An Experimental Comparison of Different Methods for Combining Biometric Identification Systems

Emanuela Marasco and Carlo Sansone

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

Abstract. Several works in the recent literature on biometrics demonstrate the efficiency of the multimodal fusion to enhance performance and reliability of the automatic recognition. In this paper, we experimentally compare the behavior of different rules for integrating different biometric identification systems. We investigated how the benefits of the fusion change by varying the set of the fused modalities, the adopted fusion scheme and the performance of the individual matchers. The experiments were carried out on two multimodal databases, using face and fingerprint. We considered trained and fixed fusion methods at score, rank and decision level.

1 Introduction

In the recent literature on biometrics, several researches demonstrate the efficiency of the multimodal fusion to enhance performance and reliability of the automatic recognition [1]. Integrating biometric information from multiple sources, multimodal biometric systems are able to improve the authentication performance, increase the population coverage, offer user choice, make biometric authentication systems more reliable and robust to spoofing.

However, the benefits of multibiometrics depend on the accuracy, complementarity, reliability and quality measurement of their component biometric experts. Moreover, when designing a multibiometric system, several factors should be considered. These concern the choice and the number of biometric traits, the level of integration and the mechanism adopted to consolidate the information provided by multiple traits. Fusion at match score level is usually preferred due to the easy to access and combine the scores presented by different modalities. The parallel fusion strategy has been extensively explored, however serial and hybrid architectures present important advantages. In particular, the serial fusion considers the biometric matchers one at a time, and makes a reliable decision by employing few experts and activating the remaining experts only for difficult cases. In general, it is desirable that a fusion scheme involves statistically independent modality matchers. In a multimodal fusion, the set of expert outputs

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 255–264, 2011.

is expected to be statistically independent, while in intramodal fusion, where the component matchers rely on the same biometric trait, a high dependency is expected among the expert outputs [2]. The merit of both multimodal and intramodal fusion has been demonstrated in [3].

Moreover, although individual modalities have proven to be reliable in ideal environments, they can be very sensitive to real environmental conditions. In real scenarios, it is difficult to acquire high quality samples, then biometric authentication errors are inevitable. The impact of adverse environmental conditions on the characteristics of the collected biometric data can be quantified by *quality measures*. It is evident that, a degradation in the quality level of the biometric signal input may affect the reliability of the matching process. The performance of the single modality matcher may change as the data quality changes and different modality matchers are sensitive to different aspects of the signal quality. Then, the opinion of a matcher in the decision of the ensemble have to be appropriately weighted, by assigning a higher weight to the matcher with higher quality data. The same observation has to be considered for the reliability, accuracy and competence of each component matcher [4]. The effectiveness of using quality measures in the fusion has been demonstrated in [5].

The key to create a secure multimodal biometric system is in how the information from different modalities is fused [1]. In identification mode, the consolidation of biometric information can be performed at various levels: sensor level, feature extraction level, match score level, rank level and decision level. Consolidating data at an early stage of the recognition process involves a higher informative contain concerning the biometric input. Thus, it is potentially able to provide better recognition results, but in practice concatenating data at a level before matching may result difficult or not possible. In particular, images captured from sensors with a different resolution are not compatible and feature vectors may be not accessible (they often are proprietary). Combining match scores provided from different matchers is the most effective fusion strategy because they offer the best trade-off between the informative contain and the ease to implement the fusion.

In this paper, we experimentally compare the behavior of different integration rules at different fusion levels for biometric identification systems. Considering the state-of-the-art this kind of comparison has been done only for the verification task and not for the identification [6]. We investigated how the benefits of the fusion change by varying the set of the fused modalities, the fusion scheme exploited and the performance of the individual matchers. The paper is organized as follows. Section 2, presents the analyzed combination rules. Section 3 describes the data, the experimental procedure and it reports our results. Section 4 draws our conclusions.

2 The Considered Fusion Methods

The comparison was carried out by considering the most commonly adopted fusion mechanisms for identification scenario. Among the possible levels, we focus on fusion approaches at score, rank and hybrid rank-score level.

2.1 Fusion Approaches at Score-Level

Fusion at match score level concerns combining the match scores generated by multiple classifiers in order to make a decision about the identity of the subject. In literature, the fusion at score level is performed by employing different approaches [7] based on different models [8]. We considered the *transformation-based schemes* which are described below.

