ISSN 1688-2806


Universidad de la República
Facultad de Ingeniería


# Outliers in Biometrics: An A-contrario Approach

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Luis Di Martino

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Eléctrica.

Directores de Tesis

Federico Lecumberry ..... Universidad de la República
Alicia Fernández ..... Universidad de la República
Javier Preciozzi ..... Universidad de la República

Tribunal

Dr. Rafael Molina ..... Universidad de Granada, España
Dr. Marcelo Fiori ..... Universidad de la República
Dr. Pablo Musé ..... Universidad de la República

Director Académico

Federico Lecumberry ..... Universidad de la República

Montevideo
Friday 29th December, 2017

*Outliers in Biometrics: An A-contrario Approach*, Luis Di Martino.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 90 páginas.
Compilada el Friday 29th December, 2017.
http://iie.fing.edu.uy/

# Acknowledgements

Several people made this work possible.
First, I will like to express my gratitude to my tutors: Federico Lecumberry, Alicia Fernández and Javier Preciozzi. Their expertise and insights allowed me to unlock problems all along the way. Their personal support and kindness made all this process more enjoyable (and even funny). I will also like to thank Rafael Grompone who I consider an additional tutor. His contributions to this thesis are invaluable. And his kindness and great sense of humor made my short internship in Paris very pleasant.
Secondly, I am very grateful to all the people at the Instituto de Ingeniería Eléctrica (*IIE*) whose comments and opinions allowed me to see the problems with another perspective. Especially I would like to thank Pablo Musé with whom I have discussed several details on the work done in this thesis. He guided me in the right direction by sharing his knowledge in the subject and bringing up fruitful discussions.
I am very grateful to my family, both my parents and brothers were backing me all the way. First in obtaining my engineering degree and then along the execution of this master's degree. Every personal achievement is the result of my parent's efforts and dedication in raising me and my brothers.
Last, but not least, I will like to thank my lovely wife Romina, her constant support and patience allowed me to dedicate time to this endeavor. She is, without any doubt, the strong foundations on which I can always rely.

This page intentionally left blank

# Abstract

This thesis addresses the problems of biometrics: *how a person's identity could be determined or validated by using some physical or behavioral characteristic.* Biometry is one of the main research topics in the field of pattern recognition due to its impact on several applications in security and human-machine interaction environments. Several works focus on the improvement of the features extracted in the particular system being presented (face, fingerprint or speech recognition among others), or the metrics used to compare such features, in this work the classification stage is particularly tackled.

A statistical approach is presented based on a well-known *a-contrario* validation strategy. Techniques based on such framework have been widely used in the fields of image processing and computer vision for the detection and matching of visual features. In this work, the method ability to detect outliers/inliers is exploited to detect when two compared biometric samples correspond to the same person. This method is adapted and applied to each of the usual biometric tasks.

First, it is applied to the task of biometric verification, modeling it as a two-class classification problem. The introduced strategy was validated using different datasets and compared against other state-of-the-art commonly used classification methods. Findings of this work have been presented at the *2014 International Conference on Pattern Recognition Applications and Methods (ICPRAM-2014)*, by applying the framework to the face recognition problem in particular. An extension of the conference article has been published in the *Journal of Neurocomputing* publised by Elsevier. In this thesis, the presented strategy is reviewed with an experimental evaluation done in several larger datasets.

Second, the *a-contrario* framework is applied to the identification task. The method is used to validate the confidence of an identification system outputs. What is normally called in the literature as *System Response Reliability* (*SRR*). Such problem has been thoroughly studied lately, the key advantages of using such control are analyzed and discussed. The obtained performance is validated on multiple datasets by comparing with other state-of-the-art approaches. This work has been presented on the *2016 International Conference of the Biometrics Special Interest Group (BIOSIG-2016).*

Finally, the framework is applied to biometric fusion. The key differences in such scenario and the corresponding proposed framework adaptations are analyzed. The proposed technique is evaluated in both artificially generated as real-scenario datasets. The performance is compared against other state-of-the-art statistical fusion strategies.

This page intentionally left blank

# Resumen

En el presente trabajo se abordan los principales problemas en biometría: *cómo se puede determinar o validar la identidad de una persona a partir de una muestra conductual o física de la misma.* La biometría es uno de las áreas con mayor popularidad dentro del estudio de reconocimiento de patrones. Esto se debe a que tiene un gran impacto en diversas aplicaciones de seguridad e interacción entre las personas y sistemas automáticos. Muchos trabajos enfocan estos problemas mediante la mejora de las características extraídas por el sistema biométrico utilizado (sistemas de reconocimiento facial, de huellas, de voz, entre otros). Otros, atacan el problema mediante la mejora de las métricas utilizadas para comparar dichas caraterísticas, en este trabajo se presentan soluciones enfocandose en la etapa de clasificación.

Se presenta un enfoque estadístico basado en el método *a-contrario*. Técnicas basadas en este método han sido utilizadas en los campos de procesamiento de imágenes y visión por computador para la detección de características en imágenes. En este trabajo, se utiliza la capacidad del método para detectar datos salientes para determinar cuando la comparación de dos muestras biométricas corresponden a la misma persona. Se adapta y aplica la técnica *a-contrario* a las tareas usuales en biometría.

Primero, se aplica al problema de verificación, modelando dicho problema como uno de clasificación entre dos clases. La estrategia presentada se evalúa en diferentes bases de datos, comparando su desempeño con otras alternativas de clasificación del estado del arte. Los avances obtenidos fueron presentados en la tercera edición de la *Conferencia Internacional de Aplicaciones y Métodos de Reconocimiento de Patrones (ICPRAM-2014)* aplicando el método al problema de reconocimiento facial en particular. Una extensión del trabajo presentado en la conferencia fue incluido como artículo de revista en el *Journal of Neurocomputing* publicado por Elsevier. En el manuscrito se extiende el trabajo presentado en la conferencia mediante el análisis experimental en diversas bases de datos de mayor tamaño.

En segundo lugar, el método *a-contrario* es adaptado al problema de identificación. La estrategia presentada se utiliza para validar la confianza en el resultado de un sistema de identificación. Lo que normalmente se conoce en la literatura como *confianza de la respuesta de un sistema* (*SRR* por sus siglas en inglés). Este problema ha sido estudiado en detalle en el estado del arte, en el manuscrito se discuten las principales ventajas de utilizar un control de este tipo. La estrategia presentada es validada en varias bases de datos comparando su desempeño con el obtenido por

otras técnicas del estado del arte. Este trabajo fue presentado en la decimoquinta *Conferencia Internacional del Grupo de Interés en Biometría (BIOSIG-2016)*. Finalmente, la técnica *a-contrario* es aplicada en el problema de fusión biométrica. Las diferencias y adaptaciones necesarias para dicho escenario son analizadas. La técnica presentada es evaluada tanto en datos generados artificalmente como en bases de datos reales, comparando el desempeño con el obtenido utilizando otras técnicas del estado del arte.

# Contents

# Chapter 1

# Introduction

Biometrics have achieved high popularity in the last decade as its application 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 [1], in entertainment [2], security of personal devices [3] and human-computer interaction systems [4]. Additionally, security-related applications have also been on the rise. Biometrics systems are being used for automatic checkpoints at countries borders [5] and admission control at sports venues [6] among others. These applications demand constant improvement in accuracy and robustness in order to fulfill their requirements. This explains why biometrics is still one of the most important subjects in both pattern recognition and computer vision areas.

Biometrics systems work by completing three main operations [7]: enrollment, verification and identification. The enrollment is the process by which a new person identity (which will be abbreviated as *ID* in the manuscript) is added to the database of enrolled people, usually called gallery dataset. An input biometric sample (face, fingerprint, iris image or another trait) is associated with a number that represents the person on the system. In the verification process, the system is used to validate a declared *ID*. In this scenario, an input biometric sample and *ID* number are provided. The system gathers the corresponding sample on its database and compares it against the input sample. If they are similar enough the declared *ID* is validated, otherwise it is rejected. When performing the identification, the system only receives an input biometric sample, and its goal is to determine the trait corresponding *ID*. For this, it performs a search comparing the input against all enrolled users corresponding samples in the gallery. The system output, in this case, is an ordered list of the enrolled identities according to how similar each one is to the input. A detailed explanation of how these modes operate can be found on Section 3.2.

As with other pattern recognition systems, biometrics can benefit greatly from the fusion of multiple systems [8]. For instance, if working with fingerprints, multiple samples from a person could be obtained by capturing all fingers impressions. The use of all of them will improve the result obtained by using just one finger. If working with a facial recognition system, by using multiple images of an individual

face, more robustness could be obtained with respect to changes in illumination, pose, aging, etc. In other situations, it can happen that only one biometric sample is available but multiple systems are accessible to process it. In this case, the fusion of the results obtained from these systems could improve the overall result. In [9] the basis and formalization of fusion strategies for biometric applications were introduced. This work is considered as one of the founding works in pattern recognition fusion, gathering more than 2500 citations. Despite the fact that the biometrics applications are targeted in the article, the concepts introduced could be applied to any pattern recognition system. The fusion schemes presented in the article are widely used for its simplicity, ease of implementation and because they do not require model training. Various well-known articles and technical reports ( [10], [11], [12], [13], [14]) validate the presented fusion approaches. Other biometric fusion approaches make use of trained statistical models [15], [12]. They provide better performance than the simple rules introduced in with the extra cost of having training samples and selecting the correct parameters for the model being trained. In this work, we follow this line of research by adopting a well-known image processing feature detection framework called *a-contrario* [16]. First, its application to biometrics problems is presented by showing how these could be focused with an outlier-detection framework. The model particularities in each biometric application are analyzed and discussed. The results in comparison with other state-of-the-art techniques are presented by evaluating the proposed strategy on several datasets.
The rest of the manuscript is organized as follows:

- **Chapter 2 - A-contrario strategy**
  This chapter presents the *a-contrario* statistical framework by stating its key concepts and mathematical formulation. Its application to biometrics is introduced.

- **Chapter 3 - Biometrics performance evaluation**
  The datasets that are used later for validating the proposed strategy are introduced. The verification and identification biometrics applications are explained in detail and their corresponding evaluation metrics are formalized.

- **Chapter 4 - Verification**
  In this chapter, the application of the *a-contrario* method to the verification problem is presented. The details of its implementation and training of the associated statistical model are studied. Its performance is evaluated on several datasets.

- **Chapter 5 - Identification**
  The reliability of identification systems is a well-known problem in the biometrics research field. In this chapter, the *a-contrario* method application to this issue is presented. The strategy is evaluated and compared to other state-of-the-art approaches.

- **Chapter 6 - Fusion**
  This chapter extends the previous work on individual systems in a multi-biometrics fusion environment. The adaptations for this scenario are presented and the performance is evaluated and compared against other statistical fusion approaches.

This page intentionally left blank

# Chapter 2

# A-Contrario Strategy

This chapter presents the core framework in which the rest of this thesis is based: the *a-contrario* model. First, a brief review of the a-contrario scheme uses in other research areas is presented. Then, the strategy is mathematically formalized and the key points in the adaptation to the biometric related problems are presented.

## 2.1 Introduction

The *a-contrario* framework originates in the attempt of applying the *Gestalt Theory* [17] to the field of *Computer Vision*. In particular, by exploiting the *non-accidentalness principle* (also referred as *Helmholtz Principle*), which in its most general form states that whenever some large deviation from randomness occurs, a structure is perceived. Informally, when applied on digital images, it affirms that there is no perception in white noise. In a broader sense, it indicates that we can find significant events as those who are far from some random or background model. A thorough study of how the non-accidentalness principle is applied in the image processing field, as well as the formulation of the *a-contrario* method and its particularities in various applications can be found in [16]. Algorithms based on *a-contrario* framework were first used in the detection of alignments [18], contrasted edges and grouping [19]. Later, its use has been extended to more complex tasks, e.g., the detection of line segments [20], matching of shapes [21] and matching of SIFT-like descriptors [22].

## 2.2 Formulation

In this section, a formulation of the key concepts behind the *a-contrario* strategy are described. We refer the reader to [16] for a complete description of the method and explanatory practical examples. As introduced before, the method allows to classify a particular realization of an event $\mathbf{e}$ in two possible classes. This could be explained by means of two alternative hypotheses:

- $\mathcal{H}_1$: The realization $\mathbf{e}$ follows *some particular causality.*
- $\mathcal{H}_0$: The realization $\mathbf{e}$ could be obtained *only by chance.*

In the context of classical hypothesis testing some *test statistic* $k(\mathbf{e})$ value is computed from the sample $\mathbf{e}$ for evaluation purposes and is compared against some pre-defined threshold $\mathbf{k}$. Two possible errors could be made as described below.

- *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 $P\left(k(\mathbf{e}) \leq \mathbf{k}|\mathcal{H}_1\right)$
- *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 $P\left(k(\mathbf{e}) \geq \mathbf{k}|\mathcal{H}_0\right)$

In this case $P\left(x|\mathcal{H}_0\right)$ and $P\left(x|\mathcal{H}_1\right)$ are the likelihood of $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively over the possible values of $k(\mathbf{e})$. From these two probabilities, $P\left(x|\mathcal{H}_1\right)$ would be normally harder to compute. For instance, we should count with observations complying with the particular causality we are looking for, or an "expert" to build such a model that explains the occurrence of the causality given $k(\mathbf{e})$. This makes the two-class classification by means of a classical method as *Likelihood Ratio* [23] very difficult. In this scenario, the *a-contrario* framework allows to estimate if the observation could be obtained just *by chance* by testing against a background model that characterizes $\mathcal{H}_0$. When the occurrence of the observation at hand $\mathbf{e}$ is very unlikely under this model we could assume that the realization is relevant and perform a detection. This could be formalized as follows.

**Definition 1 ($\varepsilon$-meaningful event [24])** *We say that an event* $\mathbf{e}$ *is* $\varepsilon$*-meaningful if the expectation of the number of occurrences of this event is less than* $\varepsilon$ *under the background model* $\mathcal{H}_0$*.*

**Definition 2 (Number of False Alarms - NFA)** *Given an event* $\mathbf{e}$ *the number of false alarms (*NFA *) is the expectation of the number of occurrences of this event under the background model* $\mathcal{H}_0$*.*

Definition 1 can be rewritten in terms of the *NFA* defined before. An event $\mathbf{e}$ is $\varepsilon - meaningful$ if its associated *NFA* is less than $\varepsilon$:

$$NFA(\mathbf{e}) < \varepsilon. \tag{2.1}$$

The correct definition of this *NFA* is a central problem in all *a-contrario* methods. However, usually this definition can be reduced to an expression of the following form, which gives an upper bound of the actual *NFA* as defined before.

**Definition 3 (Number of false alarms - *NFA* )** *The number of false alarms of an event* $\mathbf{e}$ *is defined as:*

$$\mathrm{NFA}(\mathbf{e}) = N_t \cdot P\left(\mathbf{e}|\mathcal{H}_0\right), \tag{2.2}$$

*where* $N_t$ *is called the number of test and accounts for all possible configurations of the event* $\mathbf{e}$*.*

We can often show that the expectation of the number of occurrences of an event $\mathbf{e}$ satisfying $\text{NFA}(\mathbf{e}) < \varepsilon$ is actually less than $\varepsilon$ [25]. For this reason, defining an event as $\varepsilon$-meaningful, whenever $\text{NFA}(\mathbf{e}) < \varepsilon$, is still consistent with Definition 1 and ensures that the method is robust in the sense that no more than $\varepsilon$ "false detections" will be obtained due to noise. This will be explicitly formalized in the following chapters where the *a-contrario* framework is used.

