A Self-updating Multiexpert System for Face Identification

Andrea F. Abate[1], Maria De Marsico[2], Michele Nappi[1], and Daniel Riccio[1]

[1] DMI - Dipartimento di Matematica e Informatica, Università di Salerno, Fisciano (SA), Italy
{mnappi,driccio}@unisa.it
[2] DI - Dipartimento di Informatica, Sapienza Università di Roma, Italy
demarsico@di.uniroma1.it

Abstract. Multibiometric systems can solve a number of problems of single-biometry approaches. A source of flaws for present systems, both single-biometric and multibiometric, can be found in the lack of dynamic update of parameters, which does not allow them to adapt to changes in the working settings. They are generally calibrated once and for all, so that they are tuned and optimized with respect to standard conditions. In this work we investigate an architecture where single-biometry subsystems work in parallel, yet exchanging information at fixed points, according to the N-Cross Testing Protocol. In particular, the integrated subsystems work on the same biometric feature, the face in this case, yet exploiting different classifiers. Subsystems collaborate at a twofold level, both for returning a common answer and for tuning to changing operating conditions. Results demonstrate that component collaboration increases system accuracy and allows identifying unstable subsystems.

1 Introduction

Present biometric systems generally rely on a single classifier. The main drawback is that they are singly vulnerable to possible attacks, as well as little robust with respect to a number of problems (e.g. a voice altered by a cold or a dimly lit face). In the present work we will consider the combination of some popular classifiers to build a face identification system. We chose this biometry because it is contact-less, fast and fairly reliable. Face recognition, however, raises a number of non-easy to solve issues(pose, illumination or expression changes). Indeed, none of the existing approaches for face recognition is free from limitations when dealing with issues such as variation in expression, lighting, pose and acquisition time. Recognition techniques found in literature can be grouped in different classes. A first class includes linear methods. Among them, the two classic Principle Component Analysis (PCA) [1] and Linear Discriminant Analysis (LDA) [2] are widely used in the appearance-based approaches. They both solve the recognition problem within a representation space of lower dimension than image space. In general, LDA-based algorithms outperform PCA-based ones, but they suffer from the so-called *small sample size problem* (SSS) which

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 346–354, 2009.

exists in high-dimensional pattern recognition tasks. Neighborhood Preserving Embedding (NPE) [3] differs from PCA because it aims at preserving the local manifold structure instead of the global Euclidean structure though maintaining linearity of the approach. A weight matrix is built which describes the relationships between the data points in the ambient space, so that each data point can be represented as a linear combination of the neighboring ones. Then an optimal embedding is computed such that the neighborhood structure can be preserved in the dimensionality reduced space. In [4] the authors introduce the Orthogonal Locality Preserving Projections (OLPP) method, producing orthogonal basis functions called Laplacianfaces. The manifold structure is modeled by a nearest-neighbor graph preserving the local structure of the image space. Each face image in the image space is mapped into a low-dimensional face subspace, obtained by OLPP. Fractals are largely exploited in image compression and indexing, so that they have also been investigated in the field of face recognition. In [5], the fractal code of a face image is only used for training a neural network, which works as a classifier on the face database, while in [6], the Partitioned Iterated Function Systems (PIFS) have been explicitly adapted to recognize face.

A multiexpert system provides an effective alternative, as flaws of an individual system can be compensated by the availability of a higher number of cooperating algorithms [7]. This is particularly true if the subsystems exchange information at different levels, as we shall see in presenting the N-Cross Testing Protocol. According to the above considerations, not all classifiers are equally reliable on any input image. In [8] System Response Reliability (SRR) indices have been introduced to evaluate the reliability of each response of single subsystems; responses are considered for fusion only if the corresponding SRR is higher than a given threshold *th*. This improves global system performance, but we argue that we can go even further by considering the "history" of the system. We assume the existence of a *supervisor module* exploiting not only single subsystems responses and their reliability, but also the final global response, to evaluate the overall system state and update reliability thresholds of single subsystems. Such module would allow overcoming the invariance of present multibiometric architectures, by implementing an algorithm converging to an optimal parameters configuration independently from the starting one. Each subsystem takes as input an image, pertaining to the corresponding biometry, and extracts salient features; the resulting feature vectors are used to compute similarity scores, to be inputted to the supervisor module, after a normalization step. In building our multiexpert/multiclassifier system, we chose each subsystem classification method from the most popular linear and non-linear techniques in the state of the art. We aimed at combining results which suffer with significantly different measures from the image variation problems: LDA [2], OLPP [4] and NPE [3].

