NABS: Novel Approaches for Biometric Systems

Maria De Marsico, *Member, IEEE*, Michele Nappi, Daniel Riccio, and Genoveffa Tortora, *Senior Member, IEEE*

***Abstract*—Research on biometrics has noticeably increased. However, no single bodily or behavioral feature is able to satisfy acceptability, speed, and reliability constraints of authentication in real applications. The present trend is therefore toward multimodal systems. In this paper, we deal with some core issues related to the design of these systems and propose a novel modular framework, namely, novel approaches for biometric systems (NABS) that we have implemented to address them. NABS proposal encompasses two possible architectures based on the comparative speeds of the involved biometries. It also provides a novel solution for the data normalization problem, with the new quasi-linear sigmoid (QLS) normalization function. This function can overcome a number of common limitations, according to the presented experimental comparisons. A further contribution is the system response reliability (SRR) index to measure response confidence. Its theoretical definition allows to take into account the gallery composition at hand in assigning a system reliability measure on a single-response basis. The unified experimental setting aims at evaluating such aspects both separately and together, using face, ear, and fingerprint as test biometries. The results provide a positive feedback for the overall theoretical framework developed herein. Since NABS is designed to allow both a flexible choice of the adopted architecture, and a variable compositions and/or substitution of its optional modules, i.e., QLS and SRR, it can support different operational settings.**

***Index Terms*—Ear, face, fingerprints, multimodal systems, people identification.**

I. Introduction

IDENTITY checking through fast and effective automatic systems relates both to longer investigated applications, such as controlling limited access areas or locating dangerous individuals, and to the growing diffusion of electronic transactions. In biometric systems, recognition relies on physical, e.g., fingerprints or behavioral characteristics. The latter are gaining increasing attention due to their better reaction to spoofing, though suffering from low repeatability [37], [45]. At present, even the measurement of personality traits is being investigated [6]. Two operation modes exist verification (one-to-one matching to decide if the query subject actually corresponds to the claimed identity) and identification (one-to-many matching, to decide which is the identity corresponding to the query) [19].

Many present single-biometry systems are vulnerable to possible attacks and suffer from a number of problems [31]. Examples are acquisition errors, distortions (e.g., a voice altered by a cold), or the possible nonuniversality of the biometric feature (e.g., voice recognition with deaf-mute subjects). In a multimodal system, flaws of a single-biometric subsystem can be compensated by the others. Let us underline that, promising though it is, the biometric approach is mostly still limited to controlled and/or experimental settings. The "academic" scenario, where aware and consentient users are still in selected poses or limitedly move, contrasts with an "everyday" scenario with freely moving, unaware ore even not consentient users. Much research is still needed to exploit biometrics potentialities in the second scenario for massive everyday applications. Specific architectures can also be designed to better exploit the single biometries [1]. Along this line we present methodologies to combine more recognition modules into multimodal systems with different architectures. The presented architectures can be extended to a greater number and to different types of recognition modules. We introduce new tools, which are suited to any mixed combination of physical as well as behavioral traits, and demonstrate that they are able to improve system accuracy. We especially address identification, more computationally expensive than verification, though performed in usually less crucial situations. This calls for some optimization. The paper analyzes the features affecting novel approaches for biometric systems (NABS) framework design, typically characterize any biometric application setting:

1) the *biometry set*: to exemplify a possible instantiation of NABS framework, we combine face, ear, and fingerprint biometries. We try to meet both effectiveness and efficiency, since face and ear are efficient but not sufficiently effective, while fingerprints, in contrast, guarantee a high recognition rate (RR), yet in slow times;
2) the *normalization method*: each system may return results using different dimensionalities and scales; we propose a normalization function providing good results even when the maximum value to normalize is not known;
3) the *integration schema*: present systems follow three possible design choices, i.e., *parallel*, *serial*, or *hierarchic* [19]; one of the NABS architectures is a hierarchical schema, where face and ear modules work in parallel, while fingerprint module is connected in cascade; we further propose a new architectural schema called *N-cross testing protocol (NCTP)* and present the results obtained by a multimodal system implemented according to it;
4) a *reliability measure*: each subsystem in a multimodal architecture should return a reliability measure, to express how much its response can be trusted; in fact, not all subsystems might be equally reliable and/or single responses might deserve different confidence; this is important for fusing results; we propose two alternative measures, based on the composition of the stored gallery;

Manuscript received October 26, 2009; revised April 8, 2010; accepted July 3, 2010. Date of publication August 19, 2010; date of current version June 17, 2011. This paper was recommended by Associate Editor R. Alhajj.

M. D. Marsico is with the Dipartimento di Informatica, Università di Roma "La Sapienza," Roma 00198, Italy (e-mail: demarsico@di.uniroma1.it).

M. Nappi, D. Riccio, and G. Tortora are with the Dipartimento di Matematica e Informatica, Università di Salerno, Roma 84084, Italy (e-mail: mnappi@unisa.it; tortora@unisa.it; driccio@unisa.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org

Digital Object Identifier 10.1109/TSMCC.2010.2060326

5) the *fusion process*: the integration of information by different biometries is possible in three *moments*, i.e., during feature extraction, matching, or decision ([19], [35], [40]). The sooner the fusion is performed, the higher amount of information can be saved. Fusion in the matching module seems a profitable choice, since a "weighted" integration strategy can be exploited therein. As for the *fusion rule,* one could trivially decide to accept the identity with the highest returned score, or to rely on a linear combination of the single-subsystem responses, or even to apply a more complex schema. We will later describe the fusion rules in the proposed architectures.

As for the set of biometries, we just chose a group with different operational characteristics, to create a suitable test bed for our experiments. As for the remaining features, NABS provides novel contributions with respect to the state of art. Due to the framework modularity, all the adopted solutions are not binding, therefore increasing its flexibility and updatability over time. Moreover, the proposed techniques can also be used individually, due to the very loose assumptions about input and output formats of the implementing procedures. The presented original proposals address peculiar problems of multimodal systems: how to combine results from single modules, how to normalize data from different sources, and how to measure and exploit the different reliability of single subsystems. We compare some already existing solutions with our novel ones. Experimental results confirm the effectiveness of NABS approaches.

The paper proceeds as follows. Section II describes the proposed multimodal architectures. Section III and IV, respectively, introduce a new normalization function and two measures for system response reliability (SRR). Section V describes the set of biometries exploited to implement our unimodal subsystems and presents experimental results. Section VI discusses concluding remarks.

II. Proposed Architectures

One of the characterizing aspects of a multibiometric architecture is the stage at which fusion is implemented. We do not give here a complete survey about existing methods. However, it is worth introducing some observations to support the implemented design choices. Fusion at feature level is quite complex and also rigid, since it usually implies to exactly know in advance both the involved biometries and their peculiarities. Updating such kind of system is quite hard. A number of works in literature aim to compare match score level (or simply score level) fusion techniques with rank level ones. It can be deduced from them that it is not possible to decide which of these two approaches is the very optimal one: the choice is tightly bound to the working context where fusion is performed. Belkin *et al.* [4] reached the interesting conclusion that score level fusion is to be preferred to rank level one when consistent information is being fused, otherwise rank level fusion is to be preferred. We can assimilate gallery templates to documents as well and consider related results. In particular, Lee [23] observed that ranking is better if the runs in the combination have different rank-similarity curves. Renda and Straccia [34] underlined that, in rank level fusion, a reliable final result requires that each subsystem provides a rank for all candidate elements. Score-level fusion techniques can be broadly divided into classification or combination. Among the former, support vector machine (SVM) classification is quite accurate, but it is difficult to select the right kernel and its parameters. Statistical techniques such as likehood ratio (LR) and derived product of LR, and logistic regression require statistical tools, training, and a substantial amount of training data [41]. Their complexity is in the modeling of distributions, rather than fusion per se. False acceptance rate (FAR)-based techniques, e.g., product of FARs, require modeling the impostor distribution, but do not require genuine data. Scores can be normalized by transformation to FARs (using only the impostor distributions) before fusion. However, correctly modeling the impostor distribution usually remains a complex process. It is worth noticing that the effectiveness of product of likelihoods, logistic regression, and the FAR-based techniques greatly depends on how well the distributions are modeled. They give very good results, but assume a static gallery. Very dynamic settings require methods with a less complex setup [41]. Linear techniques involve addition of weighted scores. These methods do not usually require modeling of score distributions, but assume that the inputs have comparable scale, distribution, and strength, which only happens in very specific cases, such as fusion of left and right index fingers scored by the same matcher. On the other hand, Ross and Jain [35] experimentally compared score level fusion techniques based on combination or classification. They concluded that the former generally overcome the latter. We will now propose two architectures, combining different biometries based on their effectiveness and efficiency: the *NCTP* and the *hierarchical protocol (HP)*.