## 2.3 Relation with classical hypothesis testing

The statistical test being performed in the *a-contrario* framework could be easily related with the classical *statistical hypothesis testing* [26]. We explain this relation briefly, a more complete analysis is presented in [16] and [25].
If in the previous formulation a single test is made ($N_t = 1$), the threshold $\varepsilon$ accounts for the significance level of the test and the null hypothesis would be rejected whenever its *p-value* is less than $\varepsilon$. But, the *a-contrario* framework is applied in a scenario of *multiple hypothesis testing* or *multiple comparisons* [27]. The problem of *multiple comparisons* arises when testing a hypothesis separately on several tests, each of them capable of rejecting the corresponding hypothesis. The failure in compensating the effects introduced by the multiple tests being done could invalidate the statistical test. As an example of this situation, suppose we are evaluating the efficacy of a new drug. The drug will be an improvement over the existing ones if it reduces any one of a number of symptoms of the corresponding disease. We define the null-hypothesis $\mathcal{H}_0$ as follows: *"the new drug is not more efficient than the already existing ones"*. As more symptoms are considered, it becomes increasingly likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom. Therefore, the likelihood of incorrectly rejecting the null hypothesis increases.
In the context of multiple hypotheses testing the following quantities are defined:

- $N_t$ denotes the number of tests.
- $V$ denotes the number of times $\mathcal{H}_0$ was rejected.
- $FWER = P\left(V \geq 1 | \mathcal{H}_0\right)$ is the *Familywise Error Rate*, the probability of at least one false alarm.
- $PCER = \frac{E(V)}{N_t}$ is the *Per-Comparison Error Rate*, the expectation of the proportion of false alarms among the total number of tests.
- $PFER = E\left(V\right)$ is the *Per-Family Error Rate*, the expectation of the number of false alarms.

To alleviate the previously introduced problem, the *Bonferroni* correction [28] is normally used. The correction is usually presented as a way of controlling the $FWER$. By rejecting each test whenever its *p-value* is less than $\frac{\alpha}{N_t}$, the obtained $FWER$ is less than $\alpha$ due to Bonferroni inequality. When using the

*a-contrario* approach we are not evaluating if the null-hypothesis globally stands true. We just look for the average number of realizations obtained that contradict the background model. Thus, the quantity being assessed is the $PFER$ considering that, each detected realization, is a false detection according to the model. The Bonferroni corrections is still valid in this scenario, if each test is rejected whenever its *p-value* is less than $\frac{\alpha}{N_t}$, the $PFER$ is less than $\alpha$.
In summary, both the *a-contrario* as well as the Bonferroni correction end up proposing the same threshold over the probability of the event. The only difference being in the origins of both approaches and the formal enouncement of the quantity they are controlling.

## 2.4 Applicability to biometrics

The *a-contrario* framework is based on the detection of events that are very rare to occur in a background model and therefore enables to classify a particular observation of an event in two possible classes. This fundamental idea behind the *a-contrario* strategy could be easily adapted to the problem of identity verification/identification in biometrics. When comparing biometric samples, regardless of the particular modality used, the comparison of samples from different people is the usual case. This happens, at least, when the gallery dataset is sufficiently large in order to make the identification/verification difficult enough leading and requiring the search for a robust classification scheme with good performance. As the problem of identity verification is a two-class classification problem, the adaptation of the introduced framework is straightforward and is presented below. For the identification problem, the model could be adapted to perform what is called *System Response Reliability* ($SRR$), which is further detailed in section 5.2. To better illustrate this adaptation consider the following example. Suppose there is a population of $N = 1000$ people and that, for each one, we have a pair of corresponding biometric samples: $(g_i, q_i)\,, \quad i = 1 \dots N$. One sample from each pair, noted as $g_i$, is used to enroll the individual and therefore forms part of the gallery dataset. The other one, represented by $q_i$, is part of a probe dataset we later use to perform the verification task and test a biometric system at hand. Then, as usual when evaluating a biometric verification system, we compare all the elements in the probe set against the ones enrolled in the gallery. This gives place to $N^2$ comparisons, each one of them could be classified in the genuines and impostors classes. We end up having two very unbalanced classes: the impostors one will have $N\,(N-1)$ samples, and the representatives in the genuines side just $N$. This example is graphically represented in Figure 2.1.

In this scenario, it is clear that the information we have for the impostors class is higher than the one we have of the genuine class. This disparity in representatives of each class is typical in biometric verification environments. This difference in information amount could be used to train a better classifier for tagging the unknown outcomes of the system. Thus making the use of *a-contrario* strategy ideal to the particular problem at hand: one could try to asses if a particular comparison between samples corresponds to the impostors class. This class is then

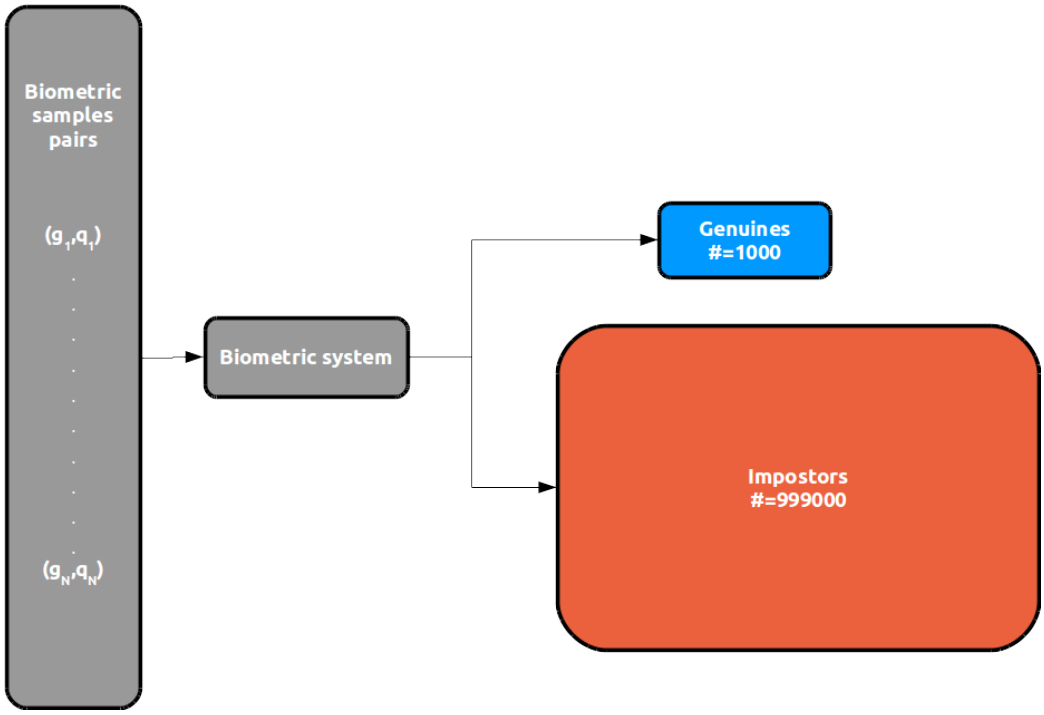



Figure 2.1: Genuine and impostors in toy example.

considered the background model for which we have a lot of information leading to precise and accurate models.

This page intentionally left blank

# Chapter 3

# Biometrics performance evaluation

The inclusion of this chapter, that describes databases and performance metrics, so soon in this thesis may seem unconventional. But this follows good reasons: throughout the next chapters the application of the *a-contrario* framework to the different biometric applications and problems is presented. In each case, the proposed strategy is evaluated across several datasets using the metrics that are described below.

## 3.1 Databases

Two datasets are used in the evaluation of the presented framework: *BSSR1* and *MFCP2-MCC* .

### 3.1.1 *BSSR1*

The *Biometric Score Set - Release 1 (BSSR1)* [29] is a multimodal database generated by the National Institute of Standards and Technology (*NIST*). The goal in creating such a database was to provide the research community in biometrics with a common dataset for testing biometric fusion techniques. The aim with this dataset is to allow the study of fusion and not to advance in the recognition of a particular biometric modality only score values are distributed. This is important as one does not have any additional information that could be used to filter the enrolled users or understand why a particular score value was obtained. It includes three partitions according to the data modalities included in each of them, these are presented below.

#### BSSR1-Face

A first partition includes the results obtained from different face recognition systems aiming to study a multi-algorithm fusion over the same biometric instances. 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. In each of these subsets the comparisons were done using two different face recognition systems called $C$ and $G$. Both systems provide a score value in order to asses how much similar the compared faces are. System $G$ returns a score in range $[0, 1]$ whereas system $C$ produce a score in range $[0, 100]$.

#### BSSR1-Fingerprint

The second partition is used to study a multi-instance fusion by using different instances of the same biometric trait and a unique biometric system. Towards that end, the dataset includes the scores obtained with a state-of-the-art fingerprint matching system when comparing pairs of right and left index fingerprint images from 6000 people.

#### BSSR1-Face & Fingerprint

A last partition allows to study a multi-modal fusion by including the results obtained using different biometric modalities and instances. For this, pairs of left and right index fingerprint images as well as pairs of facial images of 517 different subjects are retrieved. The fingerprints are compared by using a unique fingerprint matching system. The pairs of facial images are compared by using two different face recognition systems named $C$ and $G$.

The available information when using each of the partitions of the database is summarized in Table 3.1.

| Partition | Ids | Instances | Samples per instance | Matching systems | Genuines | Impostors |
|---|---|---|---|---|---|---|
| Face | 3000 | 1 | 3 | 2 | 3000 | $3000 \times 2999$ |
| Fingerprint | 6000 | 2 | 2 | 1 | 6000 | $6000 \times 5999$ |
| Face & Fingerprint | 517 | 3 | 2 | 4 | 517 | $517 \times 516$ |

Table 3.1: BSSR1 database

### 3.1.2 *MFCP2-MCC*

This dataset is obtained by processing the database *MFCP2* [30] with the *Minutia Cylinder-Code* (*MCC* ) [31] fingerprint matcher. Contrary to the previous dataset in which only the scores were provided, in this case, we have access to the actual fingerprint images that are compared to obtain the corresponding impostors and genuine scores.

The acronym *MFCP2* stands for *"Mated Fingerprint Card Pairs 2"*. This database was released by the *National Institute of Standards and Technology* (*NIST* ) for its use in the development and testing of automated fingerprint classification and matching systems. This dataset includes 27000 pairs of segmented 8-bit gray scale fingerprints images obtained from ten-prints cards. The images were scanned

using a 500 *dpi* (19.7 pixels per mm) resolution. The distributed images were compressed using an implementation of the *Wavelet Scalar Quantization*( *WSQ* ) [32] compression specification.
Because in this case, only the fingerprints images are distributed, they must be compared against each other in order to obtain the corresponding genuine and impostors scores. In order to perform the comparisons, first, the minutiae points are extracted using a proprietary system. Then, the corresponding scores are obtained with the *MCC* representation and matching technique. This algorithm has been extensively used in the field of fingerprint recognition. This is due to its great performance and the fact of being publicly available both its implementation as well as the details of how the algorithm works in the author's article. The *MCC* technique works by coding each minutiae using a $3D$ representation (called cylinders) that are built based on the distances and angles of the other minutiae in the fingerprint.
The *MFCP2-MCC* dataset was built as follows: the images included in *MFCP2* were processed in order to find their minutiae points. Then the *MCC* template corresponding to each fingerprint was obtained. By comparing the templates the corresponding impostors/genuine scores were acquired; the available information of this dataset is summarized in Table 3.2. An example of such fingerprints and the comparison of their corresponding minutiae points is shown in Figure 3.1.

| Ids | Instances | Samples per instance | Matching systems | Genuines | Impostors |
|---|---|---|---|---|---|
| 27000 | 1 | 2 | 1 | 27000 | $27000 \times 26999$ |

Table 3.2: *MFCP2-MCC* database

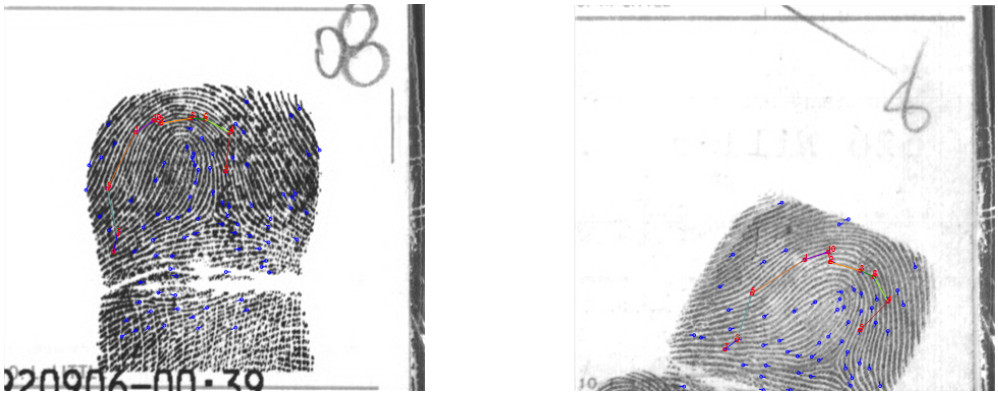


Figure 3.1: *MFCP2* example fingerprints of a same finger with minutiae points matched using *MCC*

## 3.2 Performance evaluation

In this section, the indices used to assess the performance of the proposed framework are introduced. The different measurements presented are used accordingly to the biometric task being evaluated.

