

# Automatic Template Labeling in Extensible Multiagent Biometric Systems

Maria De Marsico<sup>2</sup>, Michele Nappi<sup>1</sup>, Daniel Riccio<sup>1</sup>, and Genny Tortora<sup>1</sup>

<sup>1</sup> Biometric and Image Processing Laboratory, University of Salerno,  
Via Ponte Don Melillo, 84084 Fisciano (Salerno) Italy

<sup>2</sup> Department of Computer Science, Sapienza University of Rome,  
Via Salaria 113, 00198 Rome Italy

demarsico@di.uniroma1.it, {mnappi,driccio,tortora}@unisa.it

**Abstract.** Many works in literature have demonstrated the superiority of multibiometric systems compared to single-biometrics ones, in terms of both accuracy and robustness. Most current multibiometric systems implement a static architecture, which does not change in time. However, the ability to progressively add more modules, either to process more biometrics or to exploit additional algorithms, might contribute to further enhance recognition performance. The addition of a new module (agent) to an already fully operational multiagent system usually requires its preliminary setup and training. In particular, it must be provided with a brand-new gallery, whose templates are suitably labeled according to the represented identities; alternatively, an existing database of templates, formerly built according to the suited feature extraction procedure, might be updated to include better quality items. It would be of paramount importance if the new agent can “inherit” the “experience that was already acquired by the other agents, including the creation of its gallery without having to undergo a full enrolling phase in its turn. We present here an algorithm to align a new module to the already existing ones in an automatic and unsupervised way. Experimental results show that our algorithm is effective both when the new database must be created from scratch (sample labeling), as well as when it is pre-existing and must be updated (sample updating). The latter operation can also be iteratively performed in running modules to dynamically update their galleries. In particular, we present here results achieved for face recognition.

**Keywords:** Multibiometric system, template labeling, unsupervised learning, template updating.

## 1 Introduction

Two issues related to biometric systems, in particular to multibiometric ones, are attracting research attention: a) template labeling, where the gallery of a new added module is created from scratch, and labels must be assigned to each template to assign it to a certain identity [7], and b) template updating, where old as well as corrupted templates for an identity are substituted by more recent or representative ones [8]. The

presence of more subsystems can represent an important added value in defining effective solutions for both these problems. Differently from other template updating approaches (see for example [6]), our proposal substantially relies on the widely recognized ability of multibiometric systems to provide better recognition accuracy, thanks to the joint contribution of either different biometrics, or modalities, or algorithms. In general, the presence of more modules either implies an external coordination, to merge their results and update their state, or requires that the modules themselves have such built-in ability. The first solution allows to better exploit off-the-shelf software solutions for the single biometrics. Many architectural choices are available in literature, to combine different biometric systems. For this work, we chose N-Cross Testing architecture with Supervisor Module (NCT-SM) [1], thanks to its extensibility. The N-Cross Testing (NCT) architecture, considers a multibiometric system as a multiagent one. Each agent is an autonomous and replaceable unit, with a possibly different weight on the final response. For instance, in the version of NCT implemented herein, for each recognition operation, the identity returned by each agent is associated with a reliability measure (we use System Response Reliability – SRR [3]), which expresses how much we can trust each single response. Information from all agents is suitably fused to formulate the global response (Figure 1).