A. N-Cross Testing Protocol

N-cross testing protocol implements fusion through a novel kind of collaboration among subsystems. In this novel kind of architecture, N subsystems T_k, $k = 1, 2, \ldots, N$, work in parallel, first in *identification* mode and then in *lookup* mode, and exchange information in fixed points (see Fig. 1). Different data are acquired for the probe subject (e.g., face image, ear image, voice). The N subsystems start up independently and extract biometric features. Each T_k retrieves a list of candidates, where each list item includes the ID of a subject in the database and a score measuring its similarity with the input. The lists are ordered by increasing similarity. Each subsystem sends to the others a sublist with only the first M subjects, for a given M fixed in advance. Each T_k merges the $N - 1$ received lists in a single one. The length of a merged list will vary in the range $[M, M(N - 1)]$, depending on whether the same subjects are present in all the lists or only in some (or none) of them. For a correct fusion, scores from different subsystems are made consistent through the *quasi-linear sigmoid (QLS) function* defined later in Section III. This is a sigmoidal function, giving better results than popular normalization techniques (e.g., min–max) even when the upper extreme of values to normalize is unknown. Shared subjects get the mean of the original scores as their final

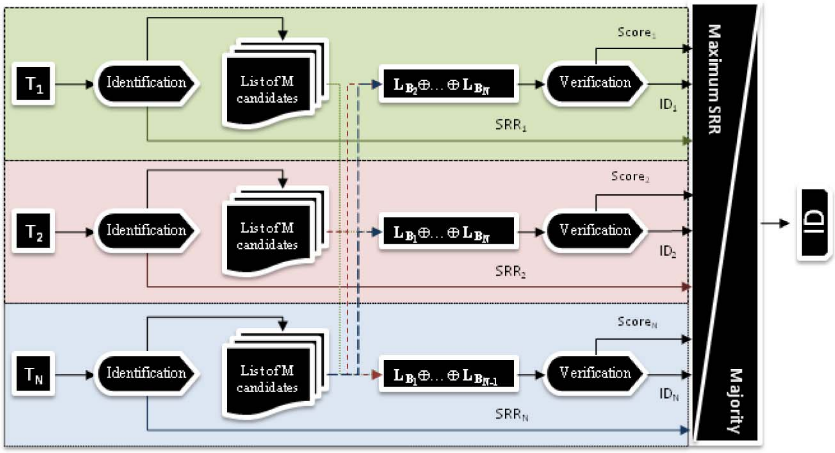

Fig. 1. Graphical representation of the NCTP.

one. Subjects in only one list retain the corresponding score. It is worth mentioning that score level fusion is adopted in NCTP because the exchanged lists of single subsystems are truncated to the first M subjects, and this makes rank level fusion less suited to our protocol, according to the observations in [34]. However, values must be consistent, as also discussed in [4], so that QLS is exploited. Each merged list is sorted again, and the N subsystems start phase II in parallel. Each T_k inspects its own original list as a lookup table, in the order given by the merged list, stops as soon as a subject distance falls below a minimum similarity threshold ε_k, and returns such subject and its computed similarity. In this way, the result returned by each T_k is influenced by the other subsystems. T_k might not return the very first subject in its list, i.e., its better choice, if this is too far behind in the merged list, causing subjects with a lower yet acceptable score to precede it. This is the main difference between NCTP and other types of voting protocols. The final response can be determined by a *majority criterion*: the identity receiving the majority of votes from the N subsystems is returned. If more subjects get the same number of votes, e.g., when the N subsystems each return a different identity, NCTP returns the one corresponding to the minimum distance $d_T = \min_k(d_T k)$, where k varies over subsystems voting for competing subjects.

The probe subject is recognized if d_T is smaller than a given threshold δ and rejected otherwise. The main limit of this approach is that many authorized users might be rejected. We verified that it is possible to improve global performances through a reliability measure for each subsystem (see Section IV). For the time being, let us assume that each subsystem also returns an estimate of its response reliability (see Fig. 1). If all identities returned in phase II are different, the protocol may decide to return the one associated with the highest reliability. In fact, most of the N acquisitions might be poor, but one is fully reliable, guaranteeing a correct recognition by itself. Also in this case, the probe subject is recognized if the distance from the returned identity is smaller than a threshold δ. The acceptance threshold δ depends on the biometric combination at hand, i.e., on the set of exploited subsystems. Therefore, it should be computed in advance according to the adopted configuration, as it happens in single mode. However, we found experimentally that a value $\delta = 0.4$ (over a range [0,1]) satisfactorily works for most tested multibiometric systems. More fusion rules involving reliability measures are discussed in Section IV. Experimental results in Section V-E refer to a *two-cross testing protocol* integrating only *contact-less* biometrics face and ear, which do not require users to interact with acquisition devices.

B. Hierarchical Protocol

Due to its parallelism, the NCTP achieves the execution time of the slowest subsystem. A hierarchical scheme would allow integrating effective but slow subsystems, when included, in a more efficient way. Fastest yet less accurate subsystems (e.g., face and ear) run first and independently produce their ordered lists. After normalizing similarity scores (see Section III), each subsystem breaks off its output list at the first K subjects, with K fixed in advance. Let us notice that K has the same role of M in NCT. We use a different letter just to underline that they need not be equal. Afterward, the lists are merged, and the resulting one is inputted to a more accurate yet slower module (e.g. fingerprint) that performs a 1:1 verification process for each input item. It stops as soon as the distance of a subject from the input is lower than a fixed threshold ε. The more this happens near to the beginning of the list, i.e., the more reliable are the other subsystems, the less is the overall time spent by the global system. Therefore, the main advantage of this architecture is that it reduces the operations required by the slowest subsystem, though preserving its RR. Experimental results in Section V-F refer to such implementation and also discuss the role of the K parameter. Different architectures, serial or parallel, can be derived from the presented schema, by inhibiting subsystems in turn.

III. Data Normalization and Fusion

Normalization is very important [38], especially in score level combination techniques. The output lists by single subsystems may contain numeric values, which result from measuring different features, using different procedures and different scales; their direct combination would give incorrect results because scores need to be comparable. For instance, in the NCT described earlier implements a step of data exchange among systems, which is absent in other methods. A single-ordered list is produced by each subsystem merging results from the others and represents the result of an intermediate identification step. Such list must be equivalent to a self-produced one. From this point of view, it is necessary that scores from equally reliable companions have the same incidence on the merged one, irrespective of the original scale. In other words, a list of returned scores within the [0,1] interval would otherwise influence values in the merged list much less than those from a system returning scores within the [0,100] interval. Large differences in the first list would become negligible in the combination with the second list. Nevertheless, it is to notice that some authors suggest to apply a normalization process even in conjunction with SVM or LR. For example [14] demonstrates than SVM performances are improved in this way. Different solutions to normalization have been proposed in the literature. In [18], a review of the most popular normalization techniques can be found, together with a discussion about their merits and limits. Among the others, the cited work compares min–max, z-score, a scheme using median and median absolute deviation (MAD), a scheme based on tanh-estimators, and a double sigmoid function. The limit of min–max technique is that it assumes that the minimum and maximum ever generated by a matching module are known. Moreover, it is influenced by outliers [14]. The z-score does not always guarantee a common interval for values from different subsystems. As compared with z-score, the quite robust median/MAD is more effective when values have a Gaussian distribution, otherwise median and MAD are poorly significant; moreover, the normalization technique neither preserves the original distribution nor transforms the values in a common interval. The schema based on hyperbolic tangent tanh guarantees projection in the open interval (0,1), but normalized data tend to concentrate around the center of that interval, and many parameters have to be determined.

In this section, we propose a normalization function derived from the family of *sigmoidal* functions, defined by

$$f(x) = \frac{1}{1 + ce^{-kx}}. \tag{1}$$

The codomain of such functions is the open interval (0,1). They have 0 as horizontal asymptote when $x \to -\infty$, and 1 when $x \to \infty$. However, they show an excessive distortion at the extremes (see Fig. 2, dotted line), and the shape depends on c and k, that in turn strongly depend on the domain of x.

Sanderson and Paliwal in [36] proposed an improvement, by considering mean and standard deviation of data. However, they assumed a Gaussian input distribution, which is often unrealistic and did not consider the distortion at the extremes. The double sigmoid in [18] improves the performances of sigmoid, but its parameters depend on the genuine and impostor score distributions, which may vary.

