Fusion of Multi-biometric Recognition Results by Representing Score and Reliability as a Complex Number

Maria De Marsico[1], Michele Nappi[2], and Daniel Riccio[3]

[1] Sapienza University of Rome, Rome – Italy
demarsico@di.uniroma1.it
[2] University of Salerno, Fisciano (SA) – Italy
mnappi@unisa.it
[3] University of Naples Federico II, Napoli – Italy
daniel.riccio@unina.it

Abstract. A critical element in multi-biometrics systems, is the rule to fuse the information from the different sources. The component sub-systems are often designed to further produce indices of input image quality and/or of system reliability. These indices can be used as weights assigned to scores (weighted fusion) or as a selection criterion to identify the subset of systems that actually take part in a single fusion operation. Many solutions rely on the estimation of the joint distributions of conditional probabilities of the scores from the single subsystems. The negative counterpart is that such very effective solutions require training and a high number of training samples, and also assume that score distributions are stable over time. In this paper we propose a unified representation of the score and of the quality/reliability index that simplifies the process of fusion, provides performance comparable to those currently offered by top performing schemes, yet does not require a prior estimation of score distributions. This is an interesting feature in highly dynamic systems, where the set of relevant subjects may undergo significant variations across time.

Keywords: Reliability, unified value score-reliability, complex numbers.

1 Introduction

Multi-biometric systems [16] are considered as one of the best viable solutions to overcome limitations of classical single biometrics, since flaws of one sub-system may be balanced by strengths of a companion one. Among the most relevant issues raised by the combined approach, we mention the need for an effective fusion strategy of the results. Information fusion in a biometric system can be performed at feature, score, or decision level [6], but most schemes in literature opt for score level fusion [5]. Score normalization is one of the important aspects to consider during fusion. Fusion schemes may also rely on treating scores as a unified feature vector, which requires a further classifier, or on transforming the scores in a posteriori probabilities [10]. A further issue is represented by the introduction of quality measures computed for the input samples [7][8] and of confidence margins [10]. The former (e.g. sharpness, lighting) allows to

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 302–309, 2013.

possibly discard problematic samples, but can also be exploited after classification, as a weight on the final obtained score. The latter can be used after classification to decide whether to trust in the system response. Two trends are currently developing, to take them into account. In the first one, all subsystems always participate in the fusion, and the quality is used to weight their responses. In the second one, only a subset of subsystems takes part from time to time in the fusion, which are selected according to reliability of their responses. In both cases, reliability measure is an additional information, and mostly handled as a separate value.

Among the many simple score fusion rules (e.g. sum, weighted sum, product, min, max) [10], a number of authors claim that simple sum is the best compromise between simplicity and performance. On the other hand, significantly better results can be obtained through more complex techniques [1]. Likelihood Ratio (LR) is one of the most interesting ones. The authors of [17] discuss how product of Likelihood Ratios represents an optimal rule to get the highest Genuine Accept Rate (GAR) for a fixed False Accept Rate (FAR) in a multi-biometric system. The main disadvantage of this rule of fusion is that it assumes an accurate estimate of the joint distribution (across all the subsystems) of the conditional probabilities of the scores achieved by genuine and impostors users. This requires a complex modeling phase (in [12] finite Gaussian Mixture Model - GMM is used to model the genuine and impostor score densities), and a significant number of training samples. Despite such complexity, performance of systems whose operational parameters are based on a preliminary estimation of score distributions, may degrade if these significantly change along time. Nevertheless, given the optimality of LR, it can be considered as an asymptotic limit for which to strive when devising a new rule of fusion, while trying to overcome its limitations.