The match scores provided by different matchers are firstly transformed into a common domain (*score normalization*) which refers to changing the location and scale parameters of the match score distributions outputs of the individual matchers [9]. Then, normalized match scores are combined using a simple fusion rule. The operators which are commonly used in the literature are *min*, *max*, *median*,*weighted sum* and *weighted product*, defined by (1), (2), (3), (4) and (5).

$$s_{min} = \min_k s_k \tag{1}$$

$$s_{max} = \max_k s_k \tag{2}$$

$$s_{median} = median_k s_k \tag{3}$$

$$s_{sum} = \sum_{k=1}^{K} w_k s_k \tag{4}$$

$$s_{prod} = \prod_{k=1}^{K} s_k^{wk} \tag{5}$$

where w_k are parameters that need to be estimated. The simple *sum* operator (or *mean*) is a special case of *weighted sum* with $w = \frac{1}{N}$, while the *product* operator is a special case of *weighted product* with $w = 1$. The operators which do not contain parameters to be tuned, are known as *fixed* combiners [10]. Based on experimental results, researchers agree that *fixed* rules usually perform well for ensemble of classifiers having similar performance, while *trained* rules handle better matchers having different accuracy. Thus, when fusing different modalities, individual matchers often exhibit different performance, then for this problem *trained* rules should perform better than *fixed* rules [6]. It has been shown that, the simple sum rule gives very good accuracy in combining multiple biometric systems [6]. Due to the diversity of scenarios encountered in the datasets, training and using a single fusion rule on the entire dataset may not be appropriate. Recently [11], the idea of dynamically selecting biometric fusion algorithms has been adopted.

2.2 Fusion Approaches at Rank-Level

For systems operating in identification mode, rank level fusion is a viable option. It provides a richer information into the decision-making process compared to

the decision level, without requiring a normalization phase before combining [12]. Let K be the number of matchers to be fused and N the number of enrolled users. Let r_{ij} be the rank assigned to the j^{th} user enrolled in the database by the i^{th} matcher, $i = 1 \dots K$, and $j = 1 \dots N$, then R_{ij}.

Highest rank scheme. For each subject, the combined rank is given by the lowest rank (6). This rank fusion technique presents the advantage of utilizing the strength of each matcher.

$$R_i = \min_{k=1}^{K} r_{ik}, \quad i = 1, 2, ...N \tag{6}$$

Borda Count scheme. For each subject, the combined rank is given by the sum of the ranks assigned by the individual matchers (7). Such a rule presents the advantage of taking into account the variability of the single matcher outputs. Its drawbacks lie in the assumptions that, the matchers are statistically independent and they perform equally well. This makes the Borda Count method highly vulnerable to the effect of weak classifiers.

$$R_i = \sum_{k=1}^{K} r_{ik}, \quad i = 1, 2, ...N \tag{7}$$

Logistic regression scheme. The fused rank is a weighted sum of the individual ranks.

$$R_i = \sum_{k=1}^{K} w_k r_{ik}, \quad i = 1, 2, ...N \tag{8}$$

The weight w_k, $i = 1 \dots K$, (see equation (8)), is determined through a training phase by logistic regression. This method is useful when the different biometric matchers have significant differences in their accuracies [8].

There is increasing interest in impact of the matcher reliability estimation in the context of fusion in biometrics. However, incorporating reliability information in rank level fusion represents a topic whose the discussion in the literature is at present still limited. The idea is to use reliability in a multibiometric system for reducing the weight of potential incorrect unimodal decisions.

2.3 Fusion Approaches at Hybrid Rank-Score Level

We considered the *predictor-based majority voting*, the *predictor-based sequential* and *predictor-based borda count* proposed in [13]. For each modality, a classifier (predictor) was trained using the hybrid information given by the ratios between scores in terms of ranks with respect to the rank one identity. Such a classifier is used to learn the decision boundary between the correct identification region and the erroneous one.

For a given probe, K unimodal matchers are employed and the winner is the identity to which the majority of matchers have assigned a rank value equal to one. The majority vote will result in an ensemble decision [14]:

$$\arg \max_{i=1...N} \sum_{k=1}^{K} d_{ik} \cdot v_k \tag{9}$$

where the binary variable d_{ik} is 1 if the k^{th} matcher outputs identity i in rank-1, and the binary variable v_k is 1 if the identification is deemed to be *correct* by the k^{th} predictor. The majority vote scheme assigns an identity to the probe only if the output of at least $\lfloor \Sigma_{k=1}^{K} v_k \rfloor + 1$ unimodal systems correspond to the same identity and are deemed to be correct by v_k.