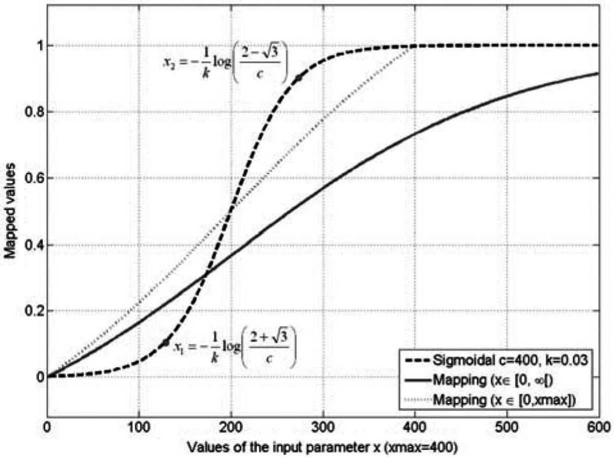

Fig. 2. Sigmoid function, the mapping function QLS when $0 \le x < \infty$, and a modified QLS when $0 \le x \le x$ max.

Data normalization in NABS follows a new approach, which first reduces the aforementioned distortion at the extremes. We derive a function $F(x)$ from the family $f(x)$, which we call QLS function. It shows a pseudolinear behavior in the whole codomain, though ensuring $F(x) \in [0, 1), \forall x$. We start by computing the extremes of the central region, where distortion is sufficiently small. Such extremes are the points x_1 and x_2, where the third derivative of $f(x)$ becomes null. Let us assume x-min as the minimum distance returned by the biometric system, while the maximum is x-max. We map these values onto x_1 and x_2, respectively, and set up a system of equations to be solved to find c and k. Here, the values in the domain of $f(x)$ are the distances between feature vectors, in the interval [0, ∞), and we can set x-min $= 0$ in the computed solutions. We obtain

$$c = 2 + \sqrt{3} \qquad \text{and} \qquad k = -\frac{1}{x\max}\log\left(\frac{2-\sqrt{3}}{2+\sqrt{3}}\right). \tag{2}$$

The influence of value x-max will be further discussed shortly and in Section V-C. The function we are looking for must have the interval [0, 1) as codomain apart from such value. Let us then define a new function $g(x) = f(x) - f(0)$ and compute its limit L when $x \to \infty$. Our QLS function is then

$$F(x) = \frac{1}{L}g(x) = \frac{1 - b^{\frac{x}{x\max}}}{ab^{\frac{x}{x\max}} + 1} \tag{3}$$

with $a = (2 + \sqrt{3})$ and $b = (7 - 4\sqrt{3})$. $F(x)$ guarantees a pseudolinear mapping for all values of x in [0, x max]; if we admit some distortion, it also allows to normalize values of x greater than x-max, still assuring $F(x) < 1$. Let us notice that this is essential when x-max is not known in advance and an estimate $\overline{x\max}$ is used instead. The graph of the function when $\overline{x\max} = 400$ is shown in Fig. 2 (solid line). When x-max is known, we can slightly modify $F(x)$ for further improvement:

$$\overline{F}(x) = \frac{ab+1}{1-b}F(x) \tag{4}$$

Fig. 2 confirms the better results when x-max $= 400$ (gray line). In Section V-C, we compare the proposed QLS function with the other discussed ones. As a further advantage, QLS is implicitly robust to outliers, due to its ability to rely on an estimate of the true maximum. No training is required, which is implicit in the modification of min–max in [14], using mean and standard deviation of genuine score distribution. A "light" training might only estimate $\overline{x\max}$ by comparing a limited number of significantly different subjects.

IV. System Response Reliability

The fusion of the results from different classifiers can be considered as one of the most significant advances in pattern classification in recent years [21]. On the other hand, subsystems might not be equally reliable, due to a possibly different accuracy of their procedures, and not all responses by a single subsystem might be equally reliable, due to a possibly different quality of input from time to time. An unreliable response, rather than producing a rejection, may represent a valid reason for a further check. A measure for response reliability is crucial for fusing the single results.

A possible approach relies on the quality of input data. However, it is hard to devise a metric estimating the confidence in an identification result, given the quality of a biometric acquisition. Such metric should also depend on the classification method. As an example, using the amount of occlusion to estimate input quality would be almost needless for a recognition method robust to occlusion. A further fault of an absolute metric is that it would not consider the relation between different data used for enrolment and for testing. If subjects are enrolled and verified in nonoptimal yet constant conditions (e.g., face images acquired by a constant light from one fixed side), we may have a high probability of a sufficient reliability. Input quality is used by Kryszczuk *et al.* [22]. Their Bayesian networks evaluate the probability of a correct verification decision by a classifier, given the available evidence from several sources. A vector of signal-domain quality measures accounts for significant features of each biometry and of each classification method.

An alternative approach would use classical performance measures, e.g., RR. However, they provide a global estimate on the recognition ability and not of the reliability of a single testing result. The single value of the distance/similarity between the probe and the returned subject is even less significant: it neglects the relation between the latter and the rest of the gallery. Results should be differently evaluated also based on the similarity of gallery subjects, i.e., on the discriminating power of the biometric feature.

Some solutions in literature use confidence measures based on margins. Each margin measures the amount of "risk" associated to a single-subsystem response, after observing its scores. Margins can be applied to any subsystem, despite the nature of its input. Poh and Bengio [32] introduce a confidence margin based on FAR and false rejection rate (FRR) of a biometric system, showing how it behaves better than those by Freund and Schapire [12] and by Vapnik [42], which are exploited for boosting and within statistical learning theory, respectively. The latter can only be calculated by supposing that the target output (class-label, i.e., impostor/genuine in this case) is known, so that they are only significant in training. On the other hand, since Poh and Bengio's margin relies on an estimate (FAR, FRR) of the actual distribution of genuine/impostor subjects' scores, a quite high number of responses are marked as reliable. The error rate of the global system is reduced, but this might not be appropriate for applications requiring a very high security level. Kryszczuk *et al.* [22] also exploit a margin similar to Poh and Bengio's one, which is defined in terms of correct acceptance and correct rejection accuracies at a given acceptance threshold. However, a frequentist approach to reliability is considered valid only assuming that the scores of the testing and of the development sets have a similar distribution, which is also stable in time.

The NABS addresses the problem of system reliability in a novel way. We propose a new kind of margin by defining a *system/gallery-dependent* metric that we will call SRR. Gallery is an integral component of an identification system, and its composition may influence recognition performances. However, present methodologies do not support tuning to the database. Our metric also attempts to limit such drawback and can be applied to any kind of biometric system. It estimates the ability of separating genuine subjects from impostors, for each probe, in the sharpest possible way. Section V shows the better results in fusing single responses than those ensured by Poh and Bengio's margin, since SRR considers the actual current separation among genuine and impostor subjects rather than static parameters relying on variable score distributions.

Let A be an identification system and G its gallery of correctly enrolled identities. Assume there are $l > 0$ acquisitions (templates) for each genuine identity. Let p be a person to be identified (probe). The system first computes the distances $d\,(p, g_i)$, $i = 1, \ldots, |G|$, between p and each template in the gallery. They are ordered so that $d(p, g_{i_1}) \leq d(p, g_{i_2}) \leq \ldots \leq d(p, g_{i_{|G|}})$. Two functions $\varphi_j(p)$, $j = 1, 2$, were used to compute two alternative SRR measures. SRR I uses the relative distance (relative distance between the scores of the first two retrieved distinct identities); SRR II uses the density ratio (relative amount of gallery templates, which are "near" to the retrieved identity). Each $\varphi_j(p)$ is independent from the specific measure d, since the QLS function F of Section III is used to normalize distances in the interval $[0, 1)$, x max is equal to the maximum distance between the present probe and the gallery images.

Relative distance was used by Yan and Bowyer in [46]. We achieved better results by the combination with QLS:

$$\varphi_1(p) = \frac{F(d(p, g_{i_2})) - F(d(p, g_{i_1}))}{F(d(p, g_{i_{|G|}})} \tag{5}$$

where g_{i_2} is the second distinct identity in the returned ordered list. Experiments showed that the relative distance tends to be small for impostors and high for genuine subjects, independently from the biometry and from the classification method. It relates to the degree of uncertainty by which the system identified person p: if $\varphi_1(p)$ is high, a person exists in the gallery, which is much more similar to p than all the others; otherwise, the retrieved g_{i_1} is only the less far in a set of identities with a similar distance from p.

