An *a-contrario* Biometric Fusion Approach

Luis Di Martino[1,2] Javier Preciozzi[1,2] Rafael Grompone von Gioi[3]
Guillermo Garella[1,2] Alicia Fernández[1] Federico Lecumberry[1]

[1] IIE, Universidad de la República, Uruguay
{dimartino,jprecio,ggarella,alicia,fefo}@fing.edu.uy
[2] Digital Sense, Uruguay
[3] Centre Borelli, ENS Paris-Saclay, Université Paris-Saclay, France
grompone@ens-paris-saclay.fr

Abstract

Fusion is a key component in many biometric systems: it is one of the most widely used techniques to improve their accuracy. Each time we need to combine the output of systems that use different biometric traits, or different samples of the same biometric trait, or even different algorithms, we need to define a fusion strategy. Independently of the fusion method used, there is always a decision step, in which it is decided if the traits being compared correspond to the same individual or not. In this work, we present a statistical decision criterion based on the a-contrario framework, which has already proven to be useful in biometric applications. The proposed method and its theoretical background is described in detail, and its application to biometric fusion is illustrated with simulated and real data.

1. Introduction

Biometrics has achieved high popularity in the last decade as its use was extended from its typical crime related scenario to a whole new spectrum of applications. It has been used in the health domain in order to efficiently deliver vaccination campaigns [13], in entertainment, security of personal devices and human-computer interaction systems [12]. Additionally, security-related applications have also been on the rise. Biometric systems are being used for automatic checkpoints at countries borders and admission control at sports venues among others. These applications demand constant improvement in accuracy and robustness in order to fulfill their requirements.

Biometrics can greatly benefit from the fusion of multiple systems [26, 27]. In [17], the basis and formalization of fusion strategies for biometric applications were introduced, becoming the reference on fusion for the pattern recognition community. The fusion schemes presented there are widely used because of its simplicity, ease of implementation and because they do not require any training process. Several articles and technical reports validate the presented fusion approaches [29, 11, 16, 28]. Other biometric fusion approaches make use of trained statistical models [30, 22]. They provide better performance than the simple rules introduced in [17], but with the extra cost of the training process and parameters selection (which is not always easy).

In this work, we present a novel statistical decision criteria for biometric fusion based on *a-contrario* framework [3]. The *a-contrario* framework was already used in the context of single biometric trait evaluation: it was used for face recognition verification [6, 5], fingerprint identification [4] and iris verification [21]. The aim of this work is to extend the use of the *a-contrario* framework as a decision criteria for multi-trait systems by providing a general biometric fusion approach.

2. The *a-contrario* approach

The *a-contrario* theory [2, 3] is a statistical framework used to set detection thresholds automatically, in order to control the number of false detections. It is based on the non-accidentalness principle [31, 19] which informally states that there should be no detection in noise. In the words of D. Lowe, “we need to determine the probability that each relation in the image could have arisen by accident, $P(a)$. Naturally, the smaller that this value is, the more likely the relation is to have a causal interpretation” [19, p. 39].

A stochastic background model $\mathcal{H}_0$ needs to be defined, where the structure of interest is not present and can only arise as an accidental arrangement. We also need to define a family of events of interest T. A statistic $k(\cdot)$ is to be evaluated on each considered test $\mathbf{e} \in T$. We are interested in events with $k(\mathbf{e}) \leq \mathbf{k}$ for a predefined threshold $\mathbf{k}$. Accordingly, we need to evaluate the probability $P(k(\mathbf{e}) \leq \mathbf{k} | \mathcal{H}_0)$.

When this probability is small enough, there exists evidence to reject the null hypothesis and declare the event meaningful. However, one needs to consider that usually multiple events are tested. If 100 tests were performed, for example, it would not be surprising to observe an event that appears with probability 0.01 under random conditions. Thus, the number of tests N_T needs to be included as a correction term, as it is done in the statistical multiple hypothesis testing framework [9].

Following the *a contrario* methodology [2, 3], we define the *Number of False Alarms* (NFA) of an event $\mathbf{e}$ as:

$$\mathrm{NFA}(\mathbf{e}) = N_T \cdot P(k(\mathbf{e}) \leq \mathbf{k} | \mathcal{H}_0). \tag{1}$$

The smaller the NFA value, the more unlikely the event $\mathbf{e}$ is to be observed by chance in the background model $\mathcal{H}_0$; thus, the more meaningful. One can show [2, 3] that under $\mathcal{H}_0$, the expected number of events with NFA $\leq \varepsilon$ is bounded by ε. As a result, ε corresponds to the mean number of false detections in $\mathcal{H}_0$. When $\mathrm{NFA}(\mathbf{e}) \leq \varepsilon$, for a predefined ε value, $\mathcal{H}_0$ is rejected as an explanation for the event $\mathbf{e}$, and an alternative hypothesis $\mathcal{H}_1$ is accepted. Note, however, that a single stochastic model $\mathcal{H}_0$ is involved and no *stochastic* model is required for $\mathcal{H}_1$.