When a new agent is to be integrated into an existing NCT architecture, its gallery is empty and must be populated from scratch through a suitable enrolling phase; moreover, agents galleries have to be updated from time to time to maintain the more representative templates. The approach we propose allows to avoid a new ad-hoc enrolling phase, and to exploit the responses of the other NCT systems instead, to “guide” the new agent in building its database until it can be considered as fully operational. We can consider our automatic labeling as a form of auto-training and unsupervised learning (since no operator intervention is required). We will also sketch how the approach can be extended to implement template updating in a multiagent system. Of course, our approach is not appropriate for a typical watch-list identification, where the aim is to exclude that the probe coincides with one of relatively few individuals (e.g. terrorists, or missing people). The probability of a positive identification in this case is too low to rely on it in order to populate the involved agents’ galleries. On the other hand, it is worth mentioning the typical category of situations where our approach is feasible. We consider a kind of closed set identification (re-identification) which requires to continuously check that the subjects within a protected area (a strategic chemical farm, or a flight deck) are all and only among those authorized to stay there. The difference with the classical closed set identification is the addition of an alarm if the returned identity is not enough similar to the probe at hand. Of course, it would be impractical to periodically perform such re-identification through explicit identity verification (with identity claim by each subject at hand). Several classifiers concur to determine the identity to label/update without human intervention, but rather borrowing from the compound system. In general, template updating is an aspect of biometric systems on which research is recently focusing, but “the state of the art related to template update is still in its infancy” [6]. NCT architecture and the labeling (template updating) algorithm have been tested according to some of the protocols suggested in [6]. We randomly selected the subjects to submit to the system and evaluated the performance variations against the number of updates.

## 2 The N-Cross Testing Protocol

### 2.1 Agent Reliability

Biometric agents might not be equally reliable, e.g. due to different performances of classifiers, or to intrinsic characteristics of biometrics at hand. Moreover, not all responses from the same agent are equally trustworthy, e.g. due to unstable input devices. Therefore, we adopt the System Response Reliability (SRR) index [3], a system/gallery dependent metric that can assess the ability of a biometric system to identify an enrolled subject, for each single probe. Two versions of it are introduced in [3], and they both exploit the full list of gallery subjects, ordered by similarity with the probe, which is returned by the identification module. In a few words, the version of SRR that we use for this work is computed considering how much “crowded” is the cloud of different individuals (each individual may correspond to more templates) which are returned in the similarity ordered list after the first one. The radius of the cloud is determined experimentally, and the more crowded the cloud, the less reliable the response (the higher the possibility of a wrong answer). Each agent  $T_k$  returns, for each of its responses  $s_{k,i}$ ,  $i=1, \dots$ , a *reliability measure*  $srr_{k,i}$  in the range  $[0, 1]$  (the higher, the better). SRR can even be computed when using an off-the-shelf biometric application, given that this returns an ordered list of (all) candidates in the gallery in response to each identification operation. See [3] for further details. Each  $T_k$  is characterized by an estimated threshold  $th_k$ , such that a response  $s_{k,i}$  is reliable only if  $srr_{k,i} \geq th_k$ . This threshold is different from the acceptance one, which would normally regulate acceptance or rejection based on a similarity/distance measure. In fact, we may obtain an acceptance response which is poorly reliable, as well as a rejection which is fully reliable. Therefore, the overall acceptance rule which results from using SRR requires that reliability of any response is controlled first. If this is sufficient, acceptance or rejection are decided according to a similarity/distance threshold. Otherwise, a repetition of the identification operation may be requested. Depending on the security requirements of the system, the reliability threshold  $th_k$  can start from an initial value of 0 (all responses are considered as reliable) and be updated over time to reach its more suitable value. Otherwise, a better initial value can be computed during a system tuning phase. In both cases, in the proposed system, the threshold of each agent  $T_k$  can be further updated over time, thanks to a supervisor module that coordinates the single biometric applications [1].

### 2.2 System Architecture

NCT architecture fits either multibiometric systems (each agent processes a different biometrics), or multimodal ones (all agents process the same biometrics, captured under different modalities), or multiexpert ones (the agents process the same biometrics, under the same modality, yet extracting different features). In the present case, we consider a multiexpert system on the face biometrics [1]. Each agent processes the same face image from which different kinds of templates are computed. Each template feeds the appropriate agent. In basic NCT,  $N$  agents  $T_k$ ,  $k=1, 2, \dots, N$ , mostly work in parallel, but exchange information at some points to reach the final result (cross operation) (Figure 1). Each agent has a database (gallery)  $G_k$  of biometric