The second function $\varphi_2(p)$ (density ratio) is computed using the ratio between the number of subjects in the gallery, distinct from the returned identity, giving a distance lower than twice

$F(d(p, g_{i_1}))$, and the cardinality G of the gallery.

$$\varphi_2(p) = 1 - \frac{|N_b|}{|G|} \tag{6}$$

where

$$N_b = \{g_{i_k} \in G | F(d(p, g_{i_k})) < 2F(d(p, g_{i_1}))\}.$$

Even in this formulation, the higher is the algorithm ability to discriminate between a genuine subject and an impostor one, the lower is the probability to find identities different from the correct one at a small distance from it. Notice that in both cases the values for $\varphi_j(p)$ fall within the interval [0, 1]. $\varphi_1(p)$ is quite easier to compute, but it is more sensible to outliers. These occur when either the first two retrieved subjects are occasionally very similar, even if quite different from the rest of the gallery, or the last subject is occasionally very far (high distance) from the rest of the gallery. $\varphi_2(p)$ is a little bit more expensive, but accounts for a significant local neighborhood of the retrieved subject. Therefore, both the preceding situations, in particular far outliers, do not affect the result.

After defining $\varphi_j(p)$, we need to identify a value $\overline{\varphi_j}$ fostering a correct separation between genuine subjects and impostor ones. Each value ($\overline{\varphi_1}$ or $\overline{\varphi_2}$) marks the point of maximum uncertainty and varies with the biometry and with the classifier, so that it must be estimated from time to time during the setting up of the single subsystems. The optimal $\overline{\varphi_j}$ is given by that value able to minimize the wrong estimates of function $\varphi_j(p)$, which can occur in two scenarios. In one case, an impostor is erroneously recognized with $\varphi_j(p)$ higher than $\overline{\varphi_j}$, so that the false acceptance is erroneously supported by a high value of $\varphi_j(p)$. Otherwise, a genuine subject is recognized with $\varphi_j(p)$ lower than $\overline{\varphi_j}$, so that the acceptance may be possibly questioned due to a low value of $\varphi_j(p)$. We also define $S(\varphi_j(p), \overline{\varphi_j})$ as the width of the subinterval from $\overline{\varphi_j}$ to the proper extreme of the overall $[0, 1)$,interval of possible values, depending on the comparison between the current $\varphi_j(p)$ and $\overline{\varphi_j}$.

$$S(\varphi_j(p), \overline{\varphi_j}) = \begin{cases} 1 - \overline{\varphi_j}, & \text{if } \varphi_j(p) > \overline{\varphi_j} \\ \overline{\varphi_j}, & \text{otherwise.} \end{cases} \tag{7}$$

The SRR index (SRR I or SRR II) can finally be defined as follows:

$$\mathrm{SRR}_j = |\varphi_j(p) - \overline{\varphi_j}| / S(\overline{\varphi_j}). \tag{8}$$

We first measure the absolute distance between $\varphi_j(p)$ and the "critical" point. Such distance gets higher values for $\varphi_j(p)$ much higher/lower than $\overline{\varphi_j}$ (genuine/impostor respectively). However, it is also important to consider its significance compared with the subinterval over which it is measured. This allows to compare reliability of different responses and of responses from different systems and to finally estimate a threshold th_j for an acceptable reliability. As an example, assume $\overline{\varphi_j} = 0.1$; a reject with $\varphi_j(p) = 0.02$ (absolute distance 0.08 and $\mathrm{SRR}_j = 0.8$) is to be considered more reliable than a reject with $\varphi_j(p) = 0.08$ (absolute distance 0.02 and $\mathrm{SRR}_j = 0.2$), and this can be reflected by the absolute distances alone. However, such response must also be considered more reliable than an accept with $\varphi_j(p) = 0.19$ (absolute distance 0.09 and $\mathrm{SRR}_j = 0.1$), which in fact spans a proportionally minor distance from the critical point toward the other end of the overall interval. It is worth noticing that SRR works without any knowledge about acquisition quality, extracted features, and classification methods. It can be used with all those (off-the-shelf) identification modules, which return an ordered list of distances (similarities).

In a multimodal system, reliability thresholds of single subsystem can be estimated in advance and remain fixed in time or can be computed and updated according to returned responses. A compromise between the number of reliable responses (NRRs) and the system error rate must be obtained. Too high thresholds make the system too restrictive, with a low error rate but also a low number of acceptances, while too low ones risk canceling the advantages of a reliability measure. Assume that T_k subsystem has executed M times producing $\{T_k^1(1), \ldots, T_k^M(1)\}$ responses with the corresponding reliability measures, which are collected in a set $\mathrm{BH}_k = \{\mathrm{SRR}_k^1, \ldots, \mathrm{SRR}_k^M\}$, representing the *history* of the system behavior. The value to assign to th_k is correlated to BH_k, in particular to its mean and its variance. A high $\overline{\mathrm{BH}_k}$ means generally reliable responses so that the corresponding threshold can be proportionally high. The variance $\sigma[\mathrm{BH}_k]$ measures the stability of T_k. The best situation is when BH_k has a high mean and a low variance, so that it is possible to fix a high value for th_k. We can summarize this in the formula

$$\mathrm{th}_k = \left| \frac{\overline{\mathrm{BH}_k}^2 - \sigma[\mathrm{BH}_k]}{\overline{\mathrm{BH}_k}} \right|. \tag{9}$$

The aforementioned formula implies some preliminary training; however, it can be periodically exploited, with no further interruption of system normal activity, to change the system reliability threshold, in order to better accommodate a dynamic setting. This is different from what normally happens with other quality or reliability measures, which require system retraining if context dramatically changes. In the following, we discuss how to integrate an SRR index into the fusion protocol. In our approach, each biometric module in a multimodal system produces such measure for each response. The integration policy of the fusion module has then to weight the single responses based on the respective SRR values. Different choices exist for integration, all equally sound, yet possibly leading to very different results. A very simple fusion rule, based on a majority criterion, was presented at the end of Section II-A. The hierarchy in some of our systems also encodes a certain specific fusion rule. Let us analyze further hypotheses, considering their features and postponing to Section V their evaluation. We start from briefly discussing two works that conceptually support our choices. The first one is about mixed group ranking (MGR) [28]. MGR exploits the power set of the set of classifiers; each subset A in it is assigned a predefined weight (preference); for each class θ, the procedure computes the weighted sum of its best ranks returned by each subset A. MGR implicitly guarantees confidence in lower ranks (the score function is more affected by variations in a lower rank). The authors show that this is in general a desirable feature. However, such technique needs a training phase for weight estimation and to readapt the system in a dynamic setting. On the other hand, the earlier findings support the suitability of our approach. Our technique not only relies on confidence in lower ranks (SRR is computed from the first and second retrieved identities), but also explicitly measures it, allowing the system to adjust the weights without any training phase.

The second work is a study by Veeramachaneni *et al.* [43] that addresses the choice of the best fusion rule for the final decision process in a binary hypotheses-testing problem. The number of explored rules grows exponentially with the number of subsystems, and, if performances of one subsystem change, the best rule must be searched again. However, two interesting observations stem from the study. AND and OR rules are a very important set (out of 16 possible ones) for fusing the results of two classifiers. Second, as the number of classifiers increases, the optimal fusion rules are constructed from AND and OR [29]. This suggests that, despite the variety of fusion policies, we can assume that AND and OR are generally a suitable choice. We will then rely on them.

Assume to have a system S composed by N subsystems $T_1, \ldots, T_N$. Each T_k uses a gallery G_k, $k = 1, 2, \ldots N$, of acquisitions for the same set of identities. However, sizes of the galleries might be different, as each biometry may require a different number of acquisitions per identity. Each subsystem can order its gallery according to the distance from the probe image p and produce a numeric value SRR_k for its response reliability. A consistent data fusion requires normalization to obtain $\sum_k \mathrm{SRR}_k = 1$. The thresholds th_k discussed earlier are normalized too and used to classify the responses as reliable/unreliable. They are also used for integration.

Our OR and AND integration policies follow a twofold schema, to better integrate reliability estimate within an identification problem. To determine global system reliability, a decision fusion under binary hypotheses (classes are reliable/not reliable) is performed first, using the OR or the AND operator (at least one or all the subsystems are reliable). Afterward, a different score fusion is used to determine the returned identity, depending on the previously used operator (the first identity of the most reliable subsystem, or the identity with the minimum weighted sum of distances from the probe, where weights are the subsystem reliabilities). SRR can also be used within more articulated schemes, such as the earlier NCTP or fuzzy fusion schemes.