In the serial scheme, the decisional process is split into two successive stages [15]. The subject to be authenticated submits the first biometric modality to the system which is processed and matched against all the templates present in the gallery. If the resulting identity is labeled to be correct by the predictor module, the input biometric trait is associated to the current identity, otherwise the system suspends the decision and an additional processing stage is performed. In the second stage, *K-1* additional biometric modalities are automatically requested and a voting strategy involving *K-1* unimodal matchers is adopted in the second stage. It can be formulated as follows:

$$Id_m = \begin{cases} Id_u, & \text{if} \quad v_u = 1 \\ \arg\max_{i=1\ldots N} \sum_{k=1}^{K-1} d_{ik} \cdot v_k & \text{if} \quad v_u = 0 \end{cases} \tag{10}$$

where Id_m is the output of the multimodal system and Id_u is the output of the unimodal system at the first stage. In order to maximize the performance of the multimodal system in terms of accuracy and recognition time, on this second stage all the matchers are combined and further stages are avoided.

In the *Borda Count* model, the rank for each identity in the database is calculated as the weighted sum of the individual ranks assigned by the K modality matchers:

$$R_i = \sum_{k=1}^{K} w_k \, r_{ik}, \qquad i = 1, 2, \ldots N \tag{11}$$

In the predictor-based fusion scheme, the unimodal outputs labeled as errors by the predictor have to be excluded from the sum in the equation above which determines the fused rank for each identity. The weight w_k was computed as the ratio between the number of correct identifications detected by the predictor and the total number of test probes.

2.4 Fusion Approaches at Decision-Level

We considered the *pure majority voting* scheme, where the final output corresponds to the most commonly occurring output. For a given probe, the outputs of K modality matchers are examined and the identity to which the majority of matchers have assigned rank one is the *winner*. The majority vote will result in an ensemble decision [14]:

$$\arg \max_{i=1\ldots N} \sum_{k=1}^{K} d_{ik} \tag{12}$$

where the binary variable d_{ik} is 1 if the k^{th} matcher outputs identity i in rank-1.

3 Experimental Results

3.1 Datasets

The performance of the proposed strategy was evaluated on two databases. The first is the West Virginia University (WVU) multimodal biometric database. A subset of this database pertaining to the fingerprint (left thumb [FL1], right thumb [FR1], left index [FL2], right index [FR2]) and face modalities of 240 subjects was used in our experiments. Five samples per subject for each modality were available. Table 1 provides the details of the database. For the *face* modality, frontal images were collected in a controlled scenario. For the *fingerprint* modality, images were collected using an optical biometric scanner, without explicitly controlling the quality [16]. The entire dataset was divided into five sets: the first sample of each identity was used to compose the *gallery* and the remaining four samples of each identity were used as *probes* (P_1, P_2, P_3, P_4). The VeriFinger software was used for generating the fingerprint scores and the VeriLook software was used for generating the face scores.

Table 1. WVU Multimodal Biometric Database

Biometric	Subjects	Samples	Scores
Face	240	5 per subject	Gen 1200×4 Imp $240 \times 239 \times 25$
Fingerprint	240	5 per finger	Gen $(1200 \times 4) \times 4$ Imp $(240 \times 239 \times 25) \times 4$

The second database is a subset of the BioSecure multimodal database. This database contains 51 subjects in the Development Set (training) and 156 different subjects in the Evaluation Set (testing). For each subject, four biometric samples are available over two sessions: session 1 and session 2. The first sample of each subject in the first session was used to compose the gallery database while the second sample of the first session and the two samples of the second session were used as probes (P_1, P_2, P_3). For the purpose of this study, we used the face and three fingerprint modalities, denoted as *fnf*, *fo1*, *fo2* and *fo3*, respectively [17]. The details about the number of match scores per person are reported in Tables 2.

3.2 Results

In this paper, we studied the behavior of trained and fixed rules when the combined matchers showed good individual performance and when the matchers achieved a significant individual percentage of errors. Our first experiments focused on comparing the performance of the fusion rules for a set of modality matchers having different a classification capability. Regarding WVU database we used one face matcher and four fingerprint matchers, while regarding Biosecure database we used one face matcher and three fingerprint matchers.