### 3.2.1 Verification

In a verification scenario, a subject presents a biometric sample and a declared identity for evaluation. The biometric system being used evaluates the features of the input sample and compares it against the enrolled ones of the declared identity. If the characteristics are similar enough, the system labels the comparison as a *match* and the declared identity is accepted. Otherwise, the obtained result is a *non-match* and the identity is rejected. For this reason, in a verification experiment the usual *False Match Rate (FMR)* and *False Non Match Rate (FNMR)* metrics are referred as *False Accept Rate (FAR)* and *False Reject Rate* respectively. A complete explanation of how this values are obtained and its details are explained in [7], but the key steps in obtaining these metrics are highlighted below.
In order to test a biometric system one must have two probe datasets: the query $Q$ and gallery $G$ sets. These are built by using a set of corresponding biometric samples pairs from some population of size $N$, the pairs are divided leaving one sample in each dataset respectively giving place to the following probe sets:

$$\begin{aligned} Q &= \{q_i\} \quad i = 1, \ldots, N_Q = N, \\ G &= \{g_j\} \quad j = 1, \ldots, N_G = N. \end{aligned} \tag{3.1}$$

Then, a verification experiment is done by comparing the samples in $Q$ against the ones $G$ in an all-versus-all manner obtaining values of distance $D(q_i, g_j)$ with $i = 1, \ldots, N_Q$ and $j = 1, \ldots, N_G$. In order to define the verification used metrics the indicator functions $\mathbb{1}_D^\tau(q_i, g_j)$ and $\mathbb{1}_{id}(q_i, g_j)$ are introduced. The first one reflects the result of the validation done by the system while the second one indicates if two compared samples are from the same person. These indicator functions are defined as follows:

$$\mathbb{1}_D^\tau(q_i, g_j) = \begin{cases} 1 & \text{if } D(q_i, g_j) \leq \tau \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

$$\mathbb{1}_{id}(q_i, g_j) = \begin{cases} 1 & \text{if } id(q_i) = id(g_j) \\ 0 & \text{if } id(q_i) \neq id(g_j) \end{cases} \tag{3.3}$$

It is worth noting that in the definition of $\mathbb{1}_D^\tau(q_i, g_j)$ the parameter $\tau$ controls the value of the indicator function. This represents the usual situation when using a verification biometric system. First the distance between two samples is computed, and then this value is compared against some threshold in order to validate if both belong to the same identity or not. The application of a threshold could be done in some other measure, based on the obtained distance but the underlying reasoning is the same. There is always a threshold parameter that

controls the classification being performed. By using the indicator functions, the previously introduced metrics can be defined as follows:

$$FAR(\tau) = \frac{\sum_{i=1}^{N_Q}\sum_{j=1}^{N_G}(1 - \mathbb{1}_{id}(q_i, g_j)) \times \mathbb{1}_D^{\tau}(q_i, g_j)}{N_Q \times (N_G - 1)}, \tag{3.4}$$

$$FRR(\tau) = \frac{\sum_{i=1}^{N_Q}\sum_{j=1}^{N_G}\mathbb{1}_{id}(q_i, g_j) \times (1 - \mathbb{1}_D^{\tau}(q_i, g_j))}{N_Q}. \tag{3.5}$$

There is a compromise in the selection of the biometric system working point defined by the threshold being applied. When a more restrictive value is used, a reduction in the *FAR* index could be achieved. But this also affects genuine users that for some particular reason (e.g. bad quality of their biometric samples) are rejected in the validation decision, increasing the *FRR*. Another commonly used verification metric is the *Genuine Accept Rate (GAR)*. This measure accounts for the amount of genuine users whose identity is correctly validated by the system. This metric is also referred in the literature as the *Verification Rate (VR)* of the system. Both *GAR* and *VR* terms are used interchangeably. The *GAR* and *FRR* indices are complementary, one could be obtained by subtracting from 1 the other one. Therefore, any of the following equations could be used to obtain the *GAR* value at some $\tau$:

$$GAR(\tau) = \frac{\sum_{i=1}^{N_Q}\sum_{j=1}^{N_G}\mathbb{1}_{id}(q_i, g_j) \times \mathbb{1}_D^{\tau}(q_i, g_j)}{N_Q} = 1 - FRR(\tau)\,. \tag{3.6}$$

When the verification performance of multiple biometric systems are compared usually their *GAR* vs *FAR* relations are shown by means of a *Receiving Operating Characteristic (ROC)* curve as shown in Figure 3.2.

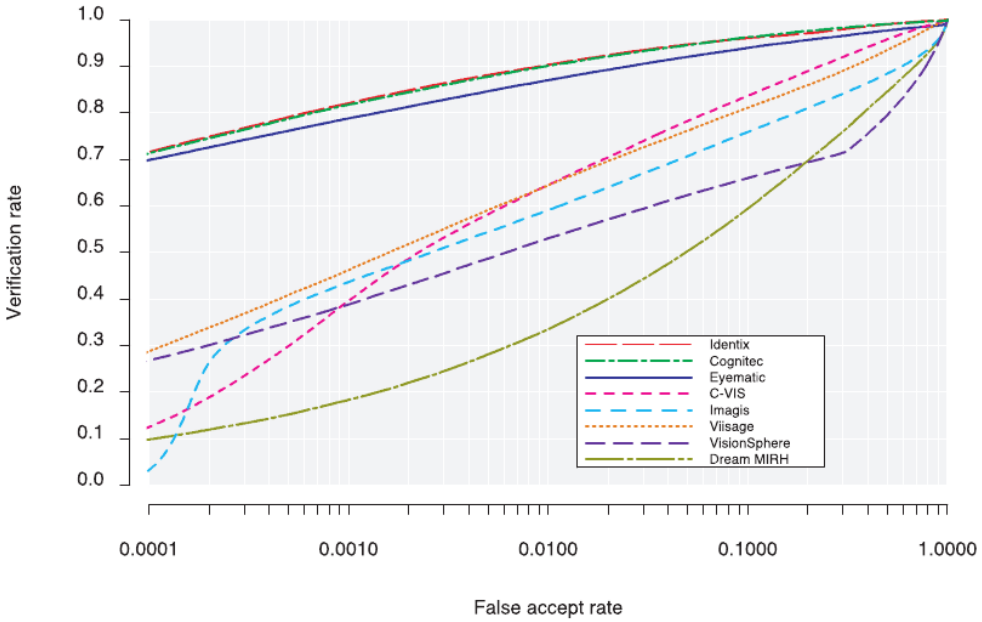



Figure 3.2: *Verification Rate vs False Accept Rate* of multiple face recognition systems in *FRVT 2002*, reprinted from [33]

### 3.2.2 Identification

In this section the performance measures used for evaluating a biometric system working in a identification mode are presented. The presentation is done in a summarized manner, a complete description of the introduced metrics and examples of their functioning can be found in [34].
The identification process can be carried out in two different scenarios: closed-set and open-set identification [7]. The former occurs when it is certain that the searched identity is enrolled in the database. In this case, the assigned identity is the one of the *nearest neighbor* (*NN* ) gallery sample, the closest one to the query sample. The second corresponds to the case where the searched identity may have been or not previously enrolled in the system. In this case, the distance of the gallery closest sample is validated against some pre-defined threshold before assigning its identity to the input sample. This threshold has to be adjusted considering the performance of the biometric system. Usually this is done by using a training dataset and the obtained value its applied globally for all the different system inputs.
In an identification scenario the used biometric system compares an input query sample $q_i$ against all enrolled users $g_j$ in the gallery dataset $G$. As a result, a vector of distances $\boldsymbol{D}(q_i)$ is obtained:

$$\boldsymbol{D}(q_i) = (D(q_i, g_1), \ldots, D(q_i, g_j), \ldots, D(q_i, g_{N_G}))\,. \tag{3.7}$$

This vector is then sorted incrementally obtaining $\boldsymbol{D}(q_i)$ as

$$\boldsymbol{D}(q_i) = \left( D(q_i, g_j^1), D(q_i, g_j^2), \ldots, D(q_i, g_j^{N_G}) \right), \tag{3.8}$$

where the superscript $x$ in $g_j^x$ indicates the position in the sorted array, thus $g_j^1$ is the closest sample to the query input and $g_j^{N_G}$ the farthest one. Then, the *rank* for the probe sample $q_i$ is defined as the position in which the corresponding gallery sample $g_i$ is located in the sorted array $\boldsymbol{D}(q_i)$. For example, $rank(q_i) = n$ if $D(q_i, g_i)$ is the $n^{th}$ smaller distance in $\boldsymbol{D}(q_i)$.
The rank provides a very intuitive measure of how good is the biometric system being used for identifying a query subject. Ideally one would like to have a system that has $rank(q_i) = 1 \quad \forall i = 1, \ldots, N_Q$. If this is not the case, but the system still achieves a small *rank* number (e.g. $n^*$ ) for any input its still could be very useful for the task. In this scenario a system operator could review the closest $n^*$ samples (commonly known as *watch list*) and search for the corresponding person id among the candidates.
In a typical experimental evaluation all the query samples are feed as input to the identification system. Then, the identification rate $P_I(n)$ for rank $n$ is defined as the proportion of query samples at rank $n$ or lower. To obtain this measure, first the cumulative count of the number of probe samples with rank $n$ or less $C(n)$ is computed as following:

$$C(n) = |\{q_i / rank(q_i) \leq n\}|, \tag{3.9}$$

where $|A|$ represents the cardinality (number of elements) of a dataset $A$. Then, the *identification rate* $P_I(n)$ is simply defined as:

$$P_I(n) = \frac{C(n)}{N_Q}. \tag{3.10}$$

In the literature the terms *identification rate* and *recognition rate* are used interchangeably. Therefore, it is common to refer $P_I(n)$ as $RR(n)$, the *recognition rate at rank n*. Usually the system performance is summarized by providing only $RR(1)$, the system ability to hit the correct identity in the first place on the candidates' list. This is simply expressed by dropping the rank index as $RR$ and named the *recognition rate* of the system. When multiple biometric systems are compared their respective *RR vs rank* curves are plotted together as shown in Figure 3.3.

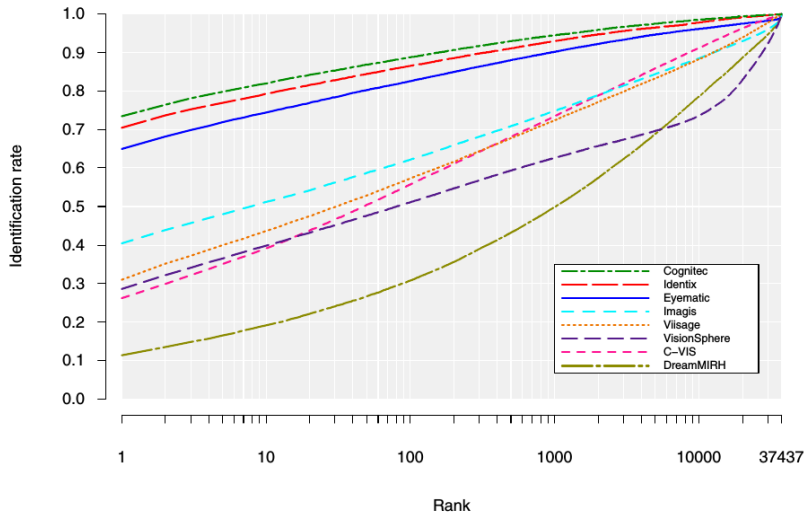



Figure 3.3: *Recognition Rate vs Rank* of multiple face recognition systems in *FRVT 2002*, reprinted from [33]

So far, the presented metrics are the usual in a *closed-set* identification scenario. In a *open-set* operation mode the distances against the query sample are validated using some pre-defined threshold $\tau$. This give place to a metric called the *detection and identification rate at rank n* $P_{DI}(n, \tau)$ defined as follows:

$$P_{DI}(n, \tau) = \frac{|\{q_i / rank(q_i) \leq n, \text{ and } D(q_i, g_i) \leq \tau\}|}{N_Q}. \tag{3.11}$$

This index takes in consideration not only position in which the corresponding gallery sample appears in the sorted distances array but also its value. Whenever the evaluated distance does not comply with the threshold the query subject is assumed to not be present in the gallery of enrolled users.
The use of a validation threshold over the result of the identification being done allows to implement a quality control and therefore estimate the confidence (or reliability) of the system output. Even in the *closed-set* identification context it is useful to have this control as the gallery closest sample could correspond to a different identity of the one corresponding to the input sample. This outcome could be caused by a bad quality input or gallery enrolled sample or just because some enrolled sample of an incorrect identity is more similar than the one of the correct identity. The use of a threshold over the input query closest sample is done in multiple works in the literature ( [35], [36], [37]) and represents and hybrid approach between the *open-set* and *closed-set* scenarios. When a particular identification result does not comply with the control being done, the system's output is considered not reliable, there is no assumption of the query subject being enrolled or not in the gallery dataset as in the *open-set* mode. Simply there is not enough confidence in order to assign its identification as the one corresponding to the closest sample. In this context a second measure called *Number of Reliable*

*Responses* ($NRR$ ) is defined as follows:

$$NRR(\gamma) = \frac{|N_{rr}(\gamma)|}{|N_Q|}, \tag{3.12}$$

where subset $N_{rr}(\gamma)$ represents the subjects $q_i$ of the query dataset $Q$ in where some defined reliability measure $r(q_i)$ complies with a minimum required confidence threshold $\gamma$, i.e.,

$$N_{rr}(\gamma) = \{q_i \in Q | r(q_i) > \gamma\}. \tag{3.13}$$

As the unreliable responses are discarded in the evaluation the recognition rate $RR$ is computed only considering those samples that comply with the confidence control and therefore also depends on the $\gamma$ parameter:

$$RR(\gamma) = \frac{|N_{match}(\gamma)|}{|N_{rr}(\gamma)|}. \tag{3.14}$$

A good confidence measure should achieve an improvement in the $RR$ as $NRR$ becomes lower. This represents the situation in which, as the number of discarded matches increases, the outputs of the system left as reliable are those in which the system identified the input sample correctly. It would not make any sense to use a confidence measure technique that, discarding matches, does not improve the obtained $RR$ . Different reliability techniques are usually compared by plotting in a same figure their respectives $RR$ vs $NRR$ as shown in Figure 3.4.

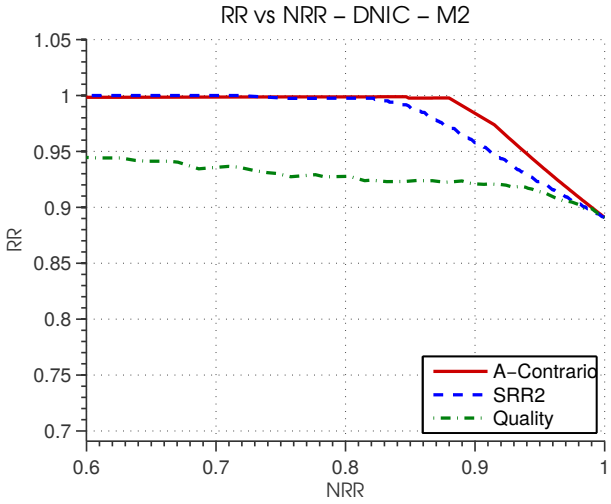



Figure 3.4: *Recognition Rate vs Number of Reliable Responses* of multiple reliability estimation systems. Reprinted from [38]

This page intentionally left blank

# Chapter 4

# Verification

Following the *a-contrario* method formalization in Section 2.2 and the considerations in Section 2.4, the framework should be adapted for the particular problem of biometric verification as explained below.
The background model (or null-hypothesis) $\mathcal{H}_0$ represents the impostors class. A particular sample that meets the hypothesis is then obtained when two biometric samples of different people are compared. Complementarily, the hypothesis of interest $\mathcal{H}_1$ is obtained when the comparison is done between samples of the same person.
Independently from the particular biometric trait, the feature extraction or the similarity measure, we always end up with a distance associated to the comparison between two biometric samples. The output of the particular system at hand could use a similarity measure instead of a dissimilarity one. This is the case when the output is a score that assess how much alike the samples are. In this particular case, the score measure could be easily converted to a distance by taking the inverse value. Special care should be taken in the case the score could be zero, in this particular scenario usually a predefined maximum distance is assigned to the comparison.
Using the notation in example 2.1, let's assume that a particular match between probe and gallery samples, $q_i$ and $g_j$ respectively, is evaluated. This is a particular realization of the event we are trying to classify. We now seek to validate if the distance between them, $D(q_i, g_j) = d_{i,j}$, is rare enough as to consider that the two samples are from the same person or it could be obtained just by chance. For this, we need to characterize the $\mathcal{H}_0$ hypothesis and compute $P(d_{i,j}|\mathcal{H}_0)$: the conditional probability of this event under the null hypothesis. The used background model depends on two factors: the samples used for the estimation and the chosen method to compute the cumulative distribution function (*cdf*) based on these samples. Different strategies on the selection of the samples and estimation of the *cdf* are discussed in Sections 4.1 and 4.3, respectively. From now on, we assume that this probability is given.
After computing the conditional probability $P(d_{i,j}|\mathcal{H}_0)$, we follow the *a-contrario* theory and define its associated *NFA* using the number of test. As explained in Section 2.2, this number depends directly on the configuration of the tests being

realized. A discussion on the setup of this value and its relation to the experiment done is presented in Section 4.2. The previous steps are summarized in Algorithm 1.