V. Experimental Results

A. Set of Biometries

The presented multibiometric systems exploit face, ear, and fingerprint subsystems. The single-mode modules, making up the core building blocks of NABS architectures, are implemented using known segmentation and classification techniques. We use open source software to detect specific interest regions in face and ear, which is a research issue in itself. Location of regions is performed using an *object detector* based on Haar features [44], implemented in the *OpenCV* library [52], [53] and exploiting, in particular, Haar Cascades provided in [47]. We think that using effective, yet "second-hand" material for the basic operation of single subsystems (modules) does not diminish the obtained results, since this paper rather focuses on the effective and suitable combination of responses. Algorithms for both face and ear recognition rely on fractal theory, in particular on partitioned iterated function system (PIFS) [8], [10], which have been adapted here to the problems at hand [2], [3], [7]. In particular, the division of the face and ear images in regions improves robustness to local variations and lighting or to occlusions. In order to validate our choice, we compared PIFS recognition performances with different widely exploited techniques: Neighborhood preserving embedding [16], locality preserving projection (LPP) [13], weighted fractal neighbor distance [39]. Further candidates included kernel-based techniques (see the recent work by Li *et al.* [24], [25]). However, our goal was not to identify the single optimal classifier, but to demonstrate that a multibiometric approach can improve the performances of a quite good classifier. We applied the chosen methods to a gallery and a probe of 100 subjects from the AR-Faces database [26]: set 1(normal) was the gallery, while sets 2 (smile), and 4 (angry) were the probes. The comparison relies on cumulative match score (CMS) and cumulative match characteristic (CMC) curve. Tables I and II give a tabular representation of the CMS values for some ranks with the different techniques. Table I shows that PIFS generally provided higher face recognition performances. Only LPP sometimes obtained comparable results on set 4.

TABLE I
Performances of Different Face Identification Methods on Sets 2 and 4 From the AR-faces Database

Method		Rank				
		1	2	3	4	5
PIFS	Set 2	**0.96**	**0.99**	**0.99**	**0.99**	**0.99**
	Set 4	**0.60**	**0.71**	**0.76**	**0.82**	**0.84**
NPE	Set 2	0.66	0.72	0.75	0.77	0.78
	Set 4	0.40	0.58	0.61	0.65	0.67
LPP	Set 2	0.54	0.64	0.66	0.69	0.72
	Set 4	0.58	**0.71**	**0.76**	0.79	0.81
WFND	Set 2	0.93	0.97	0.98	0.98	0.98
	Set 4	0.48	0.61	0.70	0.71	0.80

TABLE II
CMS Performances of the Ear Identification When Occlusions Occur: PIFS Versus PCA on 100 Subjects From the FERET Database (Profile Faces)

Method	Rank				
	1	5	10	15	20
PIFS	89%	96%	98%	100%	100%
PCA	72%	90%	95%	98%	99%

We also compared our PIFS algorithm for ear recognition to principal component analysis (PCA) [20]. A subset of 100 subjects in profile from the facial recognition technology (FERET) database [30] was used: the gallery included images in neutral conditions, while images with little occlusions (hair or earrings) were used as probe. Data in Table II confirm the superiority of PIFS even in this case. Most present fingerprint recognition systems rely on minutiae matching. In general, minutiae-based algorithms suffer from image distortion. Different approaches for fingerprint matching include the one by Bioscrypt, which has been exploited here [48]. It uses the whole fingerprint, rather than selecting a limited number of minutiae points. Pattern recognition is therefore more robust to lack of information, e.g., a damaged or dirty finger. This methodology has proven very effective during FVC 2002 (International Fingerprint Verification Competition) [50]. In fact, the Bioscrypt algorithm provided the lowest equal error rate (EER).

B. Databases, Measures, and Evaluation Criteria

We first describe the benchmark data and the criteria to highlight the obtained performance improvements. We created a multimodal database with each subject having corresponding images for face, ear, and fingerprints. We alternately used FERET [30] and AR-Faces [26] databases for the face, Notre-Dame [11], [51] for the ear and a proprietary database for fingerprints. Since Notre-Dame contains about 100 subjects, subsets were also selected from FERET and AR-Faces containing 100 subjects each, i.e., the first 100 labeled ones. The choice of a well-identifiable subset will facilitate future comparisons. In detail, three subsets of 100 subjects each came from FERET, respectively, from FAFB (changes in facial expression from gallery to probe sets), FAFC (changes in illumination), and DUP I (elapsed time between one minute and 1031 days) sets. As for AR-Faces, selection was from sets left light, sad, scarf, scream, sun glasses, using neutral as gallery. Different data were alternatively used during experiments. Fingerprints of 100 subjects were acquired using a DSP Starter Kit by Texas Instruments. TMS320C6713 DSK is based on DSP C6713 processor; it is also provided with a fingerprint authentication development tools (FADTs) module, connected to the device as an expansion board. FADT module is equipped with an AFS8600 sensor by Authentec, for fingerprints acquisition, and with a demo software for algorithms evaluation [49]. The assignment of different biometric data to the same virtual user was consistent throughout the experiment.

It is worth making a digression about the use of chimeric (virtual) users, which is often found in literature. There is an open debate about the equivalence between chimeric and true users in authentication experiments (e.g., [33] and [13]). Some results, briefly presented here, encouraged us in our choice. Poh and Bengio [33] use the XM2VTS multimodal database [27], to build 1000 samples of random identity match (chimeric). In brief, they check if half total error rate (HTER), i.e., the mean of FAR and FRR, of true identity match falls within the HTER range of the 1000 samples of random identity match. Only about two thirds of the data indicate suitability of chimeric users. However, in the other cases, computed differences are very small. More objections come from Dorizzi *et al.* [9]: a bias in the experiments could also derive from particular distributions of the chosen training and test sets. Their results suggest that experimental evaluations, using EER as a performance measure, would not be significantly affected by the use of chimeric users. This is the main reason for exploiting EER in this paper.

System performances were analyzed using RR and EER. Besides the earlier considerations about EER, these two measures provide a sufficiently complete system analysis and well represent two categories. RR mainly evaluates the ability to retrieve the correct subject of 1:m identification systems. On the other hand, measures such as EER (and FAR/FRR from which it is derived) evaluate the ability to recognize the right subject, but also to refuse impostors with a high certainty degree, so proving more suited for 1:1 verification systems. Our use of EER for identification might then be criticized in favor of more suited measures such as CMC curve. However, [5] demonstrates that there is a precise relation between CMC and FAR/FRR, given that a 1:1 matcher is used for sorting identification scores. In this case, CMC can be computed from FAR and FRR, so that the latter can be used to evaluate an identification system as well. In our experiments, we estimated them from the matrix of distances between each probe and each gallery subject; false acceptances and false rejections were detected according to a fixed a threshold for a positive match. The thresholds used for each subsystem are the same that will be mentioned shortly for the rest of the experiments. We will first report the separate effects introduced by the QLS function, the SRR measures, and the fusion approaches, and then account for the performances of the proposed architectures.

TABLE III
COMPARISON AMONG THE PERFORMANCES OF THE BIOMETRIC SYSTEMS FOR DIFFERENT NORMALIZATION FUNCTIONS

SYSTEM		PERFORMANCES						
		MIN MAX	Z SCORE	MEDIAN MAD	TANH	SIGMOID	DOUBLE SIGMOID	**QLS**
FACE	RR	93%	93%	93%	93%	93%	93%	**93%**
	EER	0.03	0.23	0.12	0.06	0.04	0.04	**0.03**
EAR	RR	72%	72%	72%	72%	72%	72%	**72%**
	EER	0.14	0.25	0.17	0.16	0.16	0.15	**0.14**
HP	RR	100%	94%	97%	95%	100%	100%	**100%**
	EER	0.003	0.22	0.05	0.02	0.01	0.01	**0.003**

C. How Much the QLS Function Affects Performances?

We compared the use of the functions in Section III in setting up the hierarchical multibiometric system presented in Section II-B (HP in the following tables). In Table III numerical performances, when minimum and maximum values are correctly estimated. The results show that performances of the multibiometric system are more widely affected by score normalization than single subsystems. Face/ear RR remains unchanged across different methods, while the ERR undergoes significant variations, probably due to an increment of the FAR. The worst performance is given by the z-scores, with EER always above 0.22. Table III highlights that, when minimum and maximum are correctly estimated, our QLS function is able to obtain results comparable to the optimal ones of min–max function.

