in similarity score spaces. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) S+SSPR 2008. LNCS, vol. 5342, pp. 821–830. Springer, Heidelberg (2008)