This work proposes a novel way of assembling the recognition score and the response reliability measure into a single complex number, facilitating the fusion in identification operations. The technique used in [18] maps the feature vectors from two biometric systems into the real and imaginary part of a complex vector. We rather use the score and the reliability, associated with an identification result by a single subsystem, to derive the module and the anomaly in the exponential representation of a complex number. The fusion of results related to the same identity relies on a modified operation of complex product among the responses from the single subsystems returning such identity. Further processing detailed in Section 2 allows to obtain a single real value as the final score assigned to a given identity by the global system. We use the *System Response Reliability* (SRR) measure [4], which does not require training, and is able to provide reliability information for each single recognition operation, differently from aggregate values like Recognition Rate.

2 Merging Scores and Reliability Values by Complex Numbers

There is a major difference between a quality measure for an input sample and a reliability measure for the response of a biometric system. The former is generally bound to a specific biometrics and to a specific classifier: for instance, a measure based on the quality of minutiae only applies to fingerprint recognition, which specifically uses minutiae for classification. Reliability measures devised without any reference to a specific biometric trait and/or algorithm can be generally used for any recognition system. The biometrics-independent reliability measure that we exploit

takes into account the composition of the gallery of the recognition system. From now on, we will use the *System Response Reliability* (SRR) [4] as a measure of reliability. The SRR relies on different versions of function φ defined in [4], which respectively exploit the relative distance and the density ratio, as well as a combination of them. All three functions measure the amount of "confusion" among possible candidates. We assume that the result of an identification operation is the whole gallery ordered by distance from the probe. Given a probe p and a system A with gallery G, the first function is:

$$\varphi_1(p) = \frac{d(p, g_{i_2}) - d(p, g_{i_1})}{d(p, g_{i_{|G|}})}, \tag{1}$$

where d is a distance function with codomain [0, 1], g_{i_k} is the k-th identity in the returned gallery ordering, and $|G|$ is the size of the gallery; distance values falling in a different codomain can be suitably normalized. Here we use the *Quasi Linear Sigmoidal (QLS)* [4]. It better preserves the original distribution of data, and is robust to a missing reliable evaluation for the maximum value. With relative distance if a person is genuine, there is a great difference between the distance from the first retrieved identity and the immediately closest one. Density ratio is instead defined as:

$$\varphi_2(p) = 1 - |N_b| / |G|, \quad \text{where} \quad N_b = \left\{ g_{i_k} \in G \mid d(p, g_{i_k}) < 2 \cdot d(p, g_{i_1}) \right\}. \tag{2}$$

The formula considers the distinct identities returned during identification as a cloud centred in p; the higher the density of this cloud, the more unreliable is the answer, as there are many individuals as potential candidates. In this paper we also adopt a variation of the density ratio. As one can observe in the definition of N_b in (2), the radius of the considered cloud depends on the distance from the probe of the first returned identity and from a constant. This function is less sensible to outliers, than φ_1, but it considers narrower clouds when the first retrieved identity is closer to the probe. On the contrary, a large distance takes to a wider cloud, which can be expected to be more crowded anyway. To further improve φ, we define here the term N_c such that the cloud radius depends on the difference between the first two distances:

$$\varphi_3(p) = 1 - |N_c| / |G|, \quad \text{where} \quad N_c = \left\{ g_{i_k} \in G \mid d(p, g_{i_k}) < \frac{(1 + d(p, g_{i_2}))(1 + d(p, g_{i_2}) - d(p, g_{i_1}))}{4} \right\}. \tag{3}$$

The new radius increases with the second distance, and with the difference between the first and the second ones. In practice, the farthest the second returned subject from the probe, also with respect to the first one, the wider the cloud we inspect. However, being all distances in [0,1], we add 1 to both terms to maintain direct proportionality. We also use the appropriate normalization factor since the value of d is in [0,1], and the maximum value for the numerator in (3) is *4*.