In addition, QLS provided even better results in further tests when the maximum value is wrongly estimated. This definitely justifies its introduction as an alternative to min–max. Let us further detail this point. The facial recognition subsystem guarantees, on FAFB, FAFC, and DUP I databases, an RR of, respectively, 93%, 16%, and 47%, and an EER of 0.03, 0.29,

TABLE IV
COMPARISON BETWEEN THE MAPPING FUNCTION AND MIN–MAX, FOR WRONG ESTIMATION OF THE MAXIMUM SCORE

SYSTEM		MAXIMUM SCORE OVERESTIMATED		MAXIMUM SCORE UNDERESTIMATED	
		MIN-MAX	QLS	MIN-MAX	QLS
FACE	RR	93%	93%	38%	93%
	EER	0.04	0.04	**0.81**	**0.034**
EAR	RR	72%	72%	72%	72%
	EER	0.14	0.14	0.14	0.14
HP	RR	78%	78%	81%	97%
	EER	0.08	0.08	**0.10**	**0.058**

0.19; the ear module provides an RR of 72% and an EER of 0.15 on Notre-Dame database. Assume a multimodal system composed by the simple sum of face and ear. Assume also that any score out of the [0, 1] interval, i.e., higher that one, is set to one. In two distinct experiments, ear scores were normalized using their correct maximum. In the first experiment, face scores were normalized by overestimating their maximum (ten times larger); in the second experiment it was underestimated (five times smaller).Table IV shows the results obtained for min–max and QLS. When maximum is overestimated, face scores tend to vanish, so EER of the final system is very close to that of ear subsystem. However, significant differences can be noticed when the maximum is underestimated. Scores given by min–max are mostly out of the [0,1] interval, so that, according to the aforementioned assumption, they are leveled toward the 1 extreme. On the contrary, scores produced by QLS function all lie in the [0,1] interval and maintain a precise distribution, so giving higher values for both RR and EER.

D. Experimental Measure of the SRR Index Contribution

We tested the ability of the SRR index to discriminate between reliable and unreliable system responses for both face and ear. Notice that an SRR index for our fingerprint module would be redundant since such biometry is highly reliable in itself. In the first place, we had to determine the optimal value $\overline{\varphi_j}$, $j = 1{,}2$, for computing SRR$_j$. To this aim, we made it vary within the interval [0,1], each time computing the percentage of responses for which the relation between the value of $\varphi_j(p)$ and $\overline{\varphi_j}$ was consistent with the correct result: correctly identified genuine subjects with $\varphi_j(p)$ higher than $\overline{\varphi_j}$, and impostors or unsurely identified subjects with $\varphi_j(p)$ lower than $\overline{\varphi_j}$. The chosen value maximizes the aforementioned percentage, somehow representing the reliability of the SRR$_j$ index. Let us remind that, according to (17), SRR$_j$ will get higher values either for $\varphi_j(p)$ much higher than $\overline{\varphi_j}$ (genuine subjects) or for $\varphi_j(p)$ much lower than $\overline{\varphi_j}$ (impostors), also depending on the position of $\overline{\varphi_j}$ within the [0,1] interval.

Once a system returns an SRR value, we have to decide if the response is reliable enough to be accepted. The next tests studied the relation between system performances and variations of the threshold th$_k$. For each subset (probe/gallery) from FERET, and from Notre-Dame, the RR, EER, and the NRRs were computed while varying th$_k$. Results are in Fig. 3. As expected, for all experimented datasets, system performances change with the SRR threshold. In all cases, the higher the threshold, the lower the NRR value. However, RR tends to improve. Fig. 3 also shows that a threshold (th) computed according to (9) (red line), though not optimal, guarantees a good compromise between NRR and RR. Of course, such choice can be further refined according to the security requirements of the system at hand.

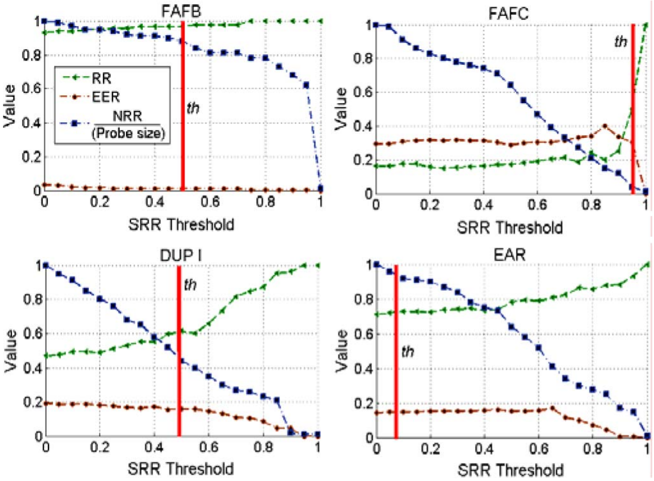

Fig. 3. RR, EER, and NRR variations with respect to the threshold (th) when SRR II is used.

TABLE V
COMPARISON BETWEEN THE PERFORMANCES OF SOME FUSION RULES

DATABASE		STATISTICS							
		NONE	FUZZY	SRR I		SRR II		POH'S MARGIN	
		SIMPLE SUM	WEIGHTED SUM	OR	AND	OR	AND	OR	AND
FAFB	RR	98%	98%	99%	100%	96%	100%	76%	99%
	EER	0.028	0.020	0.01	0.003	0.01	0.00	0.19	0.04
	NRR	100	100	75	63	94	38	99	86
FAFC	RR	55%	72%	76%	100%	84%	NA	68%	60%
	EER	0.167	0.14	0.15	0.002	0.11	NA	0.18	0.22
	NRR	100	100	85	2	74	0	99	96
DUP I	RR	75%	78%	81%	100%	87%	100%	67%	84%
	EER	0.238	0.22	0.22	0.001	0.17	0.000	0.22	0.12
	NRR	100	100	91	18	84	22	98	96

In order to evaluate the contribution of a reliability index within a multimodal system, we compared performances with those obtained by introducing it. We tested SRR I and SRR II and the related OR and AND integration policies. This experiment was performed for the datasets from FERET, combined with Notre-Dame ears. We remind that the facial recognition provides, on FAFB, FAFC, and DUP I respectively, an RR of 93%, 16%, and 47%, and an EER of 0.03, 0.29, 0.19; the ear recognition on Notre-Dame provides an RR of 72% and an EER of 0.15. Table V reports results, in terms of RR, EER, and NRRs, for the multimodal system face/ear using each of the integration policies; performances of Poh and Bengio's margin are also reported. Moreover, fuzzy decision described in [16] was tested. In the original work, external quality indexes were used as values over which to apply fuzzy rules. Indexes were related to the lighting level and to the amount of pose distortion, which might be difficult to compute. We used SRR I and SRR II instead. For reader's convenience, we once more remind that the facial

TABLE VI
COMPARISON BETWEEN THE PERFORMANCES OF SRR I AND SRR II

DISTORTIONS ON THE FACE		STATISTICS					
		FACE	EAR	FACE ⊕ EAR			
					SRR I	SRR II	POH'S MARGIN
Left Light	RR	93%	72%	RR	100%	100%	100%
	EER	0.09	0.12	EER	0.001	0.008	0.019
				NRR	37	70	83
Sad	RR	100%	72%	RR	100%	100%	100%
	EER	0.07	0.12	EER	0.005	0.002	0.030
				NRR	86	43	97
Scarf	RR	80%	72%	RR	100%	100%	97%
	EER	0.17	0.12	EER	0.015	0.020	0.080
				NRR	70	70	100
Scream	RR	47%	72%	RR	100%	100%	100%
	EER	0.18	0.12	EER	0.001	0.020	0.050
				NRR	23	46	83
Sun glasses	RR	90%	72%	RR	100%	100%	95%
	EER	0.14	0.12	EER	0.016	0.010	0.020

TABLE VII
PERFORMANCES OF THE CROSS TESTING PROTOCOL

FACE DATASET	STATISTICS						
	FACE	EAR	FACE ⊕ EAR				
			SIMPLE	SRR I		SRR II	
FERET	**RR**	**RR**	**RR**	**RR**	**NRR**	**RR**	**NRR**
FAFB	93%	72%	94%	96%	97	96%	97
FAFC	16%	72%	70%	81%	80	76%	80
DUP I	47%	72%	82%	90%	84	87%	86
			AR-FACES				
Left Light	83%	72%	97%	97%	100	97%	100
Sad	95%	72%	100%	100%	100	100%	100
Scarf	80%	72%	97%	97%	100	97%	100
Scream	47%	72%	86%	82%	97	82%	97
Sunglasses	87%	72%	97%	100%	100	100%	100

recognition subsystem guarantees, on FAFB, FAFC, and DUP I databases, an RR of, respectively, 93%, 16%, and 47%, and an EER of 0.03, 0.29, 0.19; the ear module provides an RR of 72% and an EER of 0.15 on Notre-Dame database. Let us observe that, on the whole, all the integration policies guarantee a higher RR than face or ear alone (see 87% for DUP I with OR/SRR II instead of 47% and 72%), and at the same time they considerably reduce EER (see 0.015 for FAFB with OR/SRR II instead of 0.030 and 0.150). The counterbalance with reliability estimation is the number of unreliable responses. In the first two columns all responses are considered as reliable, since fusion is performed without filtering responses. We can notice that fuzzy fusion outperforms simple sum in degraded conditions (in a significant way for FAFC and less dramatically for DUP I), but it is quite equivalent in good conditions like FAFB, were there is very little difference only for EER. However, both RR and EER in these columns are significantly worse, showing the ability of SSR to discard little reliable responses.