In contrast, in the classic hypothesis testing framework, explicit stochastic models are required for both, $\mathcal{H}_0$ and $\mathcal{H}_1$. Two possible errors can be made:

- *Non-detection:* it occurs when $\mathcal{H}_1$ is rejected for an observation $\mathbf{e}$ for which $\mathcal{H}_1$ is true. Formally, the probability of a non-detection is $P\left(k(\mathbf{e}) > \mathbf{k} | \mathcal{H}_1\right)$;

- *False alarm:* it takes place when $\mathcal{H}_1$ is accepted despite being false for the particular realization $\mathbf{e}$. The probability of a false alarm is $P\left(k(\mathbf{e}) \leq \mathbf{k} | \mathcal{H}_0\right)$.

$P\left(\cdot | \mathcal{H}_0\right)$ and $P\left(\cdot | \mathcal{H}_1\right)$ are determined by the probability distributions of $k(\mathbf{e})$ under the hypothesis $\mathcal{H}_0$ and $\mathcal{H}_1$. When both distributions are known, $\mathbf{k}$ can be fixed using classic methods such as the *Likelihood Ratio* or a *Bayesian Test*. Nevertheless, in some scenarios the structure of interest is hard to model, or the number of samples is small, resulting in a poor estimation of $\mathcal{H}_1$. In such cases, the *a-contrario* framework provides a useful alternative.

3. The *a-contrario* model for multibiometrics

When setting an *a-contrario* model, one needs to specify the family of tests, the statistic to be evaluated, and the background model $\mathcal{H}_0$. In our case, a test corresponds to the comparison of two biometric samples q_i and g_j. The result of the comparison is a vector $\mathbf{D}_{i,j}$ containing several results from every biometric system being considered. If we consider K different systems, then $\mathbf{D}_{i,j}$ is defined as follows:

$$\mathbf{D}_{i,j} = \mathbf{D}(q_i, g_j) = \left(d_{i,j}^{(1)}, \ldots, d_{i,j}^{(k)}, \ldots, d_{i,j}^{(K)}\right)$$

where $d_{i,j}^{(k)}$ represents the distance obtained between samples q_i and g_j in the biometric system k. Given a realization $\mathbf{D}_{i,j}$, the NFA is computed as follows:

$$\mathrm{NFA}(\mathbf{D}_{i,j}) = N_T \cdot P(d(\mathbf{D}_{i,j}) \leq d^* | \mathcal{H}_0) \tag{2}$$

where the term $P(d(\mathbf{D}_{i,j}) \leq d^* | \mathcal{H}_0)$ accounts for the probability of the particular observation under the background model. Finally, the event should be accepted or rejected by applying a threshold on the NFA. In the multidimensional case, there are several ways to compute the statistic or distance $d(\mathbf{D}_{i,j})$. This is a well known problem in the fusion of multiple pattern recognition systems. In [17], several options are presented and analyzed:

1. The minimum distance: $d = \min\left(d_{i,j}^{(1)}, \ldots, d_{i,j}^{(K)}\right)$;
2. The maximum distance: $d = \max\left(d_{i,j}^{(1)}, \ldots, d_{i,j}^{(K)}\right)$;
3. The product of the distances: $d = \prod_{k=1}^{K} d_{i,j}^{(k)}$;
4. The sum of the distances: $d = \sum_{k=1}^{K} d_{i,j}^{(k)}$.

As the value d is compared against a threshold d^* in the decision, the different strategies defined above will impose different criteria in the classification. Each of these options is equivalent to a function $\mathcal{F}\left(d^{(1)}, \ldots, d^{(K)}\right) : \mathbb{R}^K \to \mathbb{R}$ in which the fusion rule is characterized by all the configurations in the fusion space $\mathbb{R}^K$ that produces the same distance. From an operational point of view, the probability of occurrence of the realization being evaluated in the background model could be computed by integrating a probability distribution function that represents this model. In the one-dimensional case this integration is done simply by considering the interval of distances up to the value being evaluated. In a multidimensional setup, this integration is done over a domain $\Omega_{d^*} = \{x \in \mathbb{R}^K \,|\, \mathcal{F}(x) \leq d^*\}$ containing all possible configurations that produce the same d^* according to the criterion used. This gives place to the following expression:

$$P\big(d(\mathbf{D}_{i,j}) \leq d^* \big| \mathcal{H}_0\big) = \int_{\Omega_{d^*}} p_{\mathcal{H}_0}(\mathbf{x}) d\mathbf{x}. \tag{3}$$

We will illustrate the methodology by the simplest case of a two dimensional fusion scenario in which the obtained vector is $\mathbf{D}_{i,j} = \left(d_{i,j}^{(1)}, d_{i,j}^{(2)}\right)$. As an example, we will use the sum rule as the fusion criteria (a popular choice [17]). Figure 1 shows in red the level lines of $\mathcal{F}$ and the integration domain Ω_{d^*} (in blue) for a particular value of d^*.