**Algorithm 1** *a-contrario* verification validation

**procedure** ACONTRARIOVALIDATE($d_{i,j}, N_{test}, \varepsilon$)
  $\mathcal{H}_0 = obtainBackgroundModel(i)$ ▷ **Step 1**
  $P(d_{i,j}|\mathcal{H}_0) = computeProbability\,(d_{i,j}, \mathcal{H}_0)$ ▷ **Step 2**
  $NFA(d_{i,j}) = N_{test} P(d_{i,j}|\mathcal{H}_0)$ ▷ **Step 3**
  **if** $NFA(d_{i,j}) < \varepsilon$ **then** ▷ **Step 4**
    **Return:** $(q_i, g_j)$ validated
  **else**
    **Return:** $(q_i, g_j)$ rejected
  **end if**
**end procedure**

It is worth noting that we use the identity index $i$ when obtaining the background model. This contemplates the fact that the model could be particular for the evaluated identity as we will see in next sections. Two steps strongly depend on the particular strategy used for the obtention of the background model from the alternatives detailed in Sections 4.1 and 4.3. These are the obtention of the background model (step one) and the corresponding computation of the query distance associated background probability (step two). After these are completed, the procedure follows very simply. The *NFA* value is computed and a threshold is applied over the obtained expectancy of occurrences under the background model.

## 4.1 Modeling $\mathcal{H}_0$

As explained before, a fundamental part of the *a-contrario* framework is the correct modeling of the background model. In this section we first review the available information and then describe different strategies that could be used to model the null hypothesis $\mathcal{H}_0$. Let's recall the standard verification scenario and notation described in Section 3.2.1. We have:

- A gallery dataset $G$ of size $N$, containing the samples $g_1, \ldots, g_N$ of the system enrolled ids.
- A query dataset $Q$ of size $N$ with the corresponding samples $q_1, \ldots, q_N$ of the same ids.

### 4.1.1 Pre-computed

We could choose not to use the query samples at all for the estimation of the background model. This is a great advantage because it allows to have a model

precomputed before doing the actual verification of a query sample. In order to obtain such a model we first perform the comparison between the samples in the gallery dataset in an "all versus all" manner obtaining as result a confusion matrix of distances $D_{G,G}$,

$$D_{G,G} = \begin{pmatrix} 0 & d_{1,2}^{G,G} & \dots & \dots & d_{1,N}^{G,G} \\ d_{2,1}^{G,G} & 0 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ d_{N,1}^{G,G} & \dots & \dots & \dots & 0 \end{pmatrix}, \tag{4.1}$$

where $d_{i,j}^{G,G} = D_{G,G}(i,j)$ represents the value of distance when comparing the gallery samples with ids $i$ and $j$. This matrix has two particular features: it is symmetric and all elements in the main diagonal are zero. The first property is a consequence of the fact that distances $D_{G,G}(i,j)$ and $D_{G,G}(j,i)$ are equal as they are obtained by comparing the same samples $g_i$ and $g_j$. The values in the main diagonal are zero because 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. Because this matrix is symmetric, the useful information could be obtained by considering only the upper or lower triangular sub-matrix and excluding from it the null values in the main diagonal.

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. This could be useful in situations where one could have a lot of gallery enrolled ids with their respective biometric samples but yet does not have corresponding pairs samples as to perform the verification test. As a drawback, while the training could be done just by using the gallery samples, the obtained model could not be useful on the production environment if significant difference exist between the features of the enrolled samples and the query ones used later in a production scenario.

For example, if there is a considerable technological change between the acquisition process of the gallery and query samples, the model trained in the former could not be adjusted to the features of the latter.

Using the available information in $D_{G,G}$ matrix, two different approaches could be followed giving place to the following sub-classification: *general model* or *particular one*.

### General model

We can use all the samples, and train a unique model for all the different enrolled ids. This has the advantage that we only must record one model for all the population. But has as its main drawback that, in the generalization, the model

could miss the particularities that make an id different to the other enrolled users. Indeed, it is well known that, given a biometric trait, some people are more difficult to classify than others [39].
Formalizing, for all the different ids $i$ with $i = 1 \dots N$ the same dataset $T$ is used for training the model:

$$T = \left\{ d_{i,k}^{G,G} : i = 1 \dots N, k = 1 \dots N, i > k \right\}. \tag{4.2}$$

In this case, the training dataset has size $\frac{N \times (N-1)}{2}$.

### Particular model

The alternative to the previous strategy is to build a particular model for each identity in the database. In this way, $N$ different models should be trained. For each particular identity enrolled in the gallery dataset, the comparisons against the other $N - 1$ individuals allows to model how the particular person's biometric features differ from other people samples in the gallery dataset. For example, given a biometric trait and a biometric system at use, we could find that a particular person is very similar to the rest of the population. In this case, the estimated particular model for this individual would provide little information in order to assess a match among its corresponding biometric sample and another one from the population. In the contrary, when someone has a background model that shows greater distances values against the training samples, it would be easier to later classify a match involving this person id. For each *ID* $i$ with $i = 1 \dots N$ we will have a corresponding $T_i$ as follows:

$$T_i = \left\{ d_{i,k}^{G,G} : k = 1 \dots N, k \neq i \right\}, \quad i = 1 \dots N. \tag{4.3}$$

In this case, each training dataset $T_i$ has size $(N - 1)$.

## 4.1.2 Computed in classification time

As was already stated, the pre-computed model estimation strategy has as advantage that it could be computed offline and previous to perform any actual verification test. But, this could potentially give place to a background model that is not accurate. This could happen if, for some particular reason, the query input samples present different features that the ones of the gallery samples used in the training.
The estimation of the model in the classification stage solves this issue. The used strategy is as follows: the input query sample $q_i$ is compared against all the gallery $g_j$ samples obtaining distances $d_{i,j}^{Q,G}$. The training dataset $T_i$ is composed of all these distances with the exception of the one obtained with respect to the particular id being evaluated. In order to formalize the previous procedure, let's consider the matrix $D_{Q,G}$ of distances obtained if we compare all the query samples against the gallery ones as

$$D_{Q,G} = \begin{pmatrix} d_{1,1}^{Q,G} & d_{1,2}^{Q,G} & \dots & \dots & d_{1,N}^{Q,G} \\ d_{2,1}^{Q,G} & d_{2,2}^{Q,G} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ d_{N,1}^{Q,G} & \dots & \dots & \dots & d_{N,N}^{Q,G} \end{pmatrix}, \tag{4.4}$$

where $D_{Q,G}(i,j) = d_{i,j}^{Q,G}$ 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* modeling technique 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 (as least this is what one wants when using a biometric system) but not zero. Also, 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 obtained in classification time, one does not know beforehand which particular comparisons correspond to the impostor class (null hypothesis). Therefore, the only option is to train the background model using all the distances with the exception of the one being evaluated. This will allow to asses if the result being analyzed is rare to occur under the background model.
The training dataset $T_{i,j}$ could be formalized as

$$T_{i,j} = \left\{ d_{i,k}^{Q,G} : k = 1 \dots N, k \neq j \right\}, \quad i = 1 \dots N. \tag{4.5}$$

From the operational point of view, if the number of available samples in the gallery dataset is large, the particular distance evaluated could be also included to model the null hypothesis without changing much the numerical estimation. But, from the point of view of the *a-contrario* framework theory, this inclusion would invalidate the theory. As, in this scenario, the particular distance being assessed will have a minimum probability of occurrence. If for example, the estimation is done empirically using the ratio of distances lower or equal than the one evaluated, we will know beforehand that the distance has a probability of, at least, $1/N_G$ independently of the other samples in the gallery dataset.

### 4.1.3 Summary

The above defined training strategies are all valid from the theoretical point of view but present some key differences from the practical perspective. These are summarized in Table 4.1. How these affect the obtained results depends in various factors: how similar the gallery and query samples are, the robustness of the

biometric system being used, etc. The differences in performance are analyzed in Section 4.4.

| | Type | Training samples | Needs query samples |
|---|---|---|---|
| Pre-Computed I | general | $\frac{N\times(N-1)}{2}$ | no |
| Pre-Computed II | particular | $N-1$ | no |
| Computed online | particular | $N-1$ | yes |

Table 4.1: Model $\mathcal{H}_0$ training strategies

## 4.2 Number of tests

As was introduced in Section 2.2, the *a-contrario* framework works by thresholding the *Number of False Alarms*(*NFA*). This is the expectation of the number of occurrences of the event $\mathbf{e}$ being evaluated under the background model $\mathcal{H}_0$ . When evaluating a match between samples as a particular realization of the event, the only information available is the one provided by the probability of obtaining such realization under this model: $P(\mathbf{e}|\mathcal{H}_0)$. In order to estimate the associated $NFA(\mathbf{e})$, one can assume that the event would arise $N_t P(\mathbf{e}|\mathcal{H}_0)$ times, being $N_t$ the number of experiments performed.
This relation between the number of experiments being done and the meaningfulness of the particular event being assessed is very intuitive. For example, consider that a particular event $\mathbf{e}^*$ has a very low probability of corresponding to the impostors class. Regardless of how small this value is, there is always a corresponding number of tests for which $NFA(\mathbf{e}^*) \geq 1$. This is very intuitive in the sense that, despite the unlikeliness of seeing such event, if one performs a sufficient number of experiments, at least one realization of $\mathbf{e}^*$ could arise just by chance. Making the observation being evaluated irrelevant.
The definition of a threshold $\varepsilon$ in the *a-contrario* framework, instead of thresholding the distance between samples, has a great advantage. It represents an intuitive indicator of the expected number of false alarms and therefore allows to control the performance of the system in advance.
As was introduced in Section 3.2.1, when evaluating a biometric verification system, the performance is usually measured by comparing all the samples in a query dataset $Q$ against their corresponding representatives in the gallery dataset $G$. Therefore, $N_Q \times N_G$ experiments are being done where $N_Q$ and $N_G$ are the sizes of the query and gallery sets respectively, thus the number of tests is configured as follows:

$$N_t = N_Q \times N_G. \tag{4.6}$$

If, after evaluating a particular system, it is used in a production scenario, the number of tests and threshold applied over the *NFA* index should be adjusted accordingly. For instance, suppose we perform $N_V$ verification experiments daily. In each one, the input sample is only compared against the enrolled sample corre-

sponding to the declared identity. In this case, $N_t$ will be set up as follows:

$$N_t = N_V. \tag{4.7}$$

The selected threshold over the *NFA* will be adapted to the expected number of errors we permit according to the system security requirements.

## 4.3 Estimation of the background model probability

Following the previously introduced concepts, independently from the features used to compare two biometric samples $q_i$ and $g_j$, we always have a distance $D(q_i, g_j)$ associated to this comparison. Given a value of distance $D(q_i, g_j) = d_{i,j}$, the key to assess this event under the *a-contrario* framework lies in the computation of its probability under the background model: $P(d_{i,j}|\mathcal{H}_0)$. Having defined the different alternatives for the information used for estimating the background model on Section 4.1, the only requisite still left to compute this probability is the numerical approach used to estimate the probability value itself. We use different strategies for the computation in order to validate the presented *a-contrario* based strategies, presented later on the thesis, working on different conditions.
We start with a simple empirical approach by counting the distances in the training dataset that are smaller than the distance being assessed. This gives place to

$$P(d_{i,j}|\mathcal{H}_0) = \frac{|d_k \in T : d_k \leq d_{i,j}|}{|T|}, \tag{4.8}$$

where $T$ represents the model $\mathcal{H}_0$ used dataset shown previously ( 4.2, 4.3, 4.5). Note that with this frequentist approach we are directly approximating the *cdf* of $d_{i,j}$ under the background model. A quick sanity check could show us that this approximation is indeed correct. For instance, if we take the ideal case of a non-impostor case in which $d_{i,j} = 0$ (this will imply not only that the compared biometric samples are of the same person, but that the sample was compared against itself) there will not be any model distance $d_k$ lower than zero and therefore the probability under $\mathcal{H}_0$ hypothesis will be null. In the opposite case, if we take a value of distance that ideally represents the impostor case, say $d_{i,j} \to \infty$, all distances in $T$ will be lower than this value and therefore we will end up with $P(d_{i,j} \leq \infty|\mathcal{H}_0) = 1$. In this scenario, even by realizing a single experiment we will have at least one expected false alarm associated to the evaluated comparison and therefore no special causality could be assigned to it.
Another possible approach is to use the training data to compute the *probability density function (pdf)* $p_{\mathcal{H}_0}(x)$, as an intermediate step in the estimation of the probability. After the *pdf* is obtained, the probability under the background model for a particular match with $D(q_i, g_j) = \delta$ could be computed as follows:

$$P(d_{i,j}|\mathcal{H}_0) = \int_{-\infty}^{d_{i,j}} p_{\mathcal{H}_0}(x)dx \tag{4.9}$$

We have analyzed two different well-known approaches for estimating such function: *Kernel Density Estimation (KDE)* [40] and *Gaussian Mixture Models (GMM)* [41]. Both methods are very popular and not a key contribution of this work. Therefore, we present here a very short introduction, further details can be found in their respective references.
*KDE* is based on the estimation of the underlying probability density of the training data by assigning a kernel function to each training sample. The selection of the kernel bandwidth should be done carefully. The combination of these kernels is then normalized in order to obtain a valid *pdf* . In this work, we use a *Matlab* implementation of a *KDE* technique based on the smoothing properties of linear diffusion processes [42]. The key idea of this approach is to view the kernel from which the estimator is constructed as the transition density of a diffusion process. On the other hand, *GMM* works by adjusting a linear combination of a variable number of Gaussian distributions whose means and variances/covariances adapt to the training data. The parameters are usually tuned by using the *expectation–maximization (EM)* algorithm. We use a Matlab implementation [43] that automatically selects the best number of Gaussian in the mixture and allows to work both with unidimensional as well as multi-dimensional data. The authors propose a strategy for selecting the number of Gaussian distributions in an unsupervised way. They do this in a way that avoids the usual issues with the conventionally used *EM* strategy. The presented method is based on a criteria similar to *Minimum Message Length (MML)* giving place to a modified *EM* algorithm. The key difference in the authors' approach is that they do not use *MML* as a model selection criterion to choose one among a set of candidate models; instead, they integrate seamlessly estimation and model selection in a single algorithm.