When the SRR index is used, OR of course ensures a higher number of responses considered as reliable; the minimum is reached by the AND policy (see e.g., results on FAFC, where no reliable response is obtained). As already underlined, an unreliable response might just represent a valid reason to perform a further, possibly more expensive check, e.g., by using fingerprints. Such check would only involve a limited number of cases, so that the average computational cost of the multimodal system would be reduced anyway, with the advantage of a higher security. We can further deduce that, if the input quality is sufficient (see the FAFB subset), the combined system will reach much higher performances than the single subsystems, most of all in terms of EER. As for Poh and Bengio's margin, in general the NRRs are always very high, with an appreciable worsening of RR and EER. Finally, the two reliability indexes SRR I and SRR II as well as Poh and Bengio's margin were tested on subsets extracted from AR-Faces, combining face and ear with the AND fusion policy. The database used for ear is always the same. Results are reported in Table VI, where the first two columns report RR and EER of the single subsystems. Results in Table VI further enforce some considerations. The combined system always guarantees better performances (e.g., for distortion of the face equal to *scream* we pass from 47% and 72% to 100% for RR). However, when the single subsystems are particularly stressed by input quality, the NRRs significantly decrease, confirming the need for more biometries. Again, EER with either SRR I or SRR II is better than that with Poh and Bengio's margin.

E. N-Cross Testing Protocol

We compared RR of the two-cross testing protocol with that of face and ear alone. We considered two cases. In the former, denoted as SIMPLE, the subject is recognized if both subsystems return the same identity or, in case of two different identities, if one subsystem returns an identity with a distance from the probe lower than the other and under a given threshold. In the second case, the used criterion is maximum reliability. Results in Table VII show that the combined system in the SIMPLE case is more restrictive, so providing, on the average, a lower RR. On the contrary, RR grows when a reliability index is used. We can observe that the *two-cross testing protocol* generally outperforms both face and ear; when one subsystem behaves significantly worse than the other, the combined system tends to retain the better performances (RR of 70%, 81%, or 76% of the Face⊗Ear versus 16% and 72% of the face and ear alone on FERET FAFC); *SRR I* and *SRR II* give very similar improvements.

F. Hierarchical Schema

In the experimental implementation of the hierarchical system presented in this paper, face and ear work in parallel. Each subsystem orders its own output list, according to the distance from the probe, and then breaks it off at the first K subjects; K is fixed in advance, and in our case $K = 10$. The two lists are fused in a single-ordered one (see Section II-A and B), which is inputted to the fingerprint subsystem. This module verifies each subject in the received list and the one with the minimum distance is returned as the probe identity. The measured RR of the whole system was 99%: a subject was lost because neither face nor ear returned it among the first ten. The most interesting result from the tests was the reduction of the identification time. Face and ear require 4 s to extract a candidate list, while fingerprints spends 0.24 s to verify a subject. It is then possible to estimate the speed up introduced by the hierarchical system, yet maintaining RR comparable to the fingerprint subsystem alone. Identifying a subject over a gallery of 100 persons would require about 26 s, while the hierarchical system requires 8.8 s to obtain the same recognition, since the fingerprint module verifies at most only the first 20 subjects returned by the identification phase. The time spent by our hierarchical system can be further reduced if we consider that the fingerprint recognition algorithm, which is used in our case presents a very good ability in discriminating between the correct identity and the remaining gallery subjects. Only the gallery subject whose identity actually corresponds to the probe reaches a score higher than 0.50 (the score varies from 0 to 1); this suggests to maintain the whole lists returned by face and ear, in order to reduce false negatives, while relying on the aforementioned algorithm feature to reduce the number of subjects actually verified by the Fingerprint module. The latter verifies the subjects starting from the head of the received merged list. It halts as soon as a subject with a score higher than 0.5 is found and returns its identity. Using such policy, in some cases more than 20 subjects are verified; this is the upper limit resulting from the previously adopted policy. On the other hand, in most cases, if face and ear worked appropriately, the Fingerprint subsystem will halt much before. However, this policy presents a limit when a probe subject is not present in the gallery, because the fingerprint verification module would go through all the subjects in its input list. If the single face an ear lists were limited (to a length K), we would still obtain reasonable times, though without a final recognition. Otherwise, this would cause an unacceptable waste of time. To solve this drawback, it is possible to force a maximum recognition time, after which a negative response is returned anyway and the test can be repeated. The net result is that, on a high number of queries, system times are significantly reduced with respect to the base hierarchical protocol. This version of the recognition process takes only about 560 s, in order to identify all the 100 subjects in the probe set, instead of 880 s required by the protocol with K forced to 10. In the hierarchical model without K, RR is 100%, because in this case also the subjects retrieved beyond the tenth position both from face and from ear are verified by the fingerprint module.

VI. Conclusion

Recent research in the biometric field has been very intense, and a high number of physical and behavioral features have been studied. Some of them are more reliable than others, yet none is free from limitations. For this reason, the present trend is to combine more biometries in one system. We described two multimodal architectures: the hierarchical protocol, already known in literature, and the NCTP, introduced here. We also addressed the problem of data normalization and proposed the new QLS function, which overcomes the limits of the present ones. A further original contribution was the introduction of a new reliability index that a system can associate to its own responses, i.e., the SRR. We proposed two different indexes, also providing the conditions for their actual exploitability within the single subsystems, e.g., the *a priori* identification of thresholds. More studies will be performed to obtain further reliability measures, and a way to compute the overall reliability of a multibiometric system. Finally, we reported experimental results, which validate all our theoretical statements. We exploited three biometries, face, ear, and fingerprint. A number of experiments were performed on different databases, among those commonly used. The choice of subjects within each database aimed at facilitating future comparisons.

In this paper, we only considered face, ear, and fingerprints. Future developments will regard the extension to a higher number of biometries, so that a higher number of subsystems could be involved in our analysis. In particular, more experiments will aim at exploring different strategies for data fusion in the NCTP.

References

[1] N. Aaraj, S. Ravi, S. Raghunathan, and N. K. Jha, "Architectures for efficient face authentication in embedded systems," in *Proc. Design, Autom. Test Eur.*, Mar. 2006, vol. 2, pp. 1–6.

[2] M. D. Marsico, M. Nappi, and D. Riccio, "FARO: Face recognition against occlusions and expression variations," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 121–132, Jan. 2010.

[3] A. F. Abate, M. Nappi, D. Riccio, and G. Tortora, "RBS: A robust bimodal system for face recognition," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 17, no. 4, pp. 497–514, 2007.

[4] N. J. Belkin, P. B. Kantor, E. A. Fox, and J. A. Shaw, "Combining evidence of multiple query representation for information retrieval," *Inf. Process. Manag.*, vol. 3, no. 31, pp. 431–448, 1995.

[5] R. M. Bolle, J. H. Connell, S. Pananti, N. K. Ratha, and A. W. Senior, "The relation between the ROC curve and the CMC," in *Proc. 4th IEEE Work. Automat. Identification Adv. Technol.*, 2005, pp. 15–20.

[6] D. Delgado-Gomez, F. Sukno, D. Aguado, C. Santacruz, and A. Artes-Rodriguez, "Individual identification using personality traits," *J. Netw. Comput. Appl.*, vol. 33, no. 3, pp. 293–299, May 2010.

[7] M. D. Marsico, M. Nappi, and D. Riccio, "HERO: Human ear recognition against occlusions," in *Proc. IEEE Comput. Soc. Workshop Biometrics—In Assoc. IEEE Conf. Comput. Vis. Pattern Recognit.—CVPR*, San Francisco, CA, 18 Jun. 2010, pp. 320–325.

[8] R. Distasi, M. Nappi, and D. Riccio, "A range/domain approximation error based approach for fractal image compression," *IEEE Trans. Image Process.*, vol. 15, no. 1, pp. 89–97, Jan. 2006.

[9] B. Dorizzi, S. Garcia-Salicetti, and L. Allano, "Multimodality in biosecure: Evaluation on real versus virtual subjects," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 5, pp. 1089–1092.

[10] Y. Fisher, *Fractal Image Compression: Theory and Application*. New York: Springer-Verlag, 1994.

[11] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Proc. Int. Conf. Audio Video-Based Biometric Person Authentication (AVBPA 2003)*, Jun., pp. 44–51.