Once chosen the function φ to use, some more steps are required to compute the value of SRR for the probe at hand. For each $\varphi(p)$, we identify a value $\overline{\varphi}$ fostering a

correct separation between genuine and impostor subjects. We also define $S(\varphi(p),\bar{\varphi})$ as the width of the subinterval from $\bar{\varphi}$ to the proper extreme of the overall [0,1) interval of possible values, depending on the comparison between the current $\varphi(p)$ and $\bar{\varphi}$:

$$S(\varphi(p),\bar{\varphi})=\begin{cases} 1-\bar{\varphi} & \text{if} \quad \varphi(p)>\bar{\varphi} \\ \bar{\varphi} & \text{otherwise} \end{cases}. \tag{4}$$

SRR index can finally be defined as:

$$SRR = (\varphi(p)-\bar{\varphi})/S(\bar{\varphi}). \tag{5}$$

In detail, we measure the distance between $\varphi(p)$ and the "critical" point $\bar{\varphi}$, which gets higher values for $\varphi(p)$ much higher than $\bar{\varphi}$ (genuine), or for $\varphi(p)$ much lower than $\bar{\varphi}$ (impostors). However, it is also important to take into account how much it is significant with respect to the subinterval over which it is measured. SRR gets values in [-1, 1]. More details on computation and its motivations can be found in [4].

Numbers in the complex field can be represented as $a+ib$,or by the exponential representation $z=\rho e^{i\theta}$, where ρ is the modulus and θ is the anomaly. In our fusion, the score and the reliability measure are used to derive the ρ and the θ of this latter representation, respectively. We chose this representation because it better adapts to the kind of processing for fusion. In fact, the product operation with the real/imaginary form, would suffer from misleading cross-influence between heterogeneous parts. Given a score s and a reliability value srr, $\rho=(1+s)$ and $\theta=srr$. Since s is in the interval [0,1], and srr ranges between -1 and 1, then ρ is in the interval [1, 2] and θ is in [-1,1]. We take the set of the complex numbers obtained in this way from the values returned by the different subsystems voting for the same identity in a multibiometric identification. We define a new operation over them that we denote with ⊗, such that:

$$z=\bigotimes_{j=1}^{k} z_j=\frac{\prod_{j=1}^{k}\rho_j}{k}e^{i\frac{\sum_{j=1}^{k}\theta_j}{k}} \tag{6}$$

Thanks to the denominators, the final $\rho_{\otimes}$ and $\theta_{\otimes}$ are still in the same intervals as the initial values. The final composed score will be $s_{\otimes}=(\rho_{\otimes}/2)$ and the final reliability will be $srr_{\otimes}=\theta_{\otimes}$, and will be respectively in the interval [0,1] and [-1,1]. In the absence of a reliability measure, its value can be set to 1 for any response. The two values after fusion can again be used to obtain the exponential form of a complex number. This can be done for each group of subsystems voting for a same identity, so that at the end we will have a complex numbers for each candidate identity. However, we have to choose a winning identity, so we would prefer single and easier to compare values. To this aim, we first pass to the representation of the complex numbers in real and imaginary part $z=a+ib$, with $a,b\in$ R. The (a, b) pair can be interpreted as a couple of coordinates in a two-dimensional space, and as such can be represented in the Argand-Gauss plane (especially devised to represent complex numbers in this form).

Fig. 1 presents an example of the effects of the approach using *SRR* defined above with φ_1. The values $a+ib$ in the plots refer to the half-plane with positive x-axis (real part a, derived from scores). The first three plots represent pairs of points for the same classifier on three different biometrics (face, ear and iris). Section 4 reports details on datasets and classifiers. We see that genuine scores (red/light circles) are mainly concentrated in the first quadrant, while impostor scores (blue/dark squares) mainly lie in the fourth quadrant, with some overlap. The last plot is the result of the introduced operations over these values. Notice that the values for genuine users are distributed in the positive quadrant, while those for the impostors are concentrated in the negative one, but the interesting feature to notice is that values are much more sharply divided.