## 4.4 Experimental evaluation

In this section, the results obtained using the *a-contrario* technique in the verification scenario are presented. The evaluation is performed using the databases introduced in Section 3.1. The different background model estimation strategies, as well as methods for computing its underlying probability density explained in Sections 4.1 and 4.3 respectively are analyzed.
The performance of the proposed approach is compared against the one obtained using the usual procedure in biometrics introduced in Section 3.2.1. Recapitulating in a summarized way: the input sample is compared against the sample in the gallery dataset corresponding to the declared identity and the obtained distance is validated against a pre-defined threshold. Doing an abuse of notation, we will refer to this approach as *nearest neighbor* ($NN$) in the presented results. While it is true that in this case it does not matter really which sample in the gallery is the closest one, the verification validation greatly resembles the usual identification procedure, in which the nearest gallery sample distance is validated against a reference value. In both cases, no additional steps besides the comparison against the threshold are performed.

### 4.4.1 Results in *MFCP2-MCC* database

#### Computed in classification time (query vs gallery training)

The following results are obtained when the *a-contrario* model is computed in classification time. In Figures 4.1 and 4.2 the outcomes when $KDE$ and the empirical approach are used are respectively presented.

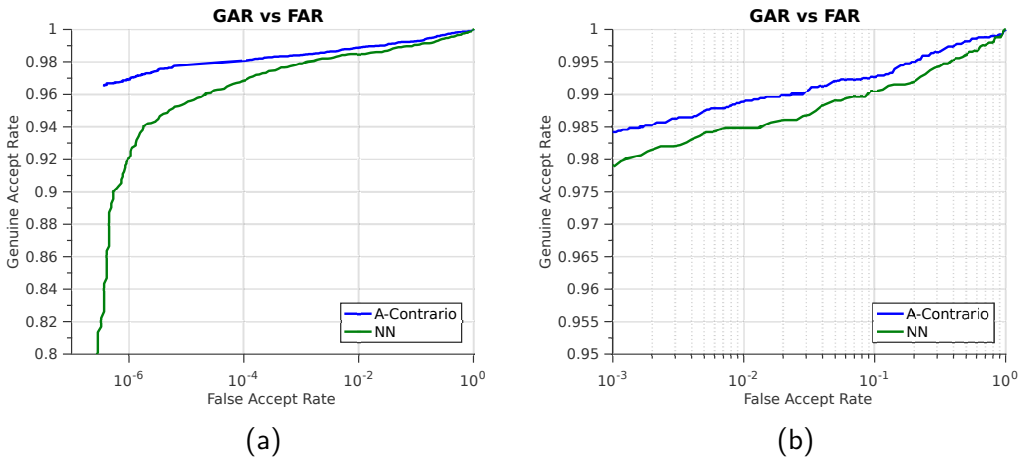



Figure 4.1: Verification results in *MFCP2* using *KDE* . 4.1a *GAR* vs *FAR* plot. 4.1b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

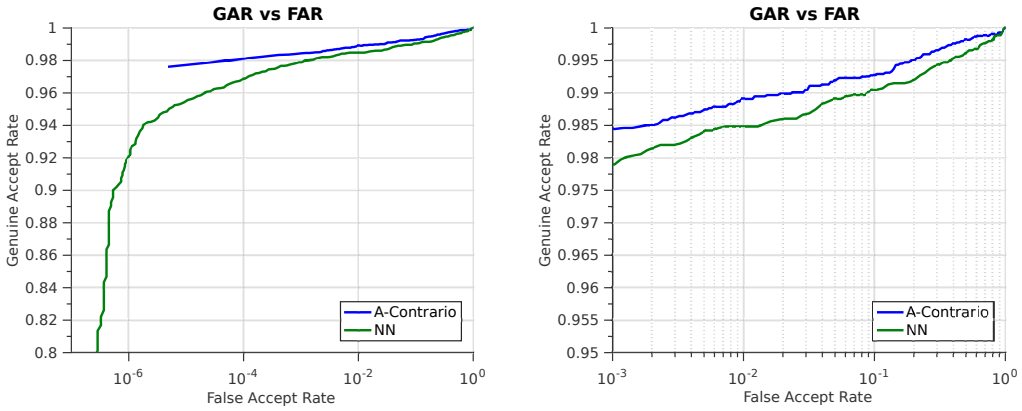



Figure 4.2: Verification results in *MFCP2*, using empirical probability estimation. 4.2a *GAR* vs *FAR* plot. 4.2b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

Two main observations can be made from the obtained results:

- The proposed *a-contrario* approach outperforms the results compared with a threshold over distances. This is to be expected, as in the proposed strategy, more information is used when a biometric match is classified in the impostor/genuine classes.

- The performance of the proposed framework is robust to the used probability estimation strategy. The results obtained using *GMM* are not presented here as they highly resemble the obtained when KDE and the empirical approach are used. Regardless of the use of different probability estimation techniques, the obtained performance remains similar. There is a detail worth noting in the comparison between both probability estimation techniques. When *KDE* is used, a value smaller than the minimum obtained with the empirical approach can be reached. This is due to the fact that using the empirical approach, the computed probability resolution is, at best, $\frac{1}{N}$ being $N$ the number of training samples. This implies less possible classification threshold values $\varepsilon$. When using *KDE* estimation, there is not, a minimum theoretical probability resolution value.

### Pre-Computed general model (gallery vs gallery general training)

In this case, the background model is trained using the comparisons of gallery samples between themselves. The results obtained with *KDE* and the empirical approach for the estimation of the *a-contrario* probability are shown in figures 4.3 and 4.4 respectively.

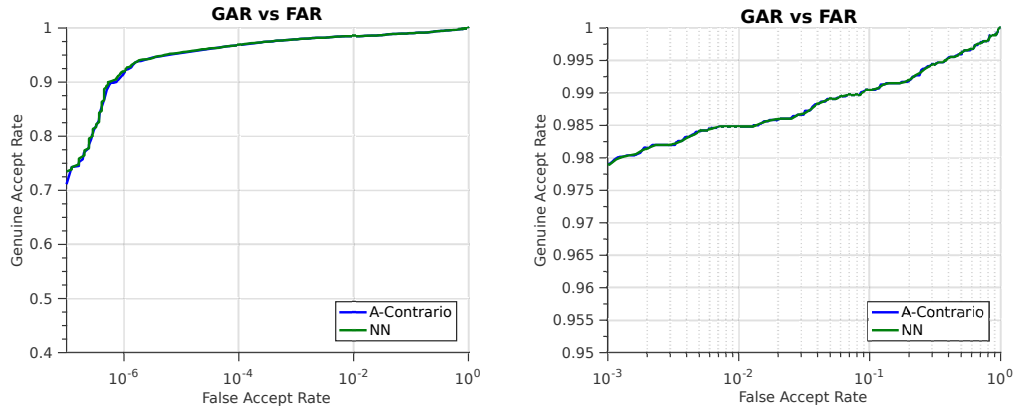



Figure 4.3: Verification results in *MFCP2* using *KDE* . 4.3a *GAR* vs *FAR* plot. 4.3b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

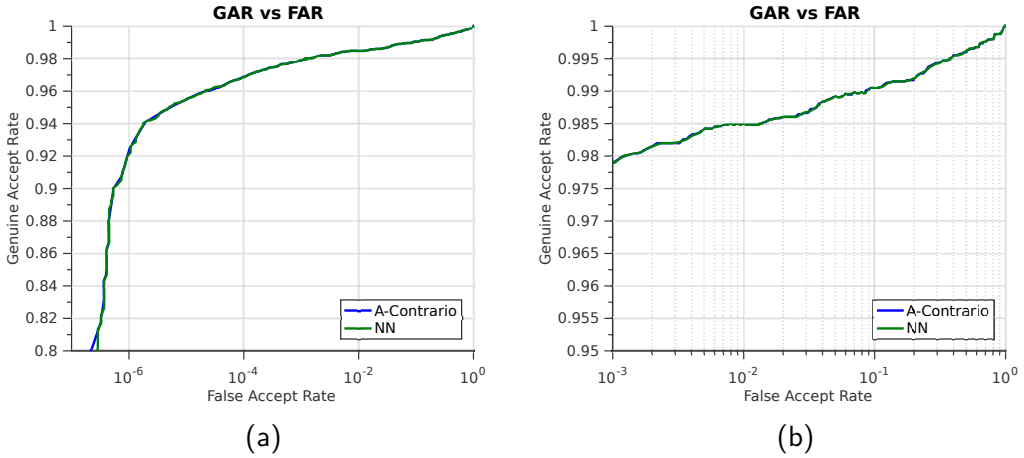
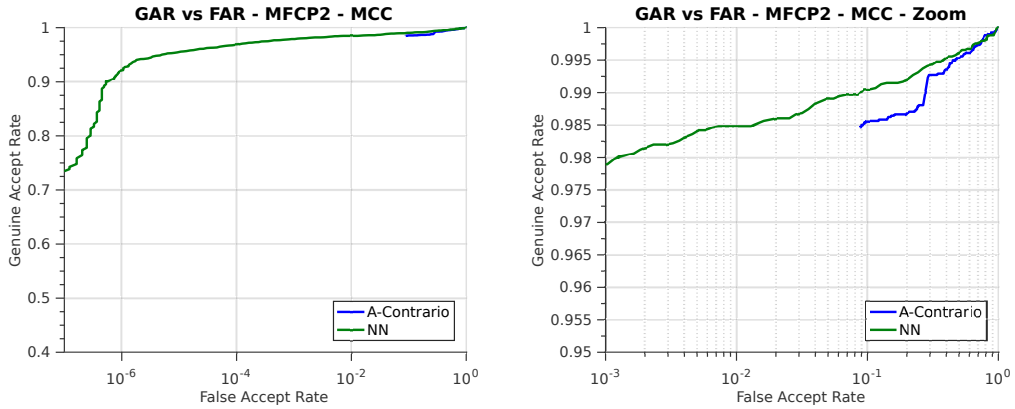



Figure 4.4: Verification results in *MFCP2*, using empirical probability estimation. 4.4a *GAR* vs *FAR* plot. 4.4b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

Regardless of the particular technique used in the estimation of the background model probability, the results with the *a-contrario* approach are the same of those obtained by using a threshold over the matches distances. This could be explained by the fact that, in this case, no extra information for each sample is obtained by using the background model.
In the previous scenario, the training was done for each particular *id* and taking into consideration how the query sample matches each gallery *id* representative. This allows the classifier to learn the particularities of each *id* query sample and its relation with the gallery dataset. In this case, the training does not provide additional information. It provides a mapping from the distance between samples to a likelihood of their comparison belonging to the impostors class based on the training done using the gallery samples. But, the distance obtained using the

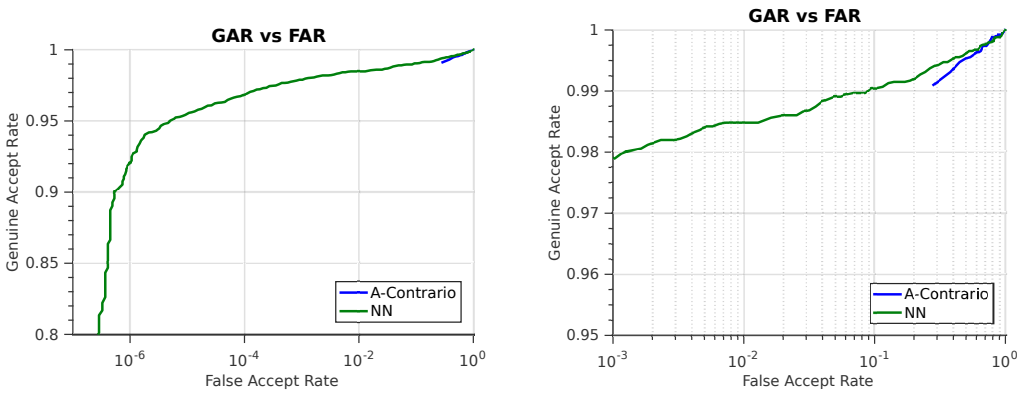
fingerprint matcher system is also directly related to the likelihood of the match belonging to each of the classes. Considering that fingerprints are a very distinctive biometric trait and the *MCC* matcher being used has a very good performance, it is to be expected that the obtained results when using a threshold over the matches distances are equally good.

## Pre-Computed particular model (gallery vs gallery particular training)




(a) (b)

Figure 4.5: Verification results in *MFCP2* database using *KDE* for the *a-contrario* background model estimation. 4.5a *GAR* vs *FAR* plot. 4.5b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.




(a) (b)

Figure 4.6: Verification results in *MFCP2* database using empirical approach for the *a-contrario* background model estimation. 4.6a *GAR* vs *FAR* plot. 4.6b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

Two main conclusions could be drawn from the obtained results.

- First, as well as with the previous training strategy, the obtained results using the *a-contrario* strategy are very similar to the ones obtained using a threshold over the comparisons distance values. Once again, no information of the query sample is used in training the background model. Therefore, an improvement in the performance due to knowing the query sample features could not be obtained.

- Secondly, the minimum achievable $FAR$ index value is greater than in the previous case. This could be explained by the fact that fewer training samples are used, in the former scenario $\frac{N\times(N-1)}{2}$ representatives were available for training the model, in this case only $N-1$. As a consequence, a coarser model is trained in this scenario. This could easily be exemplified when the empirical approach is used in estimating the background model probability. In the preceding case, the minimum achievable probability value is $\frac{2}{N\times(N-1)}$, while as when the particular training is used this minimum value goes up to $\frac{1}{N-1}$.

### 4.4.2 Results in BSSR1 database

In this section, the proposed technique is evaluated in the *BSSR1* database partitions. As it was explained in Section 3.1.1, in this dataset the biometric samples are not available: just the comparisons results between the query and gallery samples are distributed as scores values for each match. Considering this, both training strategies based on the comparisons between gallery samples could not be tested. Therefore, only the results obtained with the online training strategy are presented. The *KDE* technique has been used for estimating the background model *pdf* .

#### Fingerprint datasets

In Figures 4.7a and 4.7b the outcomes for the *left index* and *right index* sub-partitions are respectively presented.