4. The background model

A key element of the *a-contrario* framework is how to define the background model $\mathcal{H}_0$. Because the different alternatives to compute $\mathcal{H}_0$ depend on the data, let's define the usual *Query* and *Gallery* datasets:

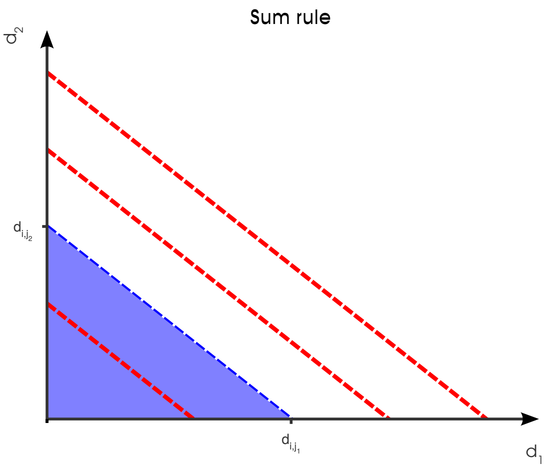


Figure 1: Integration domain Ω_{d^*} for the sum rule.

- *Gallery* dataset *G* of size N, containing the samples $g_1, \ldots, g_N$ of the system stored IDs;
- *Query* dataset *Q* of size N with the corresponding samples $q_1, \ldots, q_N$ of the same IDs (different samples of the same IDs).

4.1. Using only Gallery samples

The first option we analyze is *not to use the Query samples* at all for the background model estimation. One advantage of this option is that it allows to have a *pre-computed* model *before* knowing any Query sample. In order to obtain such a model, we compute the distances between Gallery samples in an "all versus all" manner, obtaining a matrix $D_{G,G}$ of distances, where each element $D_{G,G}\left(i,j\right) = d_{i,j}^{G,G}$ represents the distance between gallery samples i and j. This matrix has two particular features: a) it is symmetric ($d_{i,j}^{G,G} = d_{j,i}^{G,G}$): a consequence of the fact that distances $D_{G,G}(i,j)$ and $D_{G,G}(j,i)$ are equal as they are obtained using the same samples g_i and g_j; and b) all elements in the main diagonal are zero ($d_{i,i}^{G,G} = 0$): they correspond to the comparison of one sample with itself. Therefore, there are $\frac{N \times (N-1)}{2}$ useful comparisons done between samples corresponding to different IDs. They are representatives of the impostors class we are trying to model and can be used as input for the estimation of the background model. (Recall that in our problem $\mathcal{H}_0$ models the distance distribution for samples of different IDs.)

The use of this information for modeling $\mathcal{H}_0$ presents some particular advantages and drawbacks. As a benefit, it is only based on the known Gallery samples and it could be computed beforehand without actually doing any verification test. As a drawback, the obtained model may suffer a lack of precision on the production environment if significant difference exists between the features of the Gallery and Query samples (for example, if there is a considerable technological change between the acquisition process of the Gallery and Query samples).

Using the available information in $D_{G,G}$ matrix, two different approaches can be designed: a *general model*, where one $\mathcal{H}_0$ is obtained for the entire Gallery dataset or a *particular model*, where one $\mathcal{H}_0$ is obtained for each sample in the Gallery dataset.

4.1.1 General model

In this approach, all the Gallery samples are used to obtain a unique general model $\mathcal{H}_0$. In this case, the dataset used to compute the model has $\frac{N \times (N-1)}{2}$ samples of different IDs. This approach has the advantage that we only have one model for all the Gallery samples, but this is also its main drawback: in the generalization process, the model could miss the particularities that make an ID different from each other. Indeed, this is a well known problem: given a biometric trait, some people are more difficult to classify than others (Doddington Zoo) [7].

4.1.2 Particular model

An alternative to the previous strategy is to build a particular model $\mathcal{H}_0$ for each identity in the Gallery. With this approach, we will have N different background models: one for each identity. In this case, for each identity in the Gallery, there are $N-1$ samples of different individuals that allows to model how the particular person's biometric features differ from those of other people in the gallery dataset. Thus, in this case, the number of useful samples for each model is $N-1$.

4.2. Background model computed at runtime

As was already stated, the pre-computed model estimation strategy does not take into account the Query samples, leading to a potentially weaker classifier since the variability of the Query samples are not considered. To tackle this problem, we can estimate the model during verification. The strategy is as follows: given an input query sample q_i, it is compared against all the Gallery samples g_j, obtaining distances $d_{i,j}^{Q,G}$. If we compute the distance between all the query samples against all the gallery ones, we can obtain a new matrix of distances $D_{Q,G}$. Each element $D_{Q,G}\left(i,j\right) = d_{i,j}^{Q,G}$ represents the distance obtained when input query q_i is compared against gallery sample g_j. There are two important differences between this matrix and $D_{G,G}$ used in the *pre-computed* model introduced before. First, in this case the diagonal elements are not zero as they correspond to the comparison between two different samples of the same ID. This distance should be very small in relation to other ones (at least this is what we expect) but not zero. Second, the matrix is not symmetric anymore. This happens because the comparison between q_i and g_j samples is not equal to the comparison between q_j and g_i. Both comparisons being done involve the same pairs of IDs i and j but different associated biometric samples in each case.