Table 2. The Biosecure database: Development(Dev) and Evaluation(Eva) sets

Dataset	Biometric	Subjects	Samples	Scores
Dev set	Face	51	4 per subject	Gen 204×3 Imp $51 \times 50 \times 16$
Dev set	Fingerprint	51	4 per subject	Gen $(204 \times 3) \times 3$ Imp $(51 \times 50 \times 16) \times 3$
Eva set	Face	156	4 per subject	Gen 624×3 Imp $156 \times 155 \times 16$
Eva set	Fingerprint	156	4 per subject	Gen $(624 \times 3) \times 3$ Imp $(156 \times 155 \times 16) \times 3$

Tables 3, 4 compare the performance of the existing fusion schemes which are subdivided in fixed and trained rules. WVU data, acquired in optimal environment conditions, offer a scenario where fixed rules are able to achieve good performance. Biosecure data represent a challenging scenario where fixed fusion rules do not perform well, due to the presence of samples with low quality; however, the score sum presents high accuracy, such a scheme represents a good choice to make fusion. The considered trained rules also are able to achieve good multimodal performance.

Table 3. Performance of the analyzed fusion schemes averaged on the four probe sets in the WVU database

Level	Type	Rule	Accuracy
score	fixed	sum	99.58%
score	fixed	min	85.37%
score	fixed	max	99.44%
score	fixed	median	99.26%
score	fixed	product	89.81%
rank	fixed	borda count	95.42%
rank	fixed	highest rank	91.46%
decision	fixed	majority voting	98.75%
score	trained	weighted sum	98.23%
hybrid	trained	predictor-based majority voting	100%
hybrid	trained	predictor-based borda count	96.67%
hybrid	trained	predictor-based sequential	99.59%

Our further experiments aimed to compare the performance of the fusion schemes when increasing the number of the combined modality matchers (see Fig. 1). We proceeded with adding matchers in the fusion scheme based on their performance. On WVU database, the number of combined matchers ranges from

Table 4. Performance of the analyzed fusion schemes averaged on the four probe sets in the Biosecure database

Level	Type	Rule	Accuracy
score	fixed	sum	99.36%
score	fixed	min	80.56%
score	fixed	max	97.44%
score	fixed	median	97.86%
score	fixed	product	95.73%
rank	fixed	borda count	92.31%
rank	fixed	highest rank	81.62%
decision	fixed	majority voting	86.11%
score	trained	weighted sum	93.16%
hybrid	trained	predictor-based majority voting	97.22%
hybrid	trained	predictor-based borda count	92.52%
hybrid	trained	predictor-based sequential	96.58%

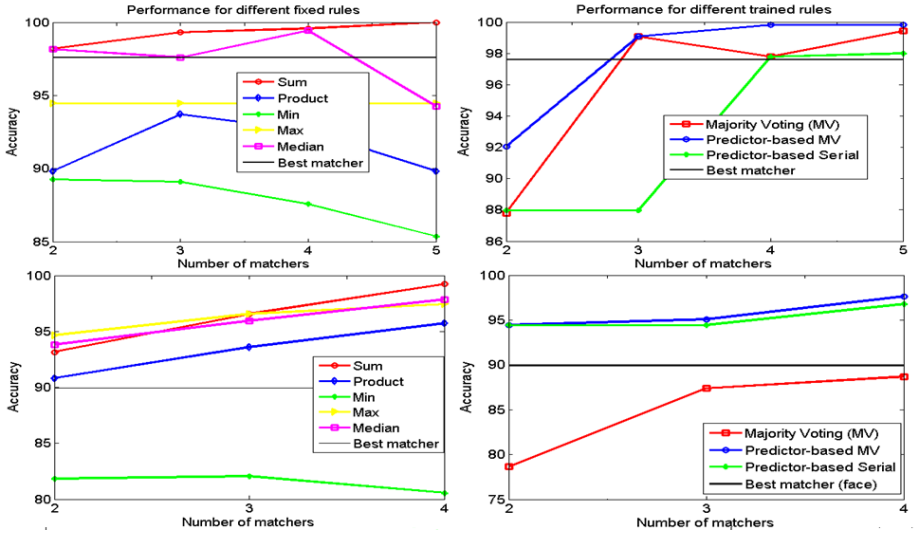

Fig. 1. Performance of fixed and trained fusion rules by varying the number of modality matchers: a) plots on the top show our results on WVU database which are obtained by averaging on the three probes, b) plots on the bottom show our results on Biosecure database which are obtained by averaging on the four probes

2 up to 5, while on Biosecure it ranges from 2 up to 4, and any considered subset was composed by the modality matchers having the best unimodal performance.