2 The Integration Scheme

Our multiexpert system significantly differs from the state of art in literature. The algorithms implemented by the recognition subsystems to classify the input

face are the ones presented above. The single subsystems cooperate at two different levels. In the first place, they exchange information about the produced subject lists. Such lists are the final result in traditional systems, while represent an intermediate result in our one. The exploited protocol is the N-Cross Testing Protocol [9]. Starting from this basic configuration, each subsystem can also return a measure of reliability for its own response, which can be used to better compute the global result. As a further increase in cooperation, a suitable supervisor module can further exploit reliability measures to obtain the final response and to update subsystems reliability parameters.

2.1 Reliability Margins

Subsystems involved in a multibiometric architecture might not be equally reliable, e.g. due to the possibly different quality of input. An unreliable response can trigger a further check. A *reliability* measure is then crucial for fusion. Some solutions use margins, measuring the "risk" associated to a single subsystem response after observing its scores. Poh and Bengio [10] introduce a confidence margin based on False Acceptance Rate (FAR) and False Rejection Rate (FRR). Many responses are marked as reliable, as the margin relies on an estimate of the actual distribution of genuine/impostor subjects' scores. This might be inappropriate when very high security is required. Moreover, frequentist approaches assume that the scores of the testing and development sets always originate from similar distributions. We adopt the new System Response Reliability (SRR) index [8], based on a system/gallery dependent metric, and measuring the ability of separating genuine subjects from impostors on a single probe basis. SRR can be computed by using as basic elements either the relative distance between the scores of the two first retrieved distinct identities, or the number of subjects near to the retrieved identity which are present in the gallery. SRR values always fall in the range $[0, 1]$. Each subsystem T_k computes, for each $s_{k,i}$, i=1, ..., in its set of responses, a *reliability measure* $srr_{k,i}$. Moreover, each subsystem T_k is characterized by an estimated threshold th_k, possibly updated over time [8], such that a response $s_{k,i}$ is considered as reliable only if $srr_{k,i} \geq th_k$. In general, th_k thresholds can be adjusted in a tuning phase, according to a set of samples with similar features to those of the set used for identification (the two sets should be different and disjoint). The limit is that thresholds do not take into account the performances of the other subsystems.

2.2 The Supervisor Module

A revisitation of the classical multibiometric schema, where subsystems act independently, given that a suitable fusion module is able to combine their single results, takes to global system self-tuning, with a much more flexible and robust architecture. One of the main limits of the subsystems combined within classical architectures, and also of the reliability measure in Section 2.1, is that they do not seize the main advantage of exploiting information coming from other subsystems. Each component works independently and final results give

no feedback for the overall system. On the contrary, assume the existence of a *supervisor module*, which still exploits single subsystem responses and their reliability to compute the final global response, but which also uses the latter to evaluate the overall system state and update its parameters. Such module would implement an algorithm to update single thresholds also according to the behavior of the other subsystems, so converging to an optimal configuration independently from the starting $\{th_1, th_2, ...th_N\}$ configuration. The algorithm distinguishes two cases:

- More identities I_j, $j \in \{1, 2, \ldots, |H|\}$ share the same maximum number of votes, e.g. when retrieved identities are all different with 1 vote each. If at least one T_k in any such group has $srr_k > th_k$, the system returns the identity retrieved by the subsystem with the higher $srr_k > th_k$, and the response is marked as reliable, otherwise the response is marked as unreliable.
- One identity I_j gets more votes than the others. I_j is returned and the response is marked as reliable.

In both cases, if the response is reliable, each subsystem T_k voting for the returned identity is rewarded by lowering its threshold th_k by an updating step us, unless its current srr_k is already above its th_k. Each other subsystem T_k is penalized by increasing its threshold th_k by us, unless its current srr_k is already below its th_k. In this way the supervisor module lowers thresholds of subsystems voting in agreement, considering such behavior a confirmation of reliability, and increases thresholds of discordant ones, compensating possible distortions (local persistent distortions like lighting variations, dirt on lens).

Such an architecture does not need an adjustment phase, since the system can start from a default configuration of its parameters and converge in any case towards an optimal one. The speed to reach such latter configuration is a significant system feature, so that it is important to define how to measure it. As we want to simulate the dynamic behavior of an online identification system, we assume that system time is beaten by performed recognition operations; we define a *probe sequence* $P = \{p_1, p_2, \ldots, p_n\}$ as a series of n probes presented to the system, sharing the same acquisition characteristics (normal conditions, right light, earrings, dirty lens). A subsystem equilibrium state (*steady state*) is given by the consecutive instants when threshold fluctuations are lower than a fixed μ, while *convergence speed* of a subsystem λ_k is defined as the ratio between the total variation of its threshold between two steady states, and the number of instants needed to obtain such transition. Total system convergence speed is defined as the minimum speed among all its subsystems, i.e. $\lambda = min_k \lambda_k$, $k \in \{1, 2, 3\}$.