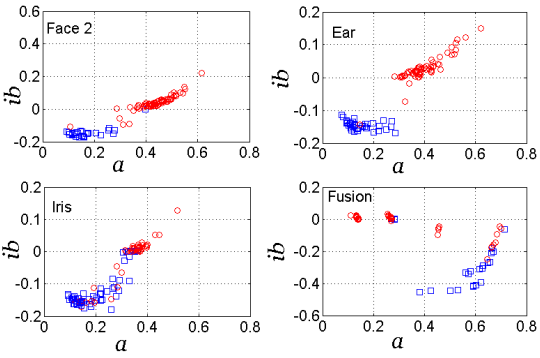

Fig. 1. First three plots: distribution of pairs real/imaginary parts obtained from the responses of a correlation based classifier (see below) over face, ear and iris datasets (left-right and top-bottom); last plot: the distribution after the product. Red/light circles are genuine scores, blue/dark squares are impostor scores.

As coordinates in a 2D space, (a, b) pairs can be further transformed in single values using Peano keys. Peano rule maps a 2D onto a 1D space such that two close points in the starting space, tend to be close also in the final one. However, the rule requires integer values, so that it is necessary to consistently map a and b onto integers with a finite number n of bits. In our implementation the new integers a_P and b_P have $n = 16$ bits. The associated Peano key K_P is obtained by interleaving bits from a_P and b_P, from the least significant to the most significant one, so to obtain a final value of 32 bits. Values for different identities can be straightforwardly compared.

3 Experimental Framework

The presented framework was tested in a multi-biometric setting (face, ear and iris) and compared with the LR discussed in [12], using the same implementation for the estimation of the GMM model. The multi-biometric database consists of Chimeric users whose biometric traits were taken from three different datasets. It is worth noticing that it is presently accepted that results obtained in this way are worthy of full reliability [9]. The number of subjects in the database is constrained by the size of the database of ears, namely 100 subjects in the Notre Dame Ear Database [13]. In order

to consider an open set identification setting, i.e. a situation where not all users are enrolled and impostors can also occur, the gallery consists of 75 enrolled subjects, for which there is a single image, while the probe is made up of 100 subjects, each accompanied by a single image. The faces are from a subset of AR-Faces database [14] (50 males and 50 females), for which 4 different datasets were considered: gallery (normal), Face-2 (smile), Face-5 (left-light) and Face-11 (scarf). The irises were from the first 100 subjects in UBIRISv1s1 database [15]. Performance were measured in terms of Recognition Rate (RR) and Equal Error Rate (EER) [2].

In order to understand the relation between the behavior of the presented framework and the classifier used, we tested it with Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and the local correlation-based classifier which is part of FACE system [3], indicated from now on as FACE for short. **Table 1** shows the performance on each dataset, which appear quite heterogeneous, as expected. This is interesting to understand later how the fusion technique works, not only when all classifiers provide optimal results, but also when one or more of them fail.

Table 1. Performance of single classifiers on each dataset, in terms of RR and EER

Dataset	PCA		LDA		FACE	
	RR	**EER**	**RR**	**EER**	**RR**	**EER**
Face-2	0.97	0.039	0.94	0.027	0.97	0.052
Face-5	0.21	0.144	0.61	0.124	0.98	0.013
Face-11	0.04	0.441	0.05	0.354	0.93	0.053
Ear	0.65	0.207	0.76	0.091	0.85	0.120
Iris	0.69	0.092	0.74	0.093	0.62	0.185

Results in **Table 1** show that PCA and LDA are much more sensible to local variations within a face image. In particular on the Face 11 set, where the lower part is completely occluded by a scarf. In combinating with other biometrics, this condition may be particularly stressing for the fusion process, making this case very interesting.
In the first experiment, we tested the best function φ. The same classifier was applied to the different biometrics and the reliability was measured from time to time by a different φ. Given the score s_j (as an inverse of distance from the probe) from biometrics j (F=face, where F2, F5 and F11 indicate the datasets from AR-Faces, E=ear and I=iris), and given srr_j its reliability value, according to the chosen φ, Complex Fusion (CF) computes the presented operation for the three $(1+sj)ei^{\cdot srr_j}$. *Simple sum* rule was also tested, and results were comparable to those of complex values with no reliability, i.e. with the imaginary part set to 1 (CF none). For sake of space, **Table 2** only reports the results of FACE, which resulted better than PCA and LDA classifiers.