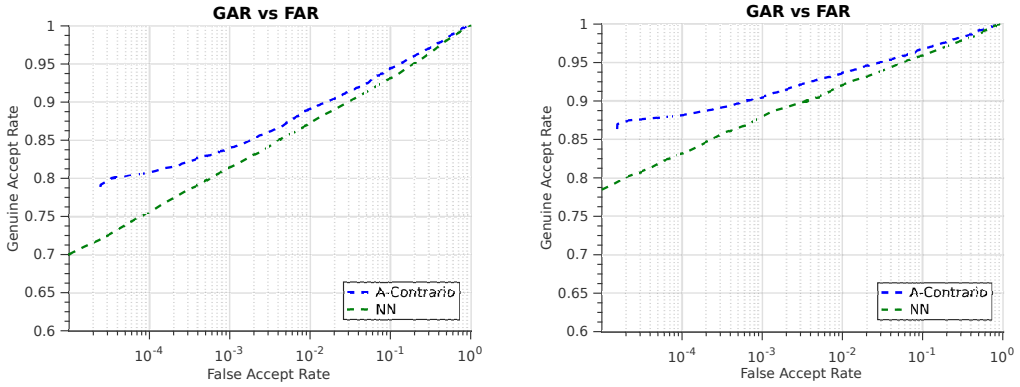
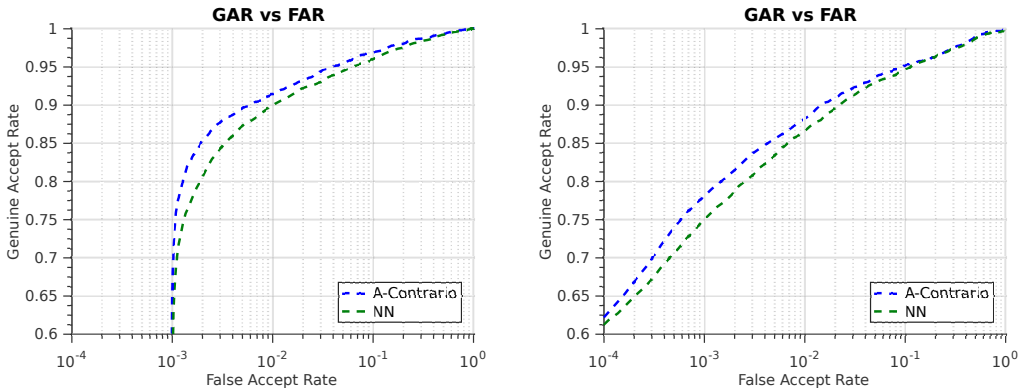



Figure 4.7: Verification results in *BSSR1-Fingerprint* database using *KDE* . 4.7a *GAR* vs *FAR* plot for *left index* sub-partition. 4.7b *GAR* vs *FAR* plot for *right index* sub-partition.

#### Face datasets

In Figures 4.8a and 4.8b the outcomes for the *System C* and *System G* sub-partitions are respectively presented.




(a) (b)

Figure 4.8: Verification results in *BSSR1-Face* database. 4.8a *GAR* vs *FAR* plot for *System C* sub-partition. 4.8b *GAR* vs *FAR* plot for *System G* sub-partition.

In both database partitions, the *a-contrario* approach performs better than the strategy based on a threshold applied to the comparisons scores. This is to be expected considering the results obtained when the same training strategy was used in *MFCP2* database 4.4.1. The outcome in these datasets also highlights a key feature of the proposed approach: it does not depend on any knowledge of the particular biometric trait being used. As was explained before, in these experiments the proposed technique only counts with the scores obtained from the comparisons between biometric samples. No face/fingerprints images or any other additional information is available. This makes the presented method widely applicable.

This page intentionally left blank

# Chapter 5

# Identification

In order to understand how the *a-contrario* framework can be used in a identification scenario it is necessary to first review the key differences between the verification and identification processes. In the former, the biometric system output should be classified in only the two possible classes introduced before: genuine and impostors. In this case, the *a-contrario* method can be easily adapted: it only needs to enable the rejection of one of the hypothesis to be useful in the classification stage. In the identification context the situation is quite different. If one tries to model the problem as one of classifying the system output into classes one should assign one label per each identity $i$ with $i = 1 \dots N$. Considering this is a $N$ classes problem, now its not clear how one could identify and build a background model against which to tests the query sample. For instance, one could formulate a hypothesis for each particular class. Then, obtain a model of the features that samples belonging to each hypothesis must comply, and finally, establish the rejection or acceptance of the input query sample to each of these cases. The problem with this approach is that, in order to obtain accurate models, various representatives of each class must be available. Usually this is not met when performing the identification in large databases with few representatives of each id as happens, for example, in national identity management databases. From this discussion it appears that the *a-contrario* framework is only useful in a verification situation and it is not applicable to the identification process. But, there is still one application where the method could be of great utility: the problem of system response reliability ($SRR$) estimation.

## 5.1 System response reliability

The *system response reliability* ($SRR$) is defined as an index that allows evaluating the confidence of an identification system output. Its use is fostered by multiple factors:

- First, the identification task is much more difficult than the verification one, even for humans.

- Secondly, the characterization of an identification system's performance by only means of *cumulative match characteristic (CMC)* curves (as the RR vs Rank) does not provide a clear scenario of the system's performance in a production environment. This is due to the fact that these performance curves are obtained while the system is being developed/tested. Such performance evaluation is done using training data whose features could greatly differ with the ones of the production data in which the system is later applied.

- Thirdly, the typically used performance *CMC* curves work in a statistical way. They evaluate the average performance of the identification system being used. But, when the system is used on-production, not all identifications have the same difficulty. Some people are more difficult to identify than others because of their inherent features. This has been explained by Doddington et al. in the context of a speaker recognition system [39]. In this article, they introduce different animal classes to exemplify various identification scenarios. Each person to be identified could be classified in one of these classes according to its particular features and its relation to the gallery dataset samples. Having an index that assesses the *reliability/confidence* in an individual output of the system allows to take different actions: ask for a re-capture of the biometric trait, assign the identification request as not valid or ask for a human-operator to validate the obtained result.

- Last, the reliability index could be adjusted to make a particular identification system comply with the particular application requirements: for instance, the identification of a person at passport issuance offices needs to be more reliable than the automatic identification on social networks for tagging.

Different techniques can be used to define a system response reliability measure. A commonly used strategy to estimate the confidence of a biometric system output is to evaluate the quality of the input biometric sample. In the case of fingerprints, several characteristics can be used to measure it. In [44] a summary of different efforts in this direction are presented, as well as the key ideas used in the definition of the *NFIQ* (NIST Fingerprint Image Quality) index. In the field of face recognition, an estimation of input quality could be done by evaluating face pose and illumination distortion as done in [35]. These strategies have the advantage of being independent of the particular feature extraction and matching techniques later used for processing the biometric sample. Nevertheless, this is also their main disadvantage: if there is little relation between the characteristics used in the quality analysis and the ones used in the matching process, the confidence measure obtained from the former may be inaccurate at the classification level.
Another common approach to solve the problem of reliability estimation is to use margins. These quantize the risk associated to a particular system output distance or score. In [37] a margin based on the *false reject rate* (*FRR*) and *false acceptance rate* (*FAR* ) indices is presented. The authors derive a threshold value where these two measures are equal, *equal error rate* (*EER*) operating point, and use the difference between the obtained output and threshold as a confidence measure. The

farther the output is from the threshold the more confidence is assigned. Finally, the match is validated or rejected according to the sign of this difference.
This margin approach differs from other margin strategies as margin in boosting [45] or Vapnik's *margin slack variable* [46] in that the last two can only be computed once the result corresponding class/label is known. Therefore, these strategies are only useful in a training phase where they could be used to select those examples that are difficult to classify and use them to retrain the classifier. The *EER* based margin only requires labeled data in a development phase to obtain the optimum threshold value and then it could be directly applied in a testing scenario. Despite this, the approach presents a major problem for its implementation. The used margin function is global in the sense that the same threshold is used for all the different biometric system inputs. A good reliability measure should be adaptable to the particular features of the input sample and its relation to the gallery enrolled samples. As it is well known that, given a biometric trait, some people are more difficult to classify than others (Doddington's Zoo).
Considering the previous statements, in [36] the authors present a list of required properties that a good confidence measure should meet:

- Take into consideration the whole gallery and the input individual query sample.
- Be well adjusted to the particular features of the biometric system being used.
- Not depend on any *a priori* knowledge of the whole query dataset.

As from the operational point of view, it must provide for each input, a unique reliability measure that can be easily interpreted and used.
The authors present two *"system response reliability"* measures, *SRR1* and *SRR2*, and apply them in the identification process of a face recognition system. Both measurements take as input the sorted list $\boldsymbol{D}(q_i)$ of distances obtained when the gallery's and query samples are compared as shown on Equation 5.1 for an example query sample $q_i$,

$$\boldsymbol{D}(q_i) = \left( D(q_i, g_j^1), D(q_i, g_j^2), \ldots, D(q_i, g_j^{N_G}) \right), \tag{5.1}$$

where the superscript $x$ in $g_j^x$ indicates the position in the sorted array, thus $g_j^1$ is the closest sample to the query input and $g_j^{N_G}$ the farthest one. As usual when performing identification, the closest sample corresponding identity is assigned to the unknown input sample. *SRR1* is based on the relative distance (relative distance between the first two retrieved identities) by using the auxiliary $\varphi_1$ function defined as

$$\varphi_1(q_i) = \frac{F\left(D(q_i, g_j^2)\right) - F\left(D(q_i, g_j^1)\right)}{F\left(D(q_i, g_j^{N_G})\right)}, \tag{5.2}$$

where $F(x)$ represents a normalization function that, when applied to the input distance/score gives as output a value in the range $[0, 1)$ as shown in Figure 5.1.

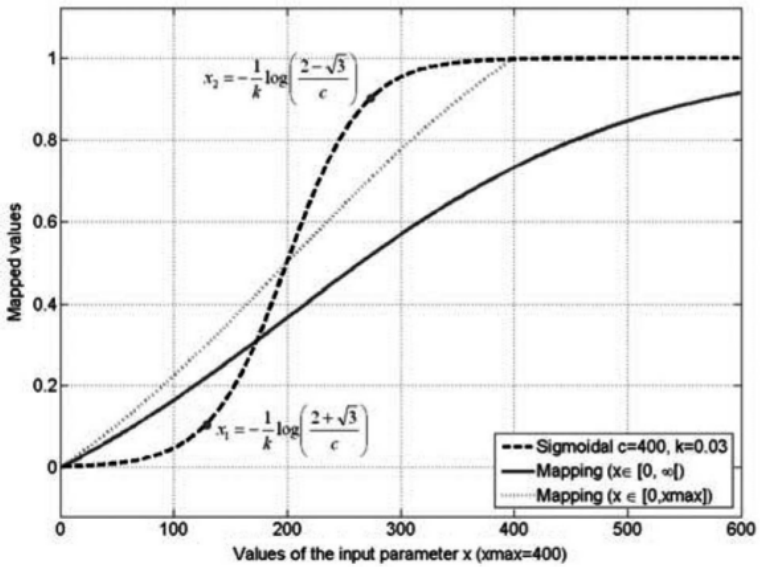



Figure 5.1: $F(x)$ function represented by the authors as "Mapping" in the legend, reprinted from [36].

This function is obtained from the family of sigmoidal functions and is defined by the authors as

$$F(x) = \frac{1 - b^{\frac{x}{x_{max}}}}{ab^{\frac{x}{x_{max}}} + 1}, \tag{5.3}$$

where $a = \left(2 + \sqrt{3}\right)$, $b = \left(7 - 4\sqrt{3}\right)$ and $x_{max}$ accounts for the maximum value of score (or minimum distance) that the system at use could give as output. If this value is not known beforehand, using an estimation $\overline{x_{max}}$ would still assure $F(x) < 1$. Details of how the authors derive this function and its advantages with respect to other normalization schemes are presented in the respective article [36]. *SRR2* use the density ratio (relative amount of gallery samples which are near the assigned identity) which, as explained by the authors, makes the measure a little harder to compute but more robust with respect to outliers. This is done by means of the auxiliary function $\varphi_2$ that is computed as

$$\varphi_2(q_i) = 1 - \frac{|N_b|}{N_G}, \qquad \text{with} \quad N_b = \left\{ g_j^k \in G | F\left(D(q_i, g_j^k)\right) < 2F\left(D(q_i, g_j^1)\right) \right\}. \tag{5.4}$$

$N_b$ is composed by those gallery samples whose distance to the query sample is at most two times the distance of the closest sample and $N_G$ is the cardinality of the gallery dataset.
Both $\varphi_1$ and $\varphi_2$ functions will tend to zero when the closest sample's distance is not very different of the other gallery samples' distances. Thus, a small value in either of the functions would indicate a *not-reliable* identification. On the other hand, when values near to one are obtained, the distance associated with the identification being done is clearly salient among the other ones, therefore indicating a

very-likely correct match. The decision lies in thresholds $\overline{\varphi_m}(m = 1, 2)$ that marks the point of maximum uncertainty and varies both with the biometry and with the classifier. These thresholds are trained from time to time by minimizing the wrong estimates of $\varphi_m$ in a training dataset. Having defined the auxiliary functions $\varphi_m$ and obtained the corresponding $\overline{\varphi_m}$, the authors define $S\left(\varphi_m\left(q_j\right), \overline{\varphi_m}\right)$ as the width of the subinterval from $\overline{\varphi_m}$ to the corresponding extreme of the interval $[0, 1)$ as shown in Equation 5.5.

$$S\left(\varphi_m\left(q_j\right), \overline{\varphi_m}\right) = \begin{cases} 1 - \overline{\varphi_m} & \text{if } \varphi_m(q_j) > \overline{\varphi_m} \\ \overline{\varphi_m} & \text{otherwise} \end{cases} \tag{5.5}$$

Finally, the presented *System Response Reliability* measures *SRR1* and *SRR2* are defined as

$$SRR_m\left(q_j\right) = \frac{|\varphi_m(q_j) - \overline{\varphi_m}|}{S\left(\varphi_m\left(q_j\right), \overline{\varphi_m}\right)}, \quad m = 1, 2. \tag{5.6}$$

The proposed reliability measures complies with the stated requirements and improve the results obtained with the margin strategy [37]. Despite this, they still have two drawbacks: first, the *SRR1* and *SRR2* indices depend on thresholds that are obtained in a training phase. If the characteristics of the gallery dataset or input biometric samples drastically change these thresholds should be retrained. Second, as both measures use different criteria to estimate the system output reliability, they normally do not perform good at the same time. Therefore, a choice of which measure to use should be made before-hand and the selected reliability measure could not be optimal for a particular input or classifier being used.

## 5.2 *A-contrario* framework adaptation

The adaptation of the *a-contrario* method to the problem of reliability estimation is very simple and follows the same strategy used in other state-of-the-art confidence measure techniques. First, the input query sample $q_i$ is compared against all the enrolled ids samples $g_j$ with $j = 1 \ldots N$ in the gallery dataset. The id corresponding to the gallery sample producing the lower distance is assigned to the input. Then, the *a-contrario* framework is used to asses if this classification should be considered as reliable or not in a similar way to the verification scenario. If, according to the model, the obtained distance could happen just by chance, the identification done is considered unreliable. On the other hand, if the result stands out from the background model then a high confidence is assigned to the evaluated biometric match. Summarizing, in the identification setup the closest gallery sample id is assigned to the input and then the obtained distance is classified in the two classes defined before: impostors and genuines. When the impostors class is assigned the response is considered unreliable and viceversa.
The procedure for the a-contrario framework based system's response reliability control is shown in Algorithm 2. It follows closely the one introduced in Algorithm 1 for the verification scenario. There is a key difference between the usage

of the introduced model in both operational modes. In the verification scenario, additional biometric comparisons than the only one required to validate the verification being done may be needed. This happens when the background model is computed in classification time as explained in Section 4.1.2. This results in an extra cost when using the proposed strategy. Meanwhile, in the identification mode, this does not happens as one already has all the distances obtained when comparing the input to the gallery samples, they were already needed in order to perform the identification. Considering this, it makes sense to use this information to customize the trained background model for each particular individual being identified by the system. This gives place to a unique strategy for modeling the null-hypothesis $\mathcal{H}_0$.