Finally, there is a key difference in how the null-hypothesis dataset is built when compared with the *pre-computed* case. When the model is computed in verification time, one does not know beforehand which particular comparisons correspond to the impostor class (null hypothesis). Therefore, the only option is to compute the background model using all the distances with the exception of the one being evaluated (thus using $N-1$ samples). This will allow to asses if the result being analyzed is rare to occur under the background model. From the operational point of view, if the number of available samples in the Gallery is large, the particular distance evaluated could also be included to model the null hypothesis without changing much the numerical estimation.

4.3. Summary

The above strategies are all valid from a theoretical point of view; in practice, however, there are differences in the setting. Table 1 provides a general view. The accuracy of each strategy depends on how similar the gallery and query samples are, the robustness of the biometric system being used, etc. In this work, we are not including a complete analysis of the different strategies. In what follows, all the experiments were done using the second approach defined in section 4.1.2.

Computed	Type	Size	Use Query dataset
Offline	general	$\frac{N\times(N-1)}{2}$	no
Offline	particular	$N-1$	no
Online	particular	$N-1$	yes

Table 1: Model $\mathcal{H}_0$ computing strategies.

5. Simulated example

Our first experiment is performed on simulated data. This allow us to analyse, in a simple and controlled manner, how the proposed method performs. Both in this section and in the following, we report the accuracy of the different approaches using the usual *False Match Rate* and *False Non-Match Rate* as defined in the standard ISO/IEC 19795-2:2007 [15].

5.1. Data generation

In order to keep the experiment as simple as possible, we simulate a 2-classifiers fusion scheme. Both classes scores distributions are assumed to follow Gaussian distributions; thus the probability densities for the impostors and genuine are:

$$p_{\mathcal{H}_0}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma_{\mathcal{H}_0}|}} e^{-\frac{1}{2}\left(\mathbf{x}-\mu_{\mathcal{H}_0}\right)^T \Sigma_{\mathcal{H}_0}^{-1}\left(\mathbf{x}-\mu_{\mathcal{H}_0}\right)}, \quad (4)$$

$$p_{\mathcal{H}_1}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma_{\mathcal{H}_1}|}} e^{-\frac{1}{2}\left(\mathbf{x}-\mu_{\mathcal{H}_1}\right)^T \Sigma_{\mathcal{H}_1}^{-1}\left(\mathbf{x}-\mu_{\mathcal{H}_1}\right)}. \quad (5)$$

The values for the mean and co-variance matrix have been set arbitrary for the impostors and genuine classes. A representative set of the samples generated with such distributions is represented in Figure 2.

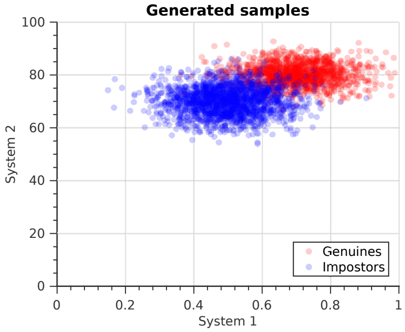


Figure 2: Generated data samples

5.2. Experimental evaluation

Using the previously defined probability distributions for both genuines and impostors distances, we can simulate a real scenario. For that, we consider a setting with N samples in the Gallery dataset and N samples in the Query dataset, with $N = 3000$ identities. In this case, the number of distances for the genuines pairs is N and the number of impostors pairs is $N(N-1)$.

In order to use *Likelihood-Ratio* as a fusion strategy, we need to train both genuines and impostors score distributions. As explained in [22], the *Likelihood-Ratio* approach ensures the best possible fusion performance when *both* genuines and impostors distributions $p_{\mathcal{H}_0}(\mathbf{x})$ and $p_{\mathcal{H}_1}(\mathbf{x})$ are known. In practice though, we only have estimations $\widetilde{p_{\mathcal{H}_0}}(\mathbf{x})$ and $\widetilde{p_{\mathcal{H}_1}}(\mathbf{x})$ for them, and the performance of the *Likelihood-Ratio* test will depend on the accuracy of such estimations. On the other hand, the *a-contrario* approach only requires the information provided by the impostors' distribution. Since the number of samples for genuines and impostors is unbalanced (N versus $N(N-1)$), the estimation of $p_{\mathcal{H}_0}(\mathbf{x})$ and $p_{\mathcal{H}_1}(\mathbf{x})$ are not equally accurate. In some cases, this results in an advantage in using *a-contrario* models for biometric applications.