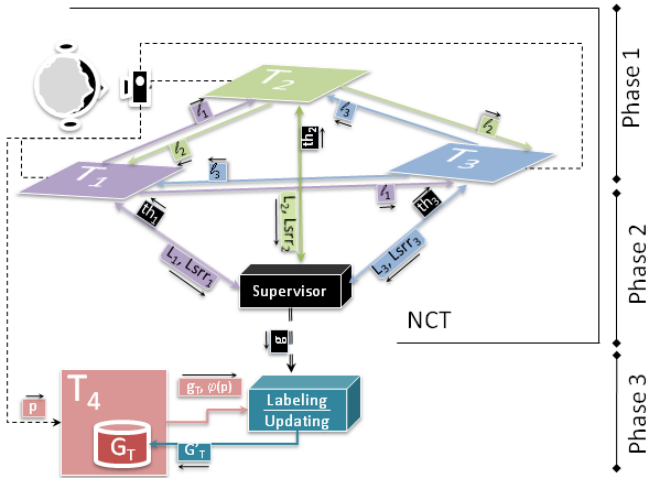
templates  $t_{k,j}$ ,  $j=1, 2, \dots, |G_k|$  which are labelled with the enrolled identities. The exchanged items are the respective lists of templates (galleries) ordered by similarity with the probe. In [1] different versions of the basic procedure are presented. For readers' convenience, we report here some details of the one adopted for this work, which exploits the SRR index discussed above. When the  $j$ -th probe  $p$  is submitted to the global system, each agent  $T_k$  independently extracts its template  $p_k$ , compares it with those in its specific gallery, and produces its ordered list of subjects. Each agent also computes a reliability measure  $srr_{k,j}$  for such response. Such value is used to determine how the agent participates to the second phase of the identification process (cross). As discussed in the previous section, each agent is characterized by a reliability threshold  $th_k$ : it sends its candidate list to the others only if its own SRR index for the current response is above such threshold. The sent list is truncated to the first  $M$  subjects, with  $M$  fixed in advance and experimentally determined. This step starts the cross phase, where each  $T_k$  first merges the lists from (reliable) companions (not including its own one, except when it is the only reliable system, as discussed below). Returned scores are suitably normalized in order to consistently merge information from different agents. Unreliable agents still work, even if they do not send their list, by receiving and merging lists from reliable companions. Each agent produces a merged list, with identities only coming from reliable ones. Unreliable agents will all produce the same list, i.e. the one obtained by merging all the reliable ones. Thus, in the special case when only one agent is reliable, all the unreliable companions return its list, and, since it does not receive any list from the others, it returns its own list as well. The length of each merged list depends on the amount of overlap, since shared subjects get a single average score. Finally, each  $T_k$  returns the first subject in its (resorted) merged list, its score and a variation of SRR index. The new reliability measure assigned to the first identity in each merged list  $L_k$ , is given by the average of reliability indexes of (reliable) agents in the set  $LS_k$  contributing to the list:

$$LSRR_k = \frac{1}{|LS_k|} \sum_{T_i \in LS_k} SRR_i \quad (1)$$

In this way the result returned by each  $T_k$  is influenced by the other agents. The decision phase produces the final response. In the present proposal, a *Supervisor Module (SM)* [1] (Figure 1) exploits single agent responses and their reliability to compute the final global response. It also evaluates the overall system state and updates its parameters. This takes to a self-tuning and co-evolution of agents.

The identity receiving the majority of votes is possibly returned as the input subject (see below). If more identities get the same number of votes, the basis for the final choice is not the score but the LSRR index. The probe is recognized if the final score got by the candidate identity (the mean of those from voting agents) complies with a threshold  $\delta$ . Our SM updates single reliability thresholds according to the behaviour of the companion agents. Agents agreeing on the correct answer are rewarded by a lower threshold (their responses will be more easily accepted by the system), while disagreeing agents are charged by increasing their thresholds (in future operations, they will need to score a higher reliability for their answers to be accepted).