[12] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, 1999.

[13] S. Garcia-Salicetti, M. A. Mellakh, L. Allano, and B. Dorizzi, "A generic protocol for multibiometric systems evaluation on virtual and real subjects," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2005, pp. 494–502.

[14] M. He, S.-J. Horng, P. Fan, R.-S. Run, R.-J. Chen, J.-L. Lai, M. K. Khan, and K. O. Sentosa, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Rec.*, no. 43, pp. 1789–1800, 2010.

[15] X. He, D. Cai, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[16] X. He, D. Cai, S. Yan, and H.-J. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2005, vol. 2, pp. 1208–1213.

[17] H. P.-K. Hui, H. M. Meng, and M.-W. Mak, "Adaptive weight estimation in multi-biometric verification using fuzzy logic decision fusion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, vol. 1, pp. 501–504.

[18] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005.

[19] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004.

[20] M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve procedure for the characterization of human faces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 103–108, Jan. 1990.

[21] J. Kittler, M. Hanef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[22] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo, "Reliability-based decision fusion in multimodal biometric verification," *EURASIP J. Adv. Signal Proc.*, vol. 2007, no. 1, pp. 74–83, 2007.

[23] J. H. Lee, "Analyses of multiple evidence combination," in *Proc. 20th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval.*, Philadelphia, PA, 1997, pp. 267–276.

[24] J-B Li, J-S Pan, and S-C Chu, "Kernel class-wise locality preserving projection," *Inf. Sci.*, no. 178, pp. 1825–1835, 2008.

[25] J-B Li, J-S Pan, and Z-M Lu, "Kernel optimization-based discriminant analysis for face recognition," *Neural Comput. Appl.*, vol. 18, pp. 603–612, 2009.

[26] A. M. Martinez, "Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 6, pp. 748–763, Jun. 2002.

[27] J. Matas, M. Hamouz, K. Jonsson, J. Kittler, Y. Li, C. Kotropoulos, A. Tefas, I. Pitas, T. Tan, H. Yan, F. Smeraldi, J. Begun, N. Capdevielle, W. Gerstner, S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Comparison of face verification results on the XM2VTS database," in *Proc. Int. Conf. Pattern Recognit.*, 2000, vol. 4, pp. 858–863.

[28] O. Melnik, Y. Vardi, and C. H. Zhang, "Mixed group ranks: Preference and confidence in classifier combination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 973–981, Aug. 2004.

[29] L. Osadciw and K. Veeramachaneni, "Sensor network management through fitness function design in multi-objective optimization," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Nov. 2007, pp. 1648–1651.

[30] J. P. Phillips, H. Moon, A. S. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[31] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002: Evaluation report," National Institute of Standards and Technology, Gaithersburg, MD, NISTIR 6965, 2003.

[32] N. Poh and S. Bengio, "Improving fusion with margin-derived confidence in biometric authentication tasks," in *Proc. 5th Int. Conf. Audio-Video-Based Biometric Person Authentication (AVBPA)*, 2005, pp. 474–483.

[33] N. Poh and S. Bengio, "Can chimeric persons be used in multimodal biometric authentication experiments?" in *Proc. Workshop Multimodal Interaction Related Mach. Learning Algorithms*, 2005, pp. 87–100.

[34] M. E. Renda and U. Straccia, "Metasearch: Rank versus score based rank list fusion methods (without training data)," CNR, Pisa, Tech. Rep. 2002-TR-07.

[35] A. Ross and A. K. Jain, "Information fusion in biometrics," *Pattern Recognit. Lett.*, vol. 24, no. 13, pp. 2115–2125, Sep. 2003.

[36] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognit.*, vol. 36, no. 10, pp. 293–302, 2003.

[37] K. Saeed and M. Nammous, "A speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image," *IEEE Trans. Ind. Electron.*, vol. 54, no. 2, pp. 887–897, Apr. 2007.

[38] R. Snelick, M. Indovina, J. Yen, and A. Mink, "Multimodal biometrics: Issues in design and testing," in *Proc. Int. Conf. Multimodal Interfaces*, Nov. 2003, pp. 68–72.

[39] T. Tan and H. Yan, "Face recognition using the weighted fractal neighbor distance," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 4, pp. 576–582, Nov. 2005.

[40] K. Toh and W. Yau, "Combination of hyperbolic functions for multimodal biometrics data fusion," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 2, pp. 1196–1209, Apr. 2004.

[41] B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan, *Studies of Biometric Fusion- Executive Summary*, National Institute of Standards and Technology, Gaithersburg, MD, NISTIR 7346, Sep. 2006.

[42] V. N. Vapnik, *Statistical Learning Theory*. New York: Springer-Verlag, 1998.

[43] K. Veeramachaneni, L. Osadciw, and P. K. Varshney, "An adaptive multimodal biometric management algorithm," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 344–356, Aug. 2005.

[44] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 511–518.

[45] R. V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *Int. J. Biometrics*, vol. 1, no. 1, pp. 81–113, 2008.

[46] P. Yan and K. W. Bowyer, "Multi-biometrics 2D and 3D ear recognition," in *Proc. Audio-Video-Based Biometric Person Authentication*, Jul. 2005, pp. 503–512.

[47] Z. E. Bhatti, Face and eyes detection using open CV. (2008, Jan. 25). [Online]. Available: http://www.codeproject.com/KB/library/eyes.aspx?fid=990485&df=90&mpp=25&noise=3&sort=Position&view=Quick&select=2514967&fr=26

[48] Bioscrypt. (2008, Jun. 6) [Online]. Available: http://www.bioscrypt.com

[49] A. A. Wardak, *Practical Guidelines and Examples for the Users of the TMS320C6713 DSK*, World Academy of Science, Engineering and Technology (45), pp. 507–512, 2008.

[50] International Fingerprint Verification Competition (2002), FVC2002 [Online]. Available: http://bias.csr.unibo.it/fvc2002,06-06-2008

[51] (2008, Jun. 6). Notre Dame Ear Database [Online]. Available: http://www.nd.edu/∼cvrl/UNDBiometricsDatabase.html

[52] (2008, Jun. 6). OpenCV [Online]. Available: http://sourceforge.net/projects/opencvlibrary/

[53] (2008, Jun. 6). SecurIDent$^{\text{TM}}$ Biometric Face Recognition Products [Online] http://www.cryptometrics.com/products_securident.php

Maria De Marsico (M'02) was born in Salerno, Italy, in 1963. She received the Laurea degree (*cum laude*) in computer science from the University of Salerno, Salerno, Italy, in 1988.

She is currently an Assistant Professor of computer science at the Dipartimento di Informatica, Università di Roma "La Sapienza," Roma, Italy. Her main research interests include image processing, multibiometric systems, human–computer interaction.

Dr. De Marsico is a member of the Association of Computing Machinery and the International Association for Pattern Recognition.

Michele Nappi was born in Naples, Italy, in 1965. He received the Laurea degree (*cum laude*) in computer science from the University of Salerno, Salerno, Italy, in 1991, the M.Sc. degree in information and communication technology from the International Institute for Advanced Scientific Studies "E.R. Caianiello," Vietri sul Mare, Salerno, and the Ph.D. degree in applied mathematics and computer science from the University of Padova, Padova, Italy.

He is currently an Associate Professor of computer science at the Università di Salerno. His research interests include pattern recognition, image processing, compression and indexing, multimedia databases, and visual languages.

Dr. Nappi is a member of the International Association for Pattern Recognition.

Daniel Riccio was born in Cambridge, U.K., in 1978. He received the Laurea degree (*cum laude*) and the Ph.D. degree in computer science from the University of Salerno, Salerno, Italy, in 2002 and 2006, respectively.

He is currently an Assistant Professor at the Università di Salerno. His research interests include biometrics, fractal image compression, and indexing.

Dr. Riccio has been a member of the Italian Group of Italian Researcher in Pattern Recognition since 2004.

Genoveffa Tortora (SM'01) received the Laurea degree in computer science from the University of Salerno, Salerno, Italy, in 1978.

She is currently a Full Professor in computer science, Università di Salerno, where she leaded the Faculty of Scienze Matematiche Fisiche e Naturali from 2000 to 2008. She is the author or coauthor of more than 150 papers and 3 books. Her research interests include software engineering, human-computer interaction, visual languages, geographic information systems, image processing, and virtual reality.

Prof. Tortora is a Senior Member of the IEEE Computer Society, an Associate Editor of several international journals, And a member of the Steering Committee of the IEEE Symposia on Human-Centric Computing Languages and Environments. She has also been a General Chair, Program Chair, and part of the Program Committee in several International Conferences.