In order to simulate the possible lack of genuine representatives, we run different tests, reducing the number of samples using to estimate the distributions. The following trained genuines distributions parameters were experimentally obtained for Sample Ratios (SR) values of 0.1 and 1. The distributions were obtained using a *Gaussian Mixture Models* (GMM), in particular the implementation in [8].

Training results for $SR = 1$ of genuine training samples When using the full dataset, the number of samples is large enough to obtain good estimations for both distributions, see Figure 3. Each estimation corresponds to a single Gaussian and the means and covariance are similar to the original ones.

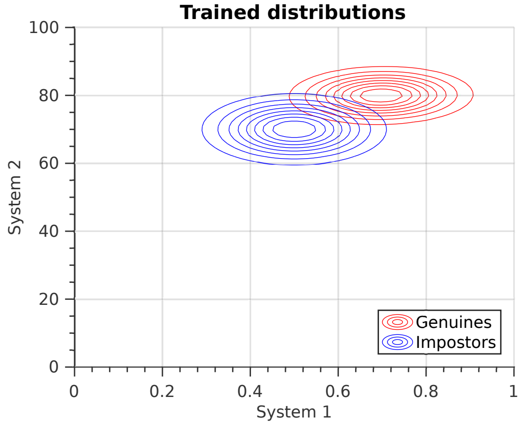


Figure 3: Impostors and genuine estimated distributions for a Sample Rate of 1

Training results for $SR = 0.1$ of genuine training samples When sampling only ten percent of the dataset, the situation is different, see Figure 4. The impostors distribution is well estimated, as before. But the genuines distribution estimation consists of the mixture of four Gaussians. The lack of samples resulted in a poor estimation of the genuine distribution.

The models trained in each case were used with both the *Likelihood-Ratio* and *a-contrario* approaches on the testing data partition. The obtained results are shown in Figures 5 and 6 for $SR = 1$ and $SR = 0.1$ respectively.

In the first case ($SR = 1$), it can be seen that the *Likelihood-Ratio* approach achieves a better performance than the obtained by each system individually and the *a-contrario* based fusion. Additionally, the behavior of the technique based on the trained probability densities is the same as the one obtained with the ground-truth distribution.

In the second case ($SR = 0.1$), the sample ratio is smaller and the accuracy of the probability densities lower, resulting in worse performances with the *Likelihood-Ratio* strategy. On the other hand, the *a-contrario* approach continues to work equally good as before. These results confirm the observations made by the authors in [22] and show the robustness of the proposed approach with respect to genuine's class available training data.

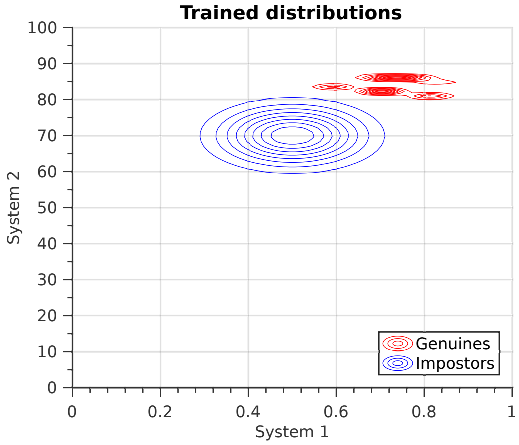


Figure 4: Impostors and genuine estimated distributions for sample rate of 0.1

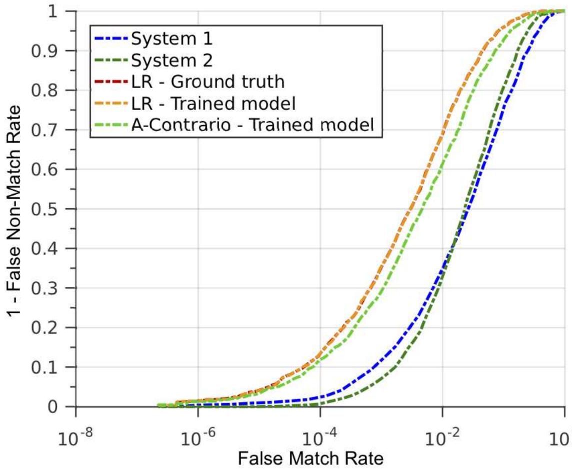


Figure 5: Verification fusion performance for $SR = 1$.

6. Experiments on real data

We have done two set of experiments on real data. We start with BSSR1 [23], a widely used dataset to test biometric fusion methods. In the second set of experiments, we focus our attention to the fusion of several samples of the same modality: in this case, image faces.