The process converges to an optimal threshold configuration independently from the starting one. More details on experiments supporting such claim can be found in [1]. In the present proposal, as a variation of the behaviour presented in [1], SM module also drives the setting up of newly added agents, as shown below.



**Fig. 1.** Functional schema of the proposed system. In the first phase (first level) the existing agents  $T_1$ ,  $T_2$  e  $T_3$ , exchange their respective lists  $l_k$  and compute the merged list with global LSRR. In the second phase (second level) such information is sent to the Supervisor Module which computes the global response (identity  $g$ ), and updates and transmits the new SRR thresholds to all three agents  $T_1$ ,  $T_2$  e  $T_3$ . It also transmits the global response to the Labeling/Updating algorithm (third level), which uses it to update the gallery of  $T_4$ .

### 3 Labeling Algorithm

Due to its highly modular nature, a NCT system can be possibly expanded with new agents, given they are trained to handle the current working setting. This implies that they must undergo a training phase to set up their feature extraction process (consider for example PCA or LDA). Afterwards, their gallery must be populated according to a specific biometric trait of registered users. In practice, the gallery must be filled with templates, and these must be labeled according to enrolled individuals. In our case, also the computation of the reliability of the new agent must be initialized. In this section we discuss how to perform these two latter activities through an automatic process. It is worth noticing that the exploited algorithm can be applied either when a gallery is empty, and the new inserted templates must be labelled, or when, during normal operation, it contains elements to be updated. Notice that template substitution and/or deletion may occur either during the initial labelling phase, or during updating, as will be discussed later. If deletion during update is allowed, and the gallery only contains one template per identity, a situation may occur with missing information about one or more subjects. While this might bias the final result, we have methods in literature to address this problem [4]. The main strength of the algorithm is the use of

SRR returned values, which drive gallery labelling/updating. Given  $PR$  a set of probes,  $G$  a gallery of templates,  $I$  the set of registered identities,  $SRR\_values$  the reliability values in the interval  $[0, 1]$ ,  $SIM\_values$  the similarity values in the interval  $[0,1]$ , and  $SIM\_threshold$  the acceptance threshold based on similarity, we first define some functions:  $id:G \rightarrow I \cup NULL$  is such that  $id(t)$  is the identity corresponding to the template  $t$  ( $NULL$  if  $t$  is not labelled yet);  $temp:I \rightarrow P(G)$  is such that  $temp(x)$  is the subset of templates in  $G$  which correspond to identity  $x$ ;  $link\_id:G \times I \rightarrow P(G)$  is such that  $link\_id(t, x)$  inserts  $t$  to the set of templates corresponding to  $x$ , i.e. assigns the identity label  $x$  to  $t$ ;  $identify: PR \rightarrow G \times SRR\_values$  is such that  $identify(p)$  is its closest template in  $G$  with an associated reliability measure;  $sim: PR \times G \rightarrow \times SIM\_values$  is such that  $sim(p,t)$  is the similarity value between  $p$  and  $t$ . For simplicity, we assume from here on that a gallery contains one template per identity at any moment. We first give the pseudo-code of the labelling algorithm.

#### Labelling pseudocode for T

**Input:** the current gallery  $G$  and related identities  $I$ , the current probe  $p$  and a (reliable) retrieved identity (label)  $g$  from the global NCT system

```

if ( $g \notin I$ ) then
   $G = G \cup p$ ;  $I = I \cup g$ ;  $link\_id(p, g)$ ;
else
  ( $ct, srr$ ) =  $identify(p)$ 
  if ( $id(ct) \neq g$ ) OR ( $id(ct) = g$  AND  $sim(p, ct) < SIM\_threshold$ )
     $G = G \setminus ct$ ;  $I = I \setminus id(ct)$ ; /*discard template and
                                     associated identity*/
  else /* $id(ct) = g$  AND  $sim(p, ct) > SIM\_threshold$ */
     $G = G \setminus \{ct\} \cup \{p\}$ ; ( $ct1, srr1$ ) =  $identify(ct)$ ;
    if ( $id(ct1) \neq g$ ) OR ( $id(ct1) = g$  AND
                              $sim(ct, ct1) < SIM\_threshold$ )
       $G = G \setminus \{p\} \cup \{ct\}$ ;
    else if ( $id(ct1) = g$ ) AND ( $srr1 > srr$ ) NOP; /*gallery
                                                    is changed*/
  endif
endif
endif