**Algorithm 2** *a-contrario* identification system response reliability

**procedure** ACONTRARIOSRR($\boldsymbol{D}(q_i), N_{test}, \varepsilon$)
- $\left[d^*_{i,j}, \widetilde{\boldsymbol{D}(q_i)}\right] = separateTrainingSamples(\boldsymbol{D}(q_i))$ ▷ **Step 1**
- $\mathcal{H}_0 = obtainBackgroundModel\left(\widetilde{\boldsymbol{D}(q_i)}\right)$ ▷ **Step 2**
- $P(d^*_{i,j}|\mathcal{H}_0) = computeProbability\left(d^*_{i,j}, \mathcal{H}_0\right)$ ▷ **Step 3**
- $NFA(d^*_{i,j}) = N_{test}P(d^*_{i,j}|\mathcal{H}_0)$ ▷ **Step 4**
- **if** $NFA(d^*_{i,j}) < \varepsilon$ **then** ▷ **Step 5**
  - **Return:** Reliable identification
- **else**
  - **Return:** Not-reliable identification
- **end if**

**end procedure**

In this case, the algorithm receives as input the sorted vector of distances $\boldsymbol{D}(q_i)$ obtained from comparing the input sample against all enrolled gallery representatives as shown in Equation 5.7. Remembering the notation in Section 3.2.2, the superscript $x$ in $g^x_j$ indicates the position in the sorted array, thus $g^1_j$ is the closest sample to the query input and $g^{N_G}_j$ the farthest one:

$$\boldsymbol{D}(q_i) = \left(D(q_i, g^1_j), D(q_i, g^2_j), \ldots, D(q_i, g^{N_G}_j)\right). \tag{5.7}$$

As we will like to assess if the closest sample distance is different enough to the other obtained values, in the first step of the algorithm these are separated giving place to $d^*_{i,j}$ and $\widetilde{\boldsymbol{D}(q_i)}$:

$$d^*_{i,j} = D(q_i, g^1_j), \qquad \widetilde{\boldsymbol{D}(q_i)} = \left(D(q_i, g^2_j), \ldots, D(q_i, g^{N_G}_j)\right). \tag{5.8}$$

From steps two to four the background model is obtained and used for computing the *NFA* associated with the identification. Finally, the *NFA* is thresholded giving place to the classification of the identification system output.

## 5.3 Experimental setup

The proposed approach based on the *a-contrario* framework is evaluated by using the datasets presented in Chapter 3. The performance is compared against the *SRR1* and *SRR2* reliability measurements introduced in Section 5.1. As was explained in Section 5.1, thresholds $\overline{\varphi_m}(m = 1, 2)$ are needed for the computation of *SRR1* and *SRR2* respectively. In order to perform the training of these, we use a $2 - fold$ cross-validation scheme in each database experiment. The presented *a-contrario* based reliability approach does not require such training, and therefore is directly applied to the testing data in each case. For each database, the corresponding *RR* at *rank 1* and *NRR* performance indices detailed in Section 3.2.2 are computed in each cross-validation fold. Finally, the presented *RR* is the one obtained by averaging in the 2 sets using a common domain of *NRR* values.

## 5.4 Results and conclusions

### 5.4.1 Results in *MFCP2-MCC* database

The results obtained by applying the different response reliability strategies to the *MFCP2-MCC* database are shown in Figure 5.2.

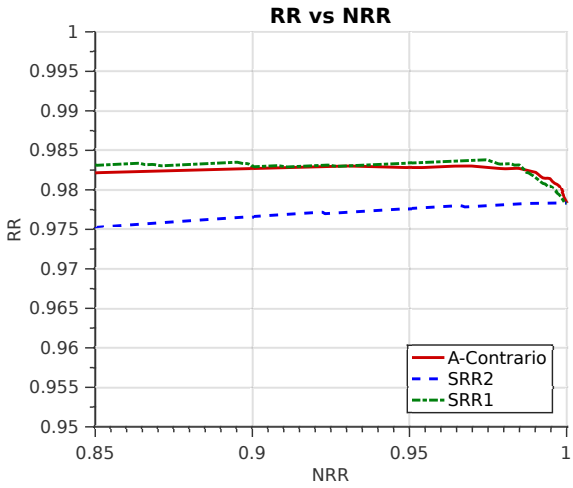



Figure 5.2: *RR* against *NRR* using different reliability estimation techniques

Several observations could be made:

- **The *MCC* fingerprint recognition system being evaluated has a very good performance**
  It reaches almost a 98% recognition rate at *rank 1* without any response reliability control being applied. This can be seen in the right-side of the plot where $NRR = 1$, indicating that all the identifications done in the experiment are considered valid. This great performance could be explained by two

main reasons: fingerprints are a very robust and distinctive biometric trait, and secondly that the *MCC* fingerprint matcher is very good at comparing fingerprints minutiae points. The former reason is the one explained by the authors in [36] of why they do not apply the *SRR1* and *SRR2* measures to fingerprint recognition systems in their experimental evaluation. As we will see later on the evaluation on *BSSR1* fingerprint datasets, this characteristic of fingerprints is only useful if the used fingerprint matching system exploits this distinctiveness. This is clearly the case in this experiment in which the *MCC* fingerprint matcher is used. This algorithm performance has been reported as very good in comparison to other fingerprint matching systems. For instance, its results in the *Fingerprint Verification Competition 2006* (*FVC 2006* [47]) are available.

- ***SRR1* and the *a-contrario* based approach are good *system response reliability* control strategies**
  It can be seen in Figure 5.2 that both approaches achieve an increase in the *RR* at *rank 1* when the number of reliable responses is decreased. As the reliability response control is applied, in the identifications considered reliable the system is correctly identifying the input samples thus making the response reliability strategies very effective. This does not happen when *SRR2* is applied. In this case, a decrease in the *RR* is obtained while the *NRR* is lowered. Thus indicating that the matches considered as not reliable are the ones in which the system is correctly labeling the input query samples.

### 5.4.2 Results in BSSR1 database

The results obtained by applying the different response reliability strategies to the *BSSR1* database are shown below. The results in the Fingerprint datasets are presented first. Secondly, the performance in the Face datasets are displayed. Finally, observations and conclusions of the achieved results are listed.

#### Fingerprint datasets

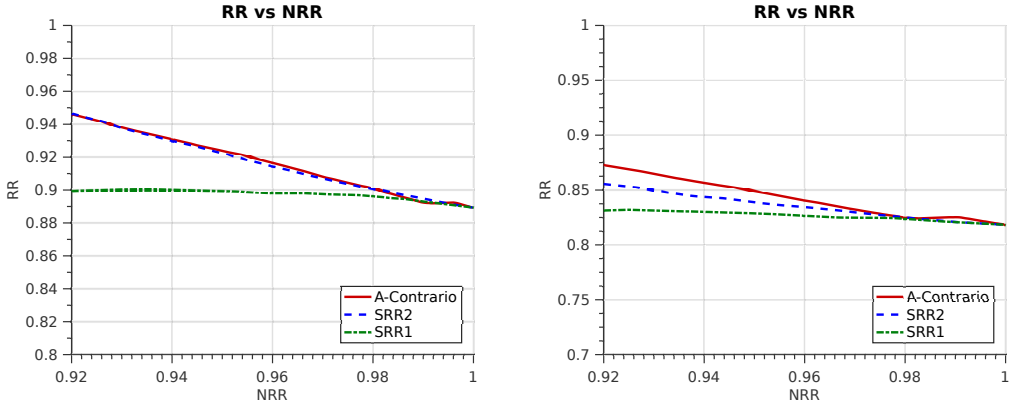



Figure 5.3: *RR* vs *NRR* using *a-contrario* , *SRR1* and *SRR2* reliability estimation techniques. 5.3a *right index* sub-partition. 5.3b *left index* sub-partition.

#### Face datasets

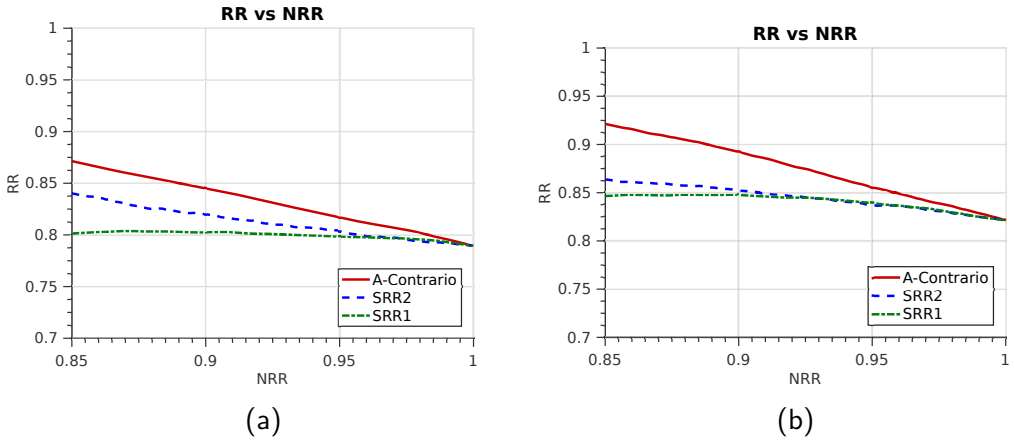



Figure 5.4: *RR* vs *NRR* using *a-contrario* , *SRR1* and *SRR2* reliability estimation techniques. 5.4a *System G* sub-partition. 5.4b *System C* sub-partition.

All performed experiments showed qualitatively similar results, giving place to the following observations:

- **The used fingerprint recognition system performs better than both face recognition systems being evaluated**
  The recognition rate at *rank 1* of each system without any response reliability limitation corresponds to the value obtained when $NRR = 1$. In that particular case, all the identification experiments done are considered valid. At first sight, one would expect that the obtained *RR* at *rank 1* in both fingerprint datasets should be lower than the obtained in the face dataset. This would be explained by the fact that each fingerprint set contains 6000 representatives while the face recognition set only contains 3000 different ids, therefore making the identification problem easier. Regardless of this disbalance in datasets size, two facts explain this equally obtained performance: first, it is a well-known fact that fingerprints suffer fewer variations than faces over time. Secondly, as fingerprints have been the most commonly selected trait for biometric applications, the fingerprint recognition systems have reached a very mature state reaching high-performance levels.

- **In these experiments, *SRR1* does not provide a good measurement of system response reliability**
  A good *system response reliability* measurement strategy should provide an increase in *RR* when the *NRR* is reduced. This would indicate that as more biometric matches are discarded as not reliable, the ones remaining are the ones in which the system performs better. In Figures 5.3a, 5.3b and 5.4a it can be seen that this does not happen when using *SRR1* index. Only on Figure 5.4b it could be seen a little improvement in the *RR* as *NRR* decreases but the increase is marginal. This bad performance could be explained by analyzing the information the $\varphi_1$ function underneath the *SRR1* measure uses. As shown in Equation 5.2, this function compares the distance difference between the two closest query samples in the gallery and normalizes this value using the farthest gallery representative. This criteria is highlighted by the authors as not very robust. For example, consider a biometric system that achieves a great identification performance, being able to identify each person in the query database correctly. Also, consider that this system does not achieve a considerable separation between the first and second closest samples in the gallery dataset. According to $\varphi_1$ function, no output will obtain a high reliability score, regardless of the great performance of the system in terms of its *RR* at *rank 1*.

- ***SRR2* and the *a-contrario* approach are good *system response reliability* measurement strategies**
  While the proposed approach and *SRR2* have equal performance in *Right index* dataset as shown in Figure 5.3a. It obtains greater performance on the other datasets as depicted on Figures 5.3b, 5.4a and 5.4b. Both strategies achieve an increase in the *RR* index when the identification experiments being done are validated, thus decreasing the *NRR* index. The obtained results could be explained by the fact that, in both strategies, all the available information is used. While in the previous case, with *SRR1*, only the infor-

mation provided by the two closest and the furthest gallery samples were taken into account, in this scenario all the gallery samples distances are used in the computation of the metrics. The difference in performance of both strategies lies in the fact that the proposed approach uses the whole distribution of gallery samples' distances against the query sample. Whereas, the function $\varphi_2$ behind *SRR2* only use a clusterization of such distances as shown in Equation 5.4.

- ***SRR1* and *SRR2* performance as response reliability control strategies is not consistently over different datasets or biometric systems.**
  While both strategies comply with the conditions a good reliability measure should have explained in Section 5.1, and besides having a solid theoretical background, the metrics fail to work consistently good in all the performed experiments. Depending on the dataset characteristics and the biometric system being used, one or the other strategies could perform well, but the problem is that knowing which measurement to use could not be determined beforehand. On the other hand, the proposed *a-contrario* approach works consistently over the different test data sets and systems configurations.

This page intentionally left blank

# Chapter 6

# Fusion

As introduced in Chapter 1, the fusion of different biometrics allows to obtain a better performance for both the identification and verification tasks. Additionally, the combination of multiple systems permits to alleviate the problems (uniqueness, universality, etc.) that each system has when working individually.