6.1. BSSR1 dataset

BSSR1 is a multimodal dataset by the National Institute of Standards and Technology (NIST), composed of scores values for different biometric modalities and algorithms. The dataset has three partitions: Face vs Face, Fingerprint

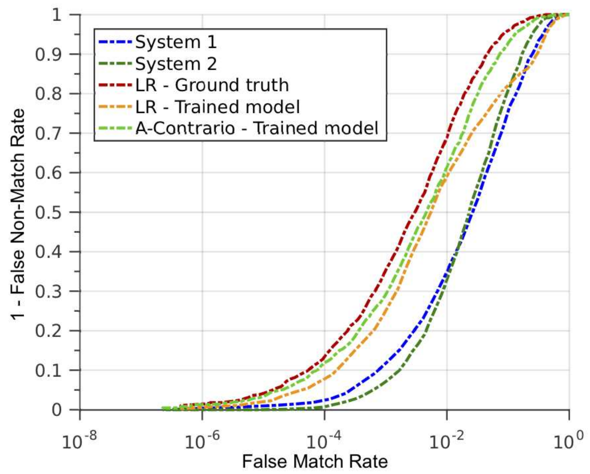


Figure 6: Verification fusion performance for $SR = 0.1$

vs Fingerprint and Face vs Fingerprint. Since we were mainly interested on face recognition, we have selected the Face vs Face partition. This partition includes the results obtained from two different face recognition systems aiming to study a multi-algorithm fusion over the same biometric modality. The data was collected from 3000 subjects retrieving 3 facial images of each. The first image of each triplet was taken as reference. The comparison against the second and third images was saved in different datasets, *First Set* and *Second Set* respectively, and two different systems were used to obtain the corresponding scores.

In this work we follow the same procedure used in [22]: the available data is used in a *2-fold cross-validation* scheme. The partition of the data is done by taking random samples for both the training and testing sets, for each experiment. To ensure that no particular partition favours one strategy over the other, the *cross-validation* experiment is repeated M times. The partition was performed, of course, into groups of different IDs (and not at individual image level) to prevent bias. For each sub-partition of the selected database and each fusion strategy, one ends up having a matrix of $FMR\,(\tau)$, where each column vector $FMR(\tau)_{m,k}$ represents the obtained results for a particular experiment m and fold k. In this context, τ represents the threshold that fixes a particular working point of the system being used. The threshold value would depend on the particular strategy being evaluated. It would be applied over the *NFA* for the *a-contrario* approach, the η for the *likelihood-ratio* strategy and, finally, over the scores when each system is working individually.

Following the experimental setup in [22], we perform 20 experiments in a 2-fold cross validation scenario, having then $M = 20$ and $K = 2$. The obtained results are summarized statistically by reporting the mean genuine accept rate $\overline{FMR}\,(\tau)$ and its 95% confidence interval $[FMR_l\,(\tau)\,, FMR_u\,(\tau)]$. In order to obtain these, the FMR metrics should be referred to a common set of FMR values.

As explained in [22], the *Likelihood-Ratio* approach is highly dependent in having an accurate estimation of the underlying impostors and genuines classes distributions:

> "However, this optimality of the likelihood ratio test is guaranteed only when the underlying densities are known. In practice, we estimate the densities $f_{gen}(x)$ and $f_{imp}(x)$ from the training set of genuine and impostor match scores, respectively, and the performance of likelihood ratio test will depend on the accuracy of these estimates." [22]

This statement is of great importance since it remarks the biggest practical difficulty in the *Likelihood-Ratio* framework: While it is the more powerful statistical test (this is assured by the Neyman-Pearson theorem), this requires good knowledge of the genuines class. This dependency is very important and can be summarized in three points:

- **Classes unbalanced:** In a typical scenario where pairs of biometric samples from a population of size N is used, just N comparisons correspond to the genuine class, whereas $N \times (N-1)$ comparisons in the impostors category are available.

- **Few samples per person:** Although in some particular databases there are multiple samples per each person (e.g. *Faces in the Wild* [10]), this is not always the case. This condition is even worse when considering citizen databases in which usually just a few samples per person exist.

- **Intra-class variations:** The biometric samples belonging to a particular person in a database could present large variations due to different factors. For example, pose or illumination variation as well as aging could be present between two face images. Or different sensors could be used between two successive fingerprint samples. Depending on the robustness of the particular biometric system being evaluated, these differences may give place to big *intra-class* variations. Such variations could make the estimation of the genuines class inaccurate.

The proposed *a-contrario* approach only depends on an accurate characterization of the impostors class, for which these issues are not present. Therefore, our goal in the experimental evaluation is to compare both classification strategies and in particular evaluate the robustness of the *Likelihood-Ratio* framework when the genuines distribution is not very accurate. In order to simulate this situation, we define a Sample Rate, to obtain a random set of test

SR	BSSR1-Face
0.01	15
0.05	75
0.1	150
0.3	450
0.7	1050
1	1500

Table 2: Genuines training samples for different Sample Ratio

samples from which the genuines distribution is computed. We then test both the *a-contrario* and *Likelihood-Ratio* fusion approaches by varying the genuines sample ratio SR. The used sample ratio values and its corresponding amount of genuines training samples for each database in a 2-fold cross-validation scheme is shown in Table 2. The results obtained for the BSSR1-Face dataset, for the different sample ratios used to train the genuines distribution are shown in Figure 7.