```

The process works as follows. The new (not yet operative) agent  $T_k$  receives the current (reliable) retrieved identity (label)  $g$  from the global NCT system, and checks if a pertaining template is already present in its own gallery  $G_k$ . If not, the current probe template  $p_k$  is inserted, and associated (labeled) with the new identity. Otherwise, if  $g$  is already associated to some template in the gallery,  $T_k$  validates it. To this aim, it performs a recognition operation on  $p_k$  to obtain the closest template in  $G_k$  and a value  $srr_{k,p}$ . The obtained template  $ct_k$  will pertain to a certain identity  $g_k$ . We have two possible situations. If the identity  $g_k$  is different from the  $g$  previously returned by the fully operational global system, or even if its similarity with the probe

is below the acceptance threshold, both template  $ct_k$  and identity  $g_k$  will be deleted. In this way, we eliminate a source of possible recognition errors. Consider that the deleted identity and a pertaining template might be possibly re-added in a future step. If this does not happen, and if a specific problem arises for the missing identities (we experimentally found that they would be very few), it will be possible to force the insertion of good quality templates for them, in a significantly reduced mini-enrollment step. This is a little price to pay compared with the decreased performances due to “bad” templates. Alternatively, if  $g$  and  $g_k$  are the same,  $T_k$  starts a second (cross) recognition step. It temporarily changes the roles of probe and gallery templates: the new recognition operation is performed using  $ct_k$  as probe, and the gallery  $G'_k = G_k - \{ct_k\} \cup \{p_k\}$ . A new identity  $h_k$  will be retrieved, and a value  $srr_{k,ct}$  will be associated to the new response. If the two recognition operations retrieved different identities, i.e.  $h_k \neq g_k$ , or even if the second recognition falls below the acceptance threshold, the original gallery  $G_k$  is restored by re-inserting  $ct_k$  and deleting  $p_k$ . This is because the original template fostered a correct identification, while the new one caused a recognition error, since it better matches a different (wrong) identity. Otherwise, if  $h_k = g_k$ , and  $srr_{k,ct} > srr_{k,p}$ ,  $G'_k$  becomes the new gallery for  $T_k$ , i.e. the just acquired template  $p_k$  replaces the former one for the identity  $g_k$ , since it fosters a better discrimination from the other identities in the gallery (a better value of  $SRR$  is obtained when  $p_k$  is a gallery template than when it is used as a probe). The reliability threshold for the new agent will start from 0 (all responses acceptable) and will be updated at any recognition operation by the SM, in the same way used for the fully operating agents (see previous section). The agent can start full cooperation with the others when its gallery becomes substantially stable during a sufficient time elapse.

The extension to  $l > 1$  gallery templates per identity is straightforward. Moreover, a template updating process can be implemented by each fully operative agent if it periodically executes a similar algorithm, possibly except for the deletion of identities in case of discordant results.