2.3 Integrating the Supervisor Module into the The N-Cross Testing Protocol

In this architecture, N subsystems T_k, k=1, 2,, N, work in parallel, first in identification mode and then in look-up mode, and exchange information in fixed

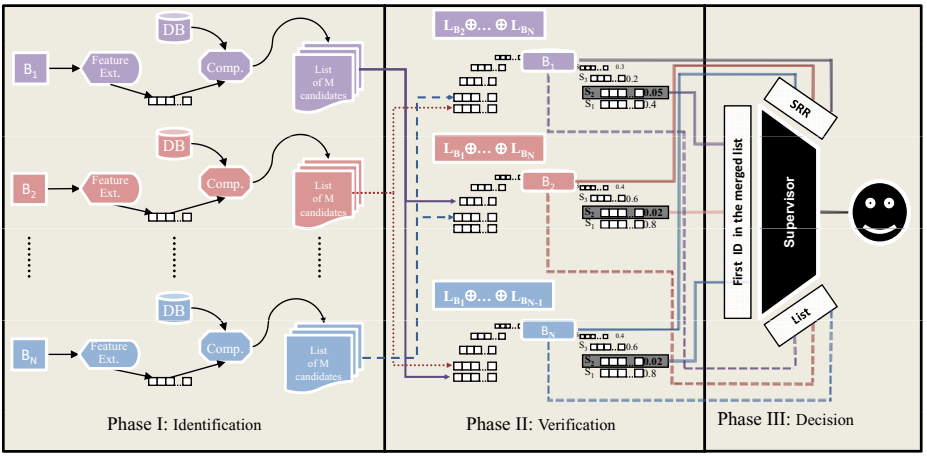

Fig. 1. The N-Cross Testing Protocol with SRR and Supervisor Module

points (Figure 1). At the beginning, each subsystem creates its own database of subjects, characterized by a peculiar set of image features. Face image data are then acquired for the probe subjects according to the specifications of each subsystem. We can identify three operation phases, namely identification, cross and testing. In identification phase, the N subsystems start up independently, by extracting biometric features and computing similarities according to the implemented classifiers. Each T_k, k=1, 2,, N, retrieves a list of candidate subjects from its specific database of enrolled subjects (gallery), ordered by similarity with the input: the lower the distance, the higher the similarity. All lists elements include the ID of a database subject and a numeric score for its similarity with the input. For a correct fusion, scores from different subsystems are made consistent through a suitable normalization function. Each subsystem also returns an estimate of its response reliability, namely a value for the SRR index. SRR is used in the cross operation. Afterwards the cross (verification) phase starts. Each T_k merges only lists coming from reliable companions. Subsystems with SRR value below their threshold still work during the fusion process, as they can receive and merge lists from reliable companions. We remind that each subsystem own list is not included in the merging operation in any case. In this way all subsystems produce a merged list, which only contains identities coming from reliable ones. A special case is when only one subsystem is reliable. During cross operation, all unreliable subsystems only receive its list, which so coincides with the merged one. The only reliable subsystem does not receive any list, so that it should return a null subject. This case is handled by adopting a specific rule, so that a reliable subsystem which does not receive any list during the cross phase can return its own list as the merged one. Shared subjects get a single score, which is the average of the original ones. Subjects belonging to only

one list retain that score. The merged list is resorted. The SRR index assigned to the first retrieved identity in each merged list L_p, is given by the average of reliability indexes of subsystems in a set $\tilde{L_p}$, namely those contributing to such list (remind that they must all be reliable), and can be defined as:

$$LSRR_p = \frac{1}{|\tilde{L_p}|} \sum_{T_i \in \tilde{L_p}} SRR_i. \quad (1)$$

Each subsystem sends its list to the supervisor module, together with its reliability index and with the first element in its merged list. The supervisor module uses the latter two to compute the final response and to update system thresholds.

3 Experimental Results

Experimental tests have been conducted in order to asses the performance of the chosen classifiers when run alone or when combined in more collaborative architectures. The classification methods to use during the experiments were selected from those described in Section 1, due to their historical reference value or to their good performances in the two-dimensional setting. They are are LDA, OLPP, NPE and PIFS.