6.2. Face recognition example

One may wonder if fusion is indeed a required functionality for a biometric system: as algorithms continue to improve, the need to fuse results may seems to be marginal. We will show that this is not the case with a face recognition example. One of the best face recognition methods is ArcFace [1], performing an astonishing 99.83% verification performance on the well-known LFW dataset [10] and protocol [18], outperforming most of the existing methods so far. Nevertheless, when we look at more challenging situations like aging, this performance drops: 95.56%: on CALFW [32] and 95.15% on AgeDB [20], two dataset where aging is a major characteristic. One of the ways to improve these results is to consider more than one pair sample: assuming that we have more than one image for each identity, then we can match a query image with each one of the images in the gallery and fuse the score results for each identity. For this test, we have used AgeDB, from where we have selected those identities with at least three images. We have 566 of those identities. We then perform the same 2-fold partition, leading to a test and train dataset of 283 elements.

Figure 8 shows the result of the experiment using two images per identity in the gallery. System 1 and System 2 correspond to the result of ArcFace using each one of the images in the dataset. The first interesting result is that even in this case where the Face Recognition system performs very well, fusion improves the results even further: both *Likelihood-Ratio* and *a-contrario* methods increase the accuracy of the system. Secondly, although there is no significant difference between both methods with a SR of 1, the *a-contrario* approach clearly outperforms *Likelihood-Ratio* strategy for a sample ratio of 0.1.

7. Conclusions

A novel decision criterion for biometric fusion was introduced, based on the *a-contrario* approach. Our experiments show that the proposed method is comparable in terms of accuracy to *Likelihood-Ratio* method when the distributions are well sampled. More importantly, the *a-contrario* approach outperforms the *Likelihood-Ratio* method when few samples are used to estimate the genuines distribution.

The *a-contrario* approach has other advantages over the *Likelihood-Ratio* strategy. First, there is no need to estimate the genuines distributions as it works only with the impostors distribution. This is an important point for practical considerations: it is not always possible to have genuine pairs to build a proper distribution. Second, the *a-contrario* method can use the information of each individual, minimizing the miss-classifications due to the Doddington's Zoo problem. Although we didn't explore this specific point here, we expect to do so in future works, comparing the results with some of well known works [14, 25, 24]. Finally, the *a-contrario* strategy can also adapt to new samples since the impostors distribution can be built between a Query sample and all the Gallery samples.

In future works, we also expect to address the multi-modal configuration. Since the proposed method is applied to the score fusion level, we expect to obtain good results also on the multi-modal case.

Acknowledgments

This work was partially supported by the ANII (Uruguay's National Agency for Research and Innovation) research grant FMV1-2014-1-103996. Luis Di Martino was partially supported by the DNIC (Uruguay's National Civil Identification Bureau) and the SCAPA-IIE.

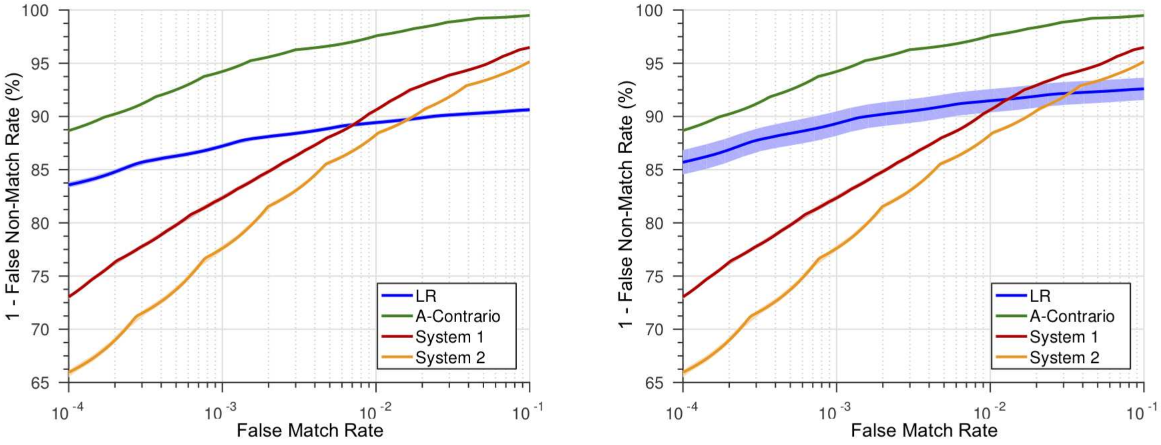
References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[2] A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.

[3] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis: A Probabilistic Approach*. Springer, 2007.