## 4 Experimental Results

The architecture that we set up to evaluate the proposed algorithm included four classifiers, each implementing a different technique for face biometry, and a supervisor module. Three out of the four agents ( $T_1$ ,  $T_2$ ,  $T_3$ ) were already working (fully labelled gallery and determined SRR thresholds) and respectively exploited Linear Discriminant Analysis (LDA), Orthogonal Locality Preserving Projections (OLPP) and Neighborhood Preserving Embedding (NPE) (see 1). The fourth agent  $T_4$ , which was added later and which gallery had to be labelled, was based on Partitioned Iterated Function Systems (PIFS) [2]. The sets of face images making up the testing benchmark were extracted from AR Faces database [5], since a good number of subjects (70 men and 56 women) with a sufficiently varied set of distortions is represented in such database. Each subject appears in two different sessions with 13 image sets each. Sets differ in expression (1 neutral, 2 smile, 3 anger, 4 scream), illumination (5 left light, 6 right light, 7 all side light), presence/absence of occlusions (8 sun glasses, 11 scarf), or combinations (9 sun glasses and left light, 10 sun glasses

and right light, 12 scarf and left light, 13 scarf and right light). Sets 14 to 26 of the second session present the same conditions at a different time. Performances were measured in terms of Recognition Rate (RR), Equal Error Rate (EER), Percentage of Templates into the Gallery (PTG) defined as  $|G_T|/|G|$ , where  $G_T$  is the gallery of a single system and  $G$  is the ideal gallery including the full set of enrolled identities (PTG=1 for fully operational agents). To show how the performances of agent  $T_4$  improved as its gallery was gradually populated/updated, it was evaluated separately from the others at regular intervals. Every 20 operations (insert, delete, substitute), the values for RR, EER, and PTG were computed for  $T_4$  using set 6 as probe, with its current gallery. Set 1 was the gallery for  $T_1, T_2, T_3$  in all experiments; the probe set depended on the experiment at hand, and probes were randomly extracted from it. An experiment ended when performances of  $T_4$  in that setting became stable.

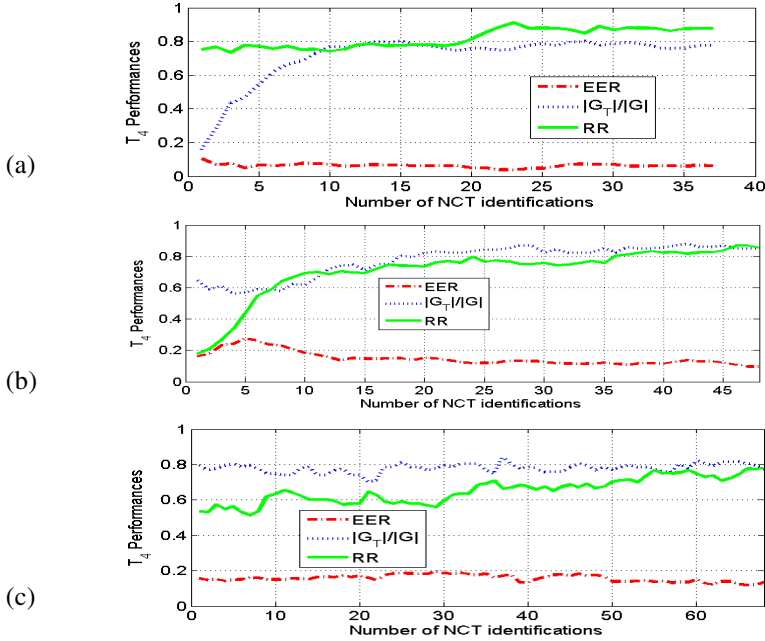
In the first experiment, the initial gallery of  $T_4$  was empty, and the probe set for all four agents was the union of sets 2 and 3. Figure 2(a) always shows  $PTG < 1$ , i.e.  $|G_T| < |G|$ , with  $|G_T|$  referring to  $T_4$ , because some responses from the NCT system ( $T_1, T_2$ , and  $T_3$ ) were marked as unreliable, and therefore the corresponding identities were not inputted to the algorithm for  $T_4$ . If we relaxed this policy, more identities might have entered the gallery in less time, but we would have had a more error-prone behaviour from the new agent. Since the gallery for  $T_4$  was initially empty, we can notice a faster convergence towards the final EER and RR. This is because many identities were created directly in the initial phase, so limiting the number of subsequent substitutions. The second experiment aimed at testing template updating in gradually degrading settings. In order to stress the system even more, we allowed template deletion. This implies that we can have  $PTG < 1$  even for a fully operational agent. We took again  $T_4$  as the agent to be affected, but updating might have involved all agents at the same time. Our choice is motivated by the higher readability of results. We introduced a strong element of difficulty, since the initial gallery of  $T_4$  included images with occlusions from set 8, while the probe set for all agents was the union of sets 2 and 3. This implied a higher number of substitutions or deletions followed by re-insertions, which increased the time needed to reach a stable configuration (Figure 2(b)) with respect to the previous experiment. We also noticed that more than 90% images with sun glasses (the initial gallery) were substituted, and this agrees with the better identifiability of images from sets 2 and 3 (smiling and angry) with respect to occluded images. The last experiment was the worst case (Figure 2(c)): we had a pre-existing gallery (set 6), while the probe also included images from set 8. Nevertheless, the algorithm showed the ability to select, among the different probes, the useful ones to be substituted in the gallery for  $T_4$ , so obtaining a substantial improvement of performances. The value for PTG is lower than in the preceding experiments, due to the higher number of responses labelled as unreliable by the NCT system composed by  $T_1, T_2$ , and  $T_3$ .