The sets of face images that have been considered as testing benchmark, were extracted from AR Faces database [11]. This database contains 126 subjects (70 men and 56 women) each of them was acquired in two different sessions with 13 image sets each. Sets differ in expression (1 neutral, 2 smile, 3 anger, 4 scream), illumination (5 left light, 6 right light, 7 all side light), presence/absence of occlusions (8 sun glasses, 11 scarf), or combinations (9 sun glasses and left light, 10 sun glasses and right light, 12 scarf and left light, 13 scarf and right light). Neutral images from the set 1 have been considered as the system gallery. Seven probes (2, 3, 4, 5, 6, 8, 11) have been used for testing. The face object detector that we used to locate the relevant regions is open source and based on Haar features [12], implemented in the OpenCV library [13]. Unfortunately, any automatic detector available nowadays makes some errors. As a matter of fact, detection methods make up a research issue by themselves. When errors happen, they are corrected by hand. After all, the focus of our work is rather on the ability by a collaborative system to enhance single classifiers performances. All tests were performed on an Intel Pentium IV, 2.8Ghz with 512 Mb of RAM. Performances were measured in terms of Recognition Rate (RR), Equal Error Rate (EER), Cumulative Match Score (CMS), and, when using SRR index, Number of Reliable Responses (NRR).

The first experiment consisted in analyzing the performances of the different subsystems, which are summarized in Table 1. In the table, the better obtained results are in boldface for readability. We can notice that PIFS outperforms the other classifiers, on a regular basis in relation with RR and in most cases in relation with EER and CMS computed for rank 5. NPE appears as a fully complementary competitor, as it reaches the best performances of all classifiers,

Table 1. Performances of the single classifiers

Datasets	Classifiers											
	LDA			OLPP			NPE			PIFS		
	RR	EER	CMS(5)	RR	EER	CMS(5)	RR	EER	CMS(5)	RR	EER	CMS(5)
Set 2	0.944	0.044	0.991	0.925	0.094	0.972	**0.953**	**0.028**	**1.000**	**0.953**	0.039	0.981
Set 3	0.822	0.075	0.944	0.879	0.135	0.925	0.869	0.043	**0.981**	**0.962**	**0.036**	**0.981**
Set 4	0.430	0.170	0.738	0.402	0.251	0.710	0.551	**0.091**	**0.850**	**0.588**	0.108	0.822
Set 5	0.626	0.095	0.841	0.879	0.102	0.972	0.262	0.118	0.617	**0.953**	**0.020**	**0.990**
Set 6	0.206	0.108	0.813	0.636	0.132	0.860	0.206	0.175	0.439	**0.869**	**0.076**	**0.934**
Set 8	0.290	0.208	0.579	0.393	0.232	0.617	0.374	0.127	0.664	**0.850**	**0.06**	**0.943**
Set 11	0.065	0.325	0.243	0.178	0.352	0.383	0.028	0.429	0.112	**0.803**	**0.123**	**0.906**

in relation with EER and CMS for rank 5 (see for example EER = 0.028 for smiling subjects in set 2) just where PIFS obtains suboptimal results, even if in all other cases it obtains changeable results. From this point of view LDA, OLPP and NPE all show performances that vary with their robustness to specific variations. For example, NPE badly reacts to changes in light (sets 5 and 6, with RR = 0.262 and RR = 0.206 respectively) and even worse to significant occlusions, with RR = 0.028 and CMS(5) = 0.112 for subjects with scarf in set 11. In the latter case, PIFS is the only classifier reaching more than acceptable results. Notice the scarce performances of all classifiers with screaming subjects in set 4. In this case, the best result in relation to RR is obtained by PIFS with RR = 0.588, which is far below its average behavior. This can be explained with the fact that, as for screaming subjects, the whole face expression undergoes a deformation process, especially the mouth region where we have something comparable to an occlusion. As a matter of fact, the same low performances are obtained by almost all classifiers with subjects wearing a scarf in set 11.

Given the discussed results, the following tests aimed at investigating if a more strict cooperation among subsystems running the given classifiers can improve the best performances obtained by the single methods. The results are reported in Table 2. We remind that NRR measures the number of reliable responses. As the simple NCT does not exploit reliability measures, the NRR column in that case always shows the total number of responses.