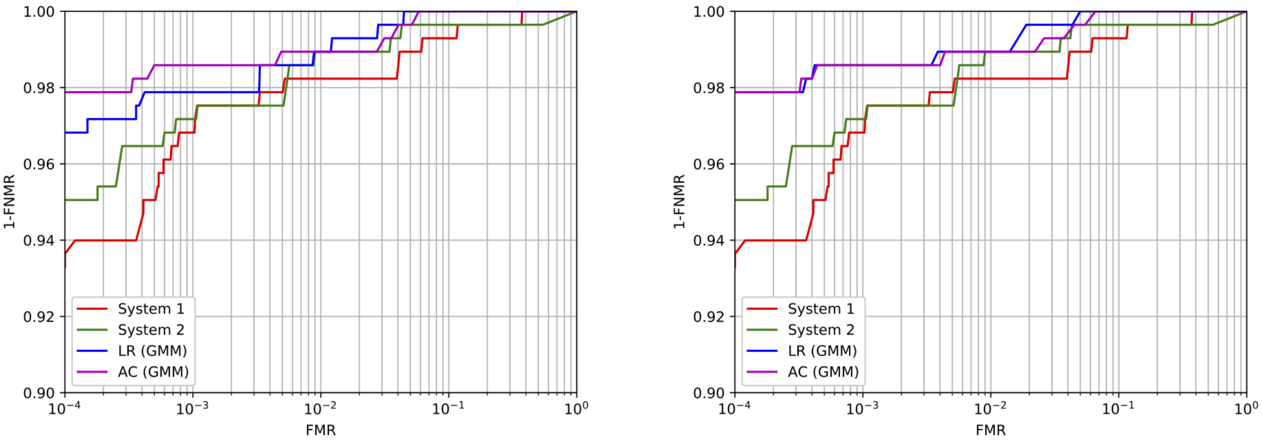
[4] L. Di Martino, A. Fernández, R. G. von Gioi, F. Lecumberry, and J. Preciozzi. A statistical approach to reliability estimation for fingerprint recognition. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–8. IEEE, 2016.



(a) Sample Ratio 0.01

(b) Sample Ratio 1.0

Figure 7: Fusion results in BSSR1-Face database. Results obtained using different samples ratios to compute the probability distributions of both the LR and *a-contrario* approaches.



(a) Sample Ratio = 0.1

(b) Sample Ratio = 1

Figure 8: Fusion results on Face Recognition using ArcFace on AgeDB dataset.

[5] L. Di Martino, J. Preciozzi, F. Lecumberry, and A. Fernández. Face matching with an a contrario false detection control. *Neurocomputing*, 173:64–71, 2016.

[6] L. D. Di Martino, J. Preciozzi, F. Lecumberry, and A. Fernández. An a-contrario approach for face matching. In *ICPRAM*, pages 377–384, 2014.

[7] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. Technical report, DTIC Document, 1998.

[8] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396, 2002.

[9] A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *The Annals of Applied Statistics*, 1(1):179–190, 2007.

[10] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[11] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38:2270–2285, 2005.

[12] A. Jain, A. A. Ross, and K. Nandakumar. *Introduction to Biometrics*. Springer, 2011.

[13] A. K. Jain, K. Cao, and S. S. Arora. Recognizing infants and toddlers using fingerprints: Increasing the vaccination coverage. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014.

[14] A. K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I. IEEE, 2002.

[15] S. S. . B. Joint Technical Committee ISO/IEC JTC 1, Information technology. Iso/iec 19795-2 information technology biometric performance testing and reporting part 2: Testing methodologies for technology and scenario evaluation, 2007.

[16] A. Kale, A. K. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identification. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5, 2004.

[17] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

[18] G. B. H. E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

[19] D. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, 1985.

[20] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.

[21] M. Mottalli, M. Tepper, and M. Mejail. A contrario detection of false matches in iris recognition. In *Iberoamerican Congress on Pattern Recognition*, pages 442–449. Springer, 2010.

[22] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain. Likelihood ratio-based biometric score fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):342–347, 2008.

[23] National Institute of Standards and Technology. Biometric Scores Set Release 1. *http://www.itl.nist.gov/iad/894.03/biometricscores*, 2004.

[24] N. Poh, A. Ross, W. Lee, and J. Kittler. A user-specific and selective multimodal biometric fusion strategy by ranking subjects. *Pattern Recognition*, 46(12):3341–3357, 2013.

[25] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the doddington zoo effect in biometric fusion. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–7. IEEE, 2009.

[26] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[27] M. Singh, R. Singh, and A. Ross. A comprehensive overview of biometric fusion. *Information Fusion*, 52:187–205, 2019.

[28] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):450–5, mar 2005.

[29] B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan. Studies of biometric fusion. *NIST Interagency Report*, 7346, 2006.

[30] M. Vatsa, R. Singh, A. Noore, and A. A. Ross. On the dynamic selection of biometric fusion algorithms. *IEEE Transactions on Information Forensics and Security*, 5(3):470–479, 2010.

[31] A. P. Witkin and J. M. Tenenbaum. On the role of structure in vision. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 481–543. Academic Press, 1983.

[32] T. Zheng, W. Deng, and J. Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.