Regarding the performance of the fixed rules, our experiments showed that adding modalities to the fusion does not always imply increasing the multimodal performance. However we observed that, for the score sum adding modalities aims to increase the multimodal performance on both databases. On WVU database, the accuracy achieves 98.33% when two matchers are combined and increases to 100% when all the available matchers are combined, while on Biosecure database the accuracy achieves 93.33% when two matchers are combined and increases to 99.33% when all the available matchers are combined. Further, the max rule presents the same performance on the considered subsets of experts on WVU database, while on Biosecure the highest accuracy is achieved by using all the three matchers (97.44%).

In particular, fusion schemes at hybrid rank-score level always are able to improve the performance of the fusion schemes at rank level. The highest multimodal performance is obtained on WVU using trained rules.

4 Conclusions and Future Work

In this paper, we investigated how the benefits of the fusion change in identification scenario by varying the set of the fused modalities, the adopted fusion scheme and the performance of the individual matchers. The experimental comparison was carried out on two multimodal databases. Our experiments showed that, adding modalities to the fusion does not always imply increasing the multimodal performance. Our experiments showed also that, the improvement achievable by employing a multimodal solution depends on the adopted rules and on the data. Future research concerning this topic may regard a comparison of the considered schemes in presence of spoof attacks in order to study their security.

References

1. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters 24, 2115–2125 (2003)
2. Poh, N., Kittler, J.: Multimodal information fusion. Multimodal Signal Processing: Theory And Applications For Human-Computer Interaction (2009)
3. Sanchez, U., Kittler, J.: Fusion of talking face biometric modalities for personal identity verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 5, pp. V-V (2006)
4. Grother, P., Tabassi, E.: Performance of biometric quality measures. IEEE Transaction On Pattern Analysis and Machine Intelligence 29(4), 531–543 (2007)
5. Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J., Drygajlo, A.: Quality dependent fusion of intramodal and multimodal biometric experts. In: SPIE Biometric Technology for Human Identification IV, vol. 6539 (2007)
6. Roli, F., Kittler, J., Fumera, G., Muntoni, D.: An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. In: Roli, F., Kittler, J. (eds.) MCS 2002. LNCS, vol. 2364, pp. 325–336. Springer, Heidelberg (2002)

7. Nandakumar, K., Chen, Y., Dass, S., Jain, A.: Likelihood ratio-based biometric score fusion. IEEE Transaction on Pattern Analysis and Machine Intelligence 30(2), 342–347 (2008)
8. Ross, A., Nandakumar, K., Jain, A.: Handbook of MultiBiometrics. Springer, Heidelberg (2006)
9. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal bio-metric systems. Pattern Recognition 38(12), 2270–2285 (2005)
10. Poh, N.: Multi-System Biometric Authentication: Optimal Fusion And User-Specific Information. Ecole Polytechnique Fédéral de Lausanne (2006)
11. Vatsa, M., Singh, R., Noore, A., Ross, A.: On the dynamic selection of biomet-ric fusion algorithms. IEEE Transaction on Information Forensics and Security 5(3), 470–479 (2010)
12. Abaza, A., Ross, A.: Quality-based rank level fusion in biometrics. In: Third IEEE International Conference on Biometrics: Theory Applications and Systems (September 2009)
13. Marasco, E., Ross, A., Sansone, C.: Predicting identification errors in a multibiometric system based on ranks and scores. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (September 2010)
14. Kuncheva, L.I.: Combining Pattern Classifiers Method and Algorithms. Wiley, Chichester (2004)
15. Marcialis, G.L., Roli, F.: Serial fusion of fingerprint and face matchers. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 151–160. Springer, Heidelberg (2007)
16. Crihalmeanu, S., Ross, A., Schuckers, S., Hornak, L.: A protocol for multibiometric data acquisition, storage and dissemination. Technical Report. West Virginia University (2007)
17. Poh, N., Bourlai, T., Kittler, J.: A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. Pattern Recognition 43, 1094–1105 (2010)