The basic N-Cross-Testing architecture provides worse performances than PIFS alone, as for RR at least, while EER follows an alternating trend. This is surely due to the fact that the global response is influenced by the scarce results obtained by the other classifiers. However, it is worth noticing how the result obtained for the screaming subjects of set 4 is better than PIFS one, that was the worst for this classifier, and the best till now for that set. In this particular case, were PIFS performances were low, the contribution of the other classifiers was significant. This seems to suggest that, when only one classifier shows a good behavior, its contribution might be overwhelmed by bad results from the others. However, in a situation where all classifiers encounter some

Table 2. Performances of N-Cross Testing in different cooperation settings

Datasets	Architectures					
	Simple N-Cross-Testing			Supervised N-Cross-Testing		
	RR	EER	NRR	RR	EER	NRR
Set 2	0.962	0.018	126	**0.990**	**0.004**	121
Set 3	0.971	0.014	126	**0.989**	**0.005**	116
Set 4	0.652	0.17	126	**0.962**	**0.018**	94
Set 5	0.744	0.127	126	**0.940**	**0.029**	118
Set 6	0.584	0.207	126	**0.905**	**0.047**	112
Set 8	0.522	0.238	126	**0.849**	**0.075**	102
Set 11	0.359	0.320	126	**0.975**	**0.012**	94

troubles, unity is strength. The best performances are reached with supervised NCT, with the addition of a supervisor module providing dynamic feedback to the single subsystems about the global performances. Both RR and EER give better results in general, and most of all on the average, and the number of reliable responses significantly increases even in critical cases. The worst results are with set 4 and set 11, where about 75% of the responses were considered as reliable anyway. Notice the values obtained with screaming subjects of set 4. In this case we pass from RR = 0.588 and EER = 0.108 of PIFS alone (nevertheless, the best result among all classifiers) to RR = 0.962 and EER = 0.018 with supervised NCT, which is the higher proportional improvement obtained over all considered image sets. Again, the more critical are the identification conditions, the higher the advantage obtained by combining more classifiers. This latter results fully confirm our working hypotheses.

4 Conclusions

Biometric systems are generally somehow penalized from limits deriving from the adopted classification techniques. Multibiometric/multiclassifier systems resolve a number of problems of single-biometry ones, but suffer from the lack of communication among subsystems, and from the invariance of their parameters, making them unable to adapt to changes in the conditions of their working environment. In this paper, we propose an architecture aiming at overcoming such limitations, through the introduction of a collaboration protocol and of a supervisor module. Such additional component collects information from the different subsystems and exploits them to modify the internal conditions (parameters) of each subsystem, aiming at improving the global response. The experimental results show that the tighter is the cooperation among different classifiers, the better results are obtained. This work opens a line to further investigations, where aspects such as a deeper action of the supervisor module on the internal

subsystems state, or template updating, represent potential integrations to the architecture presented herein.

Acknowledgements

This work has been partially supported by the Provincia di Salerno Administration, Italy.

References

1. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
2. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Trans. Pattern Analysis and Machine Intelligence 18, 831–836 (1996)
3. Yan, S., He, X., Cai, D., Zhang, H.-J.: Neighborhood preserving embedding. In: IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1208–1213 (2005)
4. Han, J., Cai, D., He, X., Zhang, H.-J.: Orthogonal laplacianfaces for face recognition. IEEE Transactions on Image Processing 15, 3608–3614 (2006)
5. He, F., Kouzani, A.Z., Sammut, K.: Fractal face representation and recognition. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 1609–1613 (1997)
6. Riccio, D., Tortora, G., Abate, A.F., Nappi, M.: Rbs: A robust bimodal system for face recognition. International Journal of Software Engineering and Knowledge Engineering 17, 497–514 (2007)
7. Ross, A., Jain, A.K., Qian, J.-Z.: Information fusion in biometrics. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, pp. 354–359. Springer, Heidelberg (2001)
8. Riccio, D., De Marsico, M., Abate, A.F., Nappi, M.: Data normalization and fusion in multibiometric systems. In: International Conference on Distributed Multimedia Systems, DMS 2007, pp. 87–92 (2007)
9. Riccio, D., De Marsico, M., Abate, A.F., Nappi, M.: Face, ear and fingerprint: Designing multibiometric architectures. In: Proceedings of the 14th International Conference on Image Analysis and Processing - ICIAP 2007, pp. 437–442 (2007)
10. Poh, N., Bengio, S.: Improving fusion with margin-derived confidence in biometric authentication tasks. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 474–483. Springer, Heidelberg (2005)
11. Martinez, A.M., Benavente, R.: The ar face database - cvc technical report n.24. Technical Report (1998)
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 511–518 (2001)
13. Opencv. Website, 2008-06-06