Besides accuracy, we also observed the frequency of the different operations (insertion, substitution, deletion) during gallery creation. The algorithm alternated phases with a significant amount of creations of new templates, and phases with more frequent updates. In the first case the gallery was being populated, so that there was an increase of  $G_T/G$ , but this also corresponded to an increase of EER and to a reduction of RR (i.e., to a temporary reduction in accuracy). However, during the second kind



of phase, though preserving a good  $G_T/G$ , EER decreased and RR increased (performances improved). On the other hand, deletions were more infrequent. When the agent reached stability, the number of operations became negligible.

Figure 2(c) shows that the system converges more slowly in a critical situation, however the reached state is satisfactory enough, since, using set 6 as probe, it presents an RR of 80% compared with the 87% value in the optimal case (set 1, i.e. neutral, as gallery and set 6 as probe 2).



**Fig. 2.** Performance of the algorithm with different starting galleries: a) empty; b) images with occlusions; c) images with right light and occluded probes

## 5 Conclusions

We presented an algorithm for unsupervised template labeling for an agent, when it enters a fully operational multiagent architecture. We showed its ability to select significant templates even in case of template updating, and in non optimal conditions.

## References

1. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: A multiexpert collaborative biometric system for people identification. *JVLC* 20(2), 91–100 (2009)
2. De Marsico, M., Nappi, M., Riccio, D.: FARO: Face Recognition Against Occlusions and Expression Variations. *IEEE Trans. On Systems, Man and Cybernetics–Part A* 40(1), 121–132 (2010)

3. De Marsico, M., Nappi, M., Riccio, D., Tortora, G.: NABS: Novel Approaches for Biometric Systems. Accepted for Publication. In: IEEE Trans. on Systems, Man, and Cybernetics—Part C (available online)
4. Fatukasi, O., Kittler, J., Poh, N.: Estimation of Missing Values in Multimodal Biometric Fusion. In: 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, BTAS 2008, Arlington, VA, pp. 1–6 (2008)
5. Martinez, A.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Trans. on PAMI 24(6), 748–763 (2002)
6. Rattani, A., Freni, B., Marcialis, G.L., Roli, F.: Template Update Methods in Adaptive Biometric Systems: A Critical Review. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 847–856. Springer, Heidelberg (2009)
7. Roli, F., Didaci, L., Marcialis, G.: Adaptive Biometric Systems That Can Improve with Use. In: Ratha, N.K., Govindaraju, V. (eds.) Advances in Biometrics - Sensors, Algorithms and Systems. Springer, Heidelberg (2008)
8. Uludag, U., Ross, A., Jain, A.: Biometric template selection and update: a case study in fingerprints. Pattern Recognition 37, 1533–1542 (2004)