Table 2. RR and EER, when different φ functions are used in fusion of FACE results

Method	F2/E/I		F5/E/I		F11/E/I	
	RR	**EER**	**RR**	**EER**	**RR**	**EER**
Simp. sum	1.00	0.026	1.00	**0.001**	1.00	0.067
Simp. prod	1.00	0.026	1.00	0.001	1.00	0.060
CF none	1.00	0.046	1.00	0.033	0.97	0.039
CF φ_1	**1.00**	**0.020**	**1.00**	0.006	**1.00**	**0.033**
CF φ_2	1.00	0.246	1.00	0.153	0.99	0.342
CF φ_3	0.98	0.039	1.00	0.033	0.94	0.061

Table 2 shows that φ_1 and φ_3 give the best results and will be used to compare performance of Simple product, Complex fusion e Likelihood ratio. Values in **Table 2** highlight that the simple sum provides acceptable results, when the classifier offers good performance for every fused biometrics. However, it was observed that, having all biometrics the same weight, if one of them provides poor results this significantly influences the overall system performance, as confirmed by the results in **Table 3**, where each fusion technique is evaluated with all classifiers.

Table 3. Performance when different techniques are used for fusion, in terms of RR and EER

PCA	F2/E/I		F5/E/I		F11/E/I	
	RR	**EER**	**RR**	**EER**	**RR**	**EER**
Simple sum	1.00	0.073	1.00	0.112	1.00	0.278
Complex Fusion (φ_1)	0.99	0.420	0.72	0.329	0.65	0.560
Complex Fusion (φ_3)	1.00	0.326	0.74	0.333	0.64	0.470
Likelihood Ratio	1.00	0.033	0.95	0.140	0.85	0.214
LDA	**F2/E/I**		**F5/E/I**		**F11/E/I**	
	RR	**EER**	**RR**	**EER**	**RR**	**EER**
Simple sum	1.00	0.040	0.96	0.120	0.84	0.171
Complex Fusion (φ_1)	0.99	0.427	0.86	0.170	0.73	0.359
Complex Fusion (φ_3)	0.99	0.118	0.84	0.160	0.77	0.181
Likelihood Ratio	1.00	0.040	0.99	0.112	0.95	0.171
FACE	**F2/E/I**		**F5/E/I**		**F11/E/I**	
	RR	**EER**	**RR**	**EER**	**RR**	**EER**
Simple sum	1.00	0.026	1.00	0.001	1.00	0.067
Complex Fusion (φ_1)	1.00	0.020	1.00	0.006	1.00	0.033
Complex Fusion (φ_3)	0.99	0.039	1.00	0.033	0.93	0.061
Likelihood Ratio	1.00	0.010	1.00	0.000	1.00	0.013

In **Table 2** and **Table 3** φ_1 provides the best results with a classifier robust to variations, like FACE. On the contrary, e.g., with PCA and partly with LDA, φ_3 sometimes provides better results. **Table 3** shows that, in many cases for PCA and LDA, complex fusion performance is below simple sum. This is because these two algorithms are both poorly robust to distortions, and provide poorly reliable responses. In fact, we would observe a wide overlap between genuine and impostor distributions. With FACE classifier we achieve both higher robustness, and higher reliability. The latter makes the fusion results with complex numbers better than those with simple sum, especially with function φ_1. The overall interesting aspect is that, using a robust classifier aligned with the state of the art, the proposed fusion technique is able to provide better results that simple sum and only slightly lower that the optimum LR. This is very important if we consider that it is simple like the sum, yet does not require any preliminary estimation of genuine and impostor score distributions. In other words, at the expense of a slightly lower performance, we are able to adopt a strategy which is stable over time and delivers results which are congruous for each single probe, we avoid an expensive training phase, and save computation even in operational phases.

4 Conclusions

This paper has presented a multi-biometric fusion framework based on the joint representation in the complex field of score values and reliability measures. The experimental

results show that in the case of robust classifiers the performance of the proposed framework are comparable to those of LR, which proves to be the best criterion for fusion. The product of complex values, however, has the further advantage of not needing an accurate approximation of the distributions of the scores. Future studies will focus on even better criteria to use the complex representation.

References

1. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D Face Recognition: A Survey. Pattern Recognition Letters 28(14), 1885–1906 (2007)
2. Bolle, R.M., Connell, J.H., Pananti, S., Ratha, N.K., Senior, A.W.: The Relation Between the ROC Curve and the CMC. In: Proc. of 4th IEEE Work. on Automatic Identification Adv. Technologies, pp. 15–20 (2005)
3. De Marsico, M., Nappi, M., Riccio, D.: FACE: Face analysis for commercial entities. In: Proc. of Int. Conference on Image Processing (ICIP), Honk Kong, pp. 1597–1600 (2010)
4. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: NABS: Novel Approaches for Biometric Systems. IEEE Trans. on Systems, Man, and Cybernetics–Part C: Applications and Reviews 40(6) (2011)
5. Ross, A., Jain, A.K.: Information Fusion in Biometrics. Pattern Recognition Letters 24(13), 2115–2125 (2003)
6. Jain, A.K., Nandakumar, K., Ross, A.: Score Normalization in multimodal biometric systems. Pattern Recognition 38(12), 2270–2285 (2005)
7. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J.: Discriminative Multimodal Biometric Authentication Based on Quality Measures. Pattern Recognition 38(5), 777–779 (2005)
8. Nandakumar, K., Chen, Y., Jain, A.K., Dass, S.: Quality-Based Score Level Fusion in Multibiometric Systems. In: Proc. Int'l Conf. Pattern Recognition, pp. 473–476 (August 2006)
9. Garcia-Salicetti, S., Mellakh, M.A., Allano, L., Dorizzi, B.: A Generic Protocol for Multibiometric Systems Evaluation on Virtual and Real Subjects. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 494–502. Springer, Heidelberg (2005)
10. Kittler, J., Hatef, M., Duin, R.P., Matas, J.G.: On Combining Classifiers. IEEE Trans. PAMI 20(3), 226–239 (1998)
11. Poh, N., Bengio, S.: Improving Fusion with Margin-Derived Confidence in Biometric Authentication Tasks. In: Proc. Fifth Int'l Conf. Audio Video-Based Biometric Person Authentication, pp. 474–483 (July 2005)
12. Nandakumar, K., Chen, Y., Dass, S.C., Jain, A.K.: Likelihood Ratio-Based Biometric Score Fusion. IEEE Transactions on PAMI 30(2), 342–347 (2008)
13. Notre Dame Ear Database, `http://www.nd.edu/~cvrl/UNDBiometricsDatabase.html`
14. Martinez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. PAMI 24(6), 748–763 (2002)
15. Proença, H., Alexandre, L.A.: UBIRIS: A noisy iris image database. In: Roli, F., Vitulano, S. (eds.) ICIAP 2005. LNCS, vol. 3617, pp. 970–977. Springer, Heidelberg (2005)
16. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer (2006)
17. Ulery, B., Hicklin, A.R., Watson, C., Fellner, W., Hallinan, P.: Studies of Biometric Fusion. Technical Report IR 7346, NIST (September 2006)
18. Yang, J., Yang, J., Zhang, D., Lua, J.: Feature fusion: parallel strategy vs. serial strategy. Pattern Recognition 36(6), 1369–1381 (2003)