## 6.1 *A-contrario* strategy adaptation

The *a-contrario* strategies introduced so far deal with one dimensional classification problems and must be adapted for working in a multidimensional scenario. This adaptation requires the revision of two main subjects: the characterization of the random variable whose realizations are classified and the definition of the *NFA* index used in the decision stage. The random variable being assessed accounts for the same event as before: the comparison of two biometric samples $q_i$ and $g_j$. But, in this case, the result of the comparison is not just a value of distance (or score) but instead a vector $\mathbf{D}_{i,j}$ containing the results from every biometric system being fused. If $K$ different outputs are being fused, $\mathbf{D}_{i,j}$ is defined as follows:

$$\mathbf{D}_{i,j} = \mathbf{D}(q_i, g_j) = \left( d^1_{i,j}, \ldots, d^k_{i,j}, \ldots, d^K_{i,j} \right), \tag{6.1}$$

where $d^k_{i,j}$ represents the distance obtained from the biometric system $k$. In order to classify an event, the $NFA(\mathbf{D}_{i,j})$ is computed as

$$NFA(\mathbf{D}_{i,j}) = N_{test} P(\mathbf{D}_{i,j}|H_0), \tag{6.2}$$

where the term $P(\mathbf{D}_{i,j}|H_0)$ accounts for the probability of the particular observation under the background model. Then, the event should be assessed by applying a threshold on it following the same procedure as in the one-dimensional scenario, although in this case the definition of such a threshold is not as simple as before. In the previous situation one has to asses a unique value of distance, in a multidimensional case one has to assess a particular configuration of distance values that could be evaluated by using different criteria. These will define how likely it

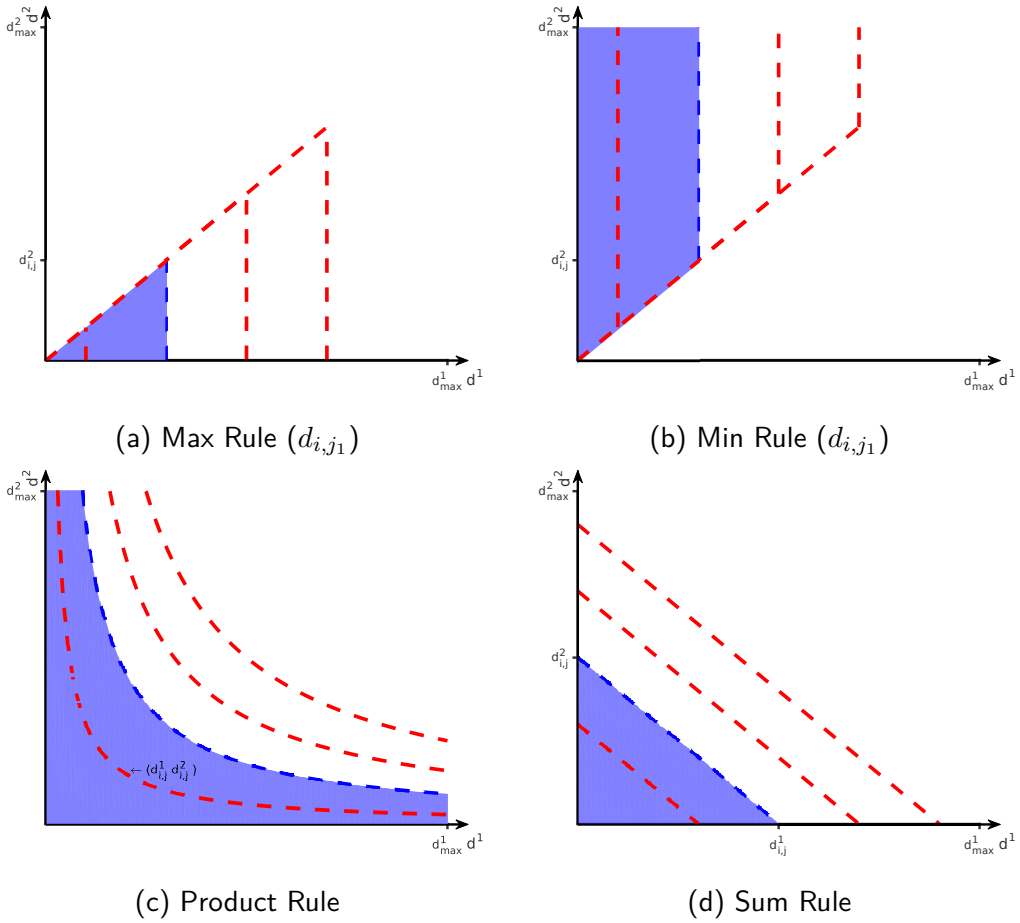
is to obtain other realizations of the random variable complying with the condition imposed by the chosen criteria and therefore define how $P(\mathbf{D}_{i,j}|H_0)$ should be computed.
The definition of a strategy used to evaluate a particular configuration $\mathbf{D}_{i,j}$ is a well known problem in the fusion of multiple pattern recognition systems. In particular, in the well known article of *J. Kittler et. al.* [9] they present several options to define a unique value $d^*$ for classifying a particular arrange of distance values:

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 $d^* = \prod_{k=1}^{k=K} d_{i,j}^k$
4. The sum of the distances $d^* = \sum_{k=1}^{k=K} d_{i,j}^k$

As the value $d^*$ is compared against some threshold in the decision, the different strategies defined above will impose different criteria in the classification. For instance, the first definition will assume that if the minimum obtained distance is small enough the event should be considered meaningful regardless of the remaining distance values. The second definition is more conservative in the sense that it requires that all the distances comply with being small according to the imposed restriction. The third criteria assume that all the combined systems are independent and therefore approximate the probability of the joint events that compose the configuration as the product of the individual probabilities. It is important to remark that, in principle, this independence condition may not be met. A statistical test of independence as the *Chi-Square Test* could be used to validate this hypothesis. Finally, using the sum, there exist a balance between the combined distances where, if one of them is big it could be compensated by a small value on other one. When using this last option the normalization of the combined distances should be carefully addressed. It does not make sense to sum values that lie in different scale ranges.
Each of these options would give place to a family of functions $\mathcal{F}(d^*) : \mathbb{R}^K \to \mathbb{R}$. Given a particular choice for computing $d^*$, its corresponding family characterize all the configurations in the fusion space $\mathbb{R}^K$ that produce the same $d^*$ value. 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 trained *pdf* 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 in a subspace $\Omega_{d^*} \in \mathbb{R}^K$ containing all possible configurations that produce, at most $d^*$, according to the used criteria. This give place to the Equation 6.3:

$$P(\mathbf{D}_{i,j}|H_0) = \int_{\Omega_{d^*}} p_{H_0}(\mathbf{x}) d\mathbf{x}. \tag{6.3}$$

For instance, consider a two dimensional fusion scenario in which the obtained vector is $\mathbf{D}_{i,j} = \left(d^1_{i,j}, d^2_{i,j}\right)$. In this case the trained *pdf* complies $p_{H_0} : \mathbb{R}^2 \to \mathbb{R}$ and the different presented criteria would define regions of integration as shown on Figure 6.1.




(a) Max Rule ($d_{i,j_1}$)

(b) Min Rule ($d_{i,j_1}$)

(c) Product Rule

(d) Sum Rule

Figure 6.1: Integration region $\Omega_{d^*}$ for the different distances fusion rules

In each case, in red is shown the resulting family of functions $\mathcal{F}(d^*)$ according to the selected strategy for evaluating $\mathbf{D}_{i,j}$. In blue the obtained function for a particular value of $d^*$ is shown as a dashed line as well as the area of integration $\Omega_{d^*}$ that it defines. As a peculiarity, both Figures 6.1a and 6.1b represent the situation where both the *max* and *min* rules introduced previously considering that $d^1_{i,j}$ is the limiting distance. Similar regions will be obtained when the distance $d^2_{i,j}$ is the one defining $d^*$. In both cases, only one distance component will define the integration region according to the maximum or minimum value respectively. Two additional remarks are important in how the introduced fusion rules are applied in the *a-contrario* fusion strategy. First, when using the *Min* rule, it seems from Figure 6.1b that the integration region $\Omega_{d^*}$ is not bounded and therefore the computation of $P(\mathbf{D}_{i,j}|H_0)$ not possible. This assertion holds from the the-

oretical point of view, e.g. there exists infinite configurations $\mathbf{D}_{i,j}$ for which the minimum value is the same. In practice, the integration region is bounded by the maximum values the fusioned systems could produce when matching two biometric templates. The scale ranges of the systems being combined could be known in advance from the system development. If this information is not available, the system's output maximum values could be approximated by testing the system at hand with sufficient data. Once the region defined by the maximum values are known, one could assume that $p_{H_0}(\mathbf{x})$ will be null outside this region. As a second remark, the normalization of scores in the sum rule is necessary when a unique threshold is used over the obtained sum value. In the presented approach in which the sum rule concept is used to derive an integration region this normalization is not mandatory. In this case, the used *pdf* $p_{H_0}(\mathbf{x})$ is already trained in the scale ranges of each system being combined.

There is an important observation in how this operational approach could be applied in the case the biometric systems assign scores values instead of distances to the biometric comparisons. In this situation a conversion from scores to distances could be done just by inverting the score values with particular attention to the null scores. An alternative would be to simply invert the reasoning done with distances for deriving the rule to compute $P(\mathbf{D}_{i,j}|H_0)$. With this on mind, one simply needs to take the complementary region of integration defined before. This is shown in Figure 6.2 for the particular case of the sum fusion strategy where the vector of scores being assessed is $\mathbf{S}_{i,j} = \left(s^1_{i,j}, s^2_{i,j}\right)$.

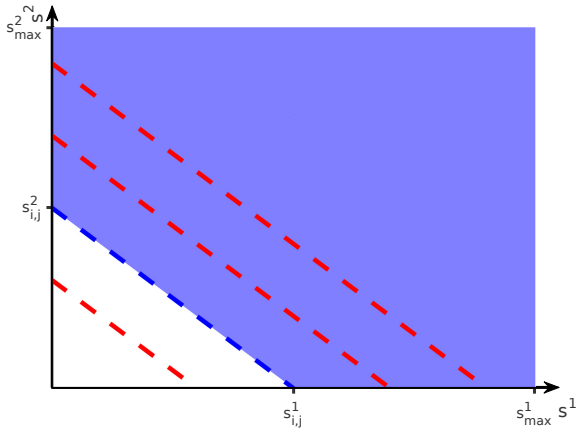



Figure 6.2: Integration region $\Omega_{d^*}$ for sum rule with scores

To better understand how these procedures are applied in practice, an example using the partition *BSSR1-Face* database is shown in Figure 6.3.

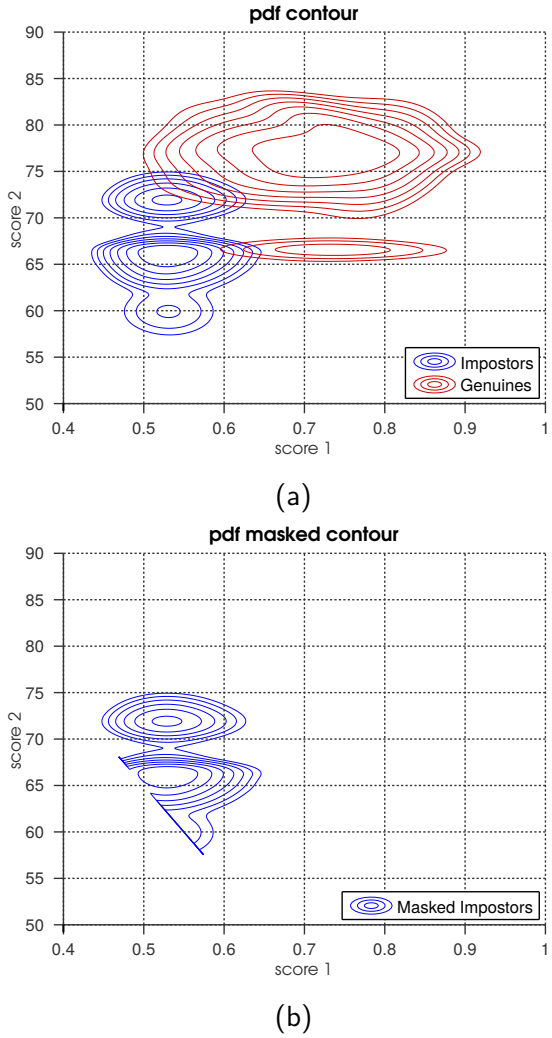



Figure 6.3: Example of masking the impostors *pdf* in *BSSR1-Face* database.6.3a Original distributions.6.3b Masked impostors distribution using *Sum* rule.

where the selected example point is $\mathbf{S}_{i,j} = (0.5, 65)$. The level lines of the original and masked distribution are represented in 6.3a and 6.3b respectively. Note that, in the last figure, only the masked impostors distribution is shown as is the only one used in the computation of $P(\mathbf{S}_{i,j}|H_0)$.

## 6.2 Experimental setup

The evaluation of the presented adaptation of the *a-contrario* framework to the fusion problem is done using the database *BSSR1* introduced in Section 3.1.1. This dataset was selected for its intrinsic features. As explained before, it was built targeting the evaluation of biometric fusion at different levels. This motivated its use

by various research groups whose reported results could be used as a benchmark. The obtained results by using the proposed fusion approach are compared with the ones obtained by using the *Likelihood-Ratio* approach presented in [12]. To the best of our knowledge, the strategy introduced in that work is the one in closer relation to the *a-contrario* based fusion. Both procedures works in a statistical way solving the problem by means of hypothesis testing.
The performance metrics used in the verification evaluation are the usual *FAR* and *GAR* introduced in Section 3.2.1. There is a particular characteristic of the *BSSR1* database that needs to be reviewed in this section as it affects directly how the evaluation protocol is done. The dataset only includes the scores of the comparisons between query and gallery samples (and not the actual samples images), therefore it only allows to estimate the background model by using the actual comparisons done when "on-production". This impose that any statistical approach should be done by using a *cross-validation* scheme. In this work we follow the same procedure used in the work considered as benchmark. The available data is used in a *2-fold cross-validation* scheme. The data partitioning is done by taking samples randomly in each of the training and testing sets in each performed experiment. To ensure that a particular data partition that favors one fusion strategy over the other does not occur, the *cross-validation* experiment is repeated $M$ times. Therefore, for each sub-partition of the selected database and each fusion strategy one ends up having a matrix of $GAR\,(\tau)$ values where each column vector $GAR(\tau)_{m,k}$ represents the obtained results for a particular experiment $m$ and fold $k$:

$$GAR(\tau) = (GAR(\tau)_{1,1} \quad \dots \quad GAR(\tau)_{m,k} \quad \dots \quad GAR(\tau)_{M,K})\,, \tag{6.4}$$

where $\tau$ 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 $\eta$ for the *Likelihood-Ratio* strategy and, finally, over the scores when each system is working individually.
Following the experimental setup in [12], 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 $\overline{GAR}\,(\tau)$ and its 95% confidence interval $[GAR_l\,(\tau)\,, GAR_u\,(\tau)]$. In order to obtain these, the *GAR* metrics are first referred to a common set of *FAR* values.
As the authors of the *Likelihood-Ratio* approach explain in their article [12], this strategy is highly dependent in having an accurate estimation of the underlying impostors and genuines classes distributions. Citing them: *"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."*
This statement is of great importance: it remarks the strongest flaw in the *Likelihood-Ratio* framework. While it is the most powerful statistical test (this is assured by means of the Neyman-Pearson theorem) its dependency on the genuines class information make its application difficult. This dependency in the information

available of the genuines class is very important. We already detailed the problems in obtaining such information in Section 2.4. They are summarized below:

- **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* database [48]), this is not the usual. This condition is even worse when considering citizen databases in which normally few samples per person are available.

- **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 introduced *a-contrario* approach only depends in the 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 different actions could be taken:

- **Distort the training samples:** Some type of noise could be added to the samples used in the estimation of the genuines class distribution. This alteration could be done at different levels: the noise could be added to the input samples images or to the results after the comparison between two biometric representatives.

- **Reduce the number of available training samples:** The amount of genuine samples used in training their inherent distribution could be reduced. For this, one could define a sample ratio $SR$ as the rate of genuines training samples that are used from all the available ones. For a particular sample ratio value the corresponding employed representatives are selected randomly.

From these two options, the first one presents some difficulties in its implementation. First, in order to distort the input samples one would need access to those images, this is clearly not possible when using the *BSSR1* database in which only the comparison scores are available. Secondly, if one chooses to alter the obtained scores, the type and amount of noise to be added has to be adjusted to the variations that could result in a real-life scenario and this is not known before-hand.

The second alternative is more plausible and could be directly applied. It clearly simulates a common scenario when working with citizen databases. In this situation, one usually have a considerable amount of people, all the population that has been enrolled in their first id document, for which only one biometric sample is available. And therefore a genuine comparison could not be done in these cases. We then choose to test both the *a-contrario* and *Likelihood-Ratio* fusion approaches by varying the genuines sample ratio $SR$. Note that, for the application of the *a-contrario* strategy $SR$ could even be null but this particular setup makes the *Likelihood-Ratio* framework infeasible, thus a minimum value $SR_{min} > 0$ is used as limit. 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 6.1.

| **SR** | **BSSR1-Face** | **BSSR1-Fingerprint** |
|---|---|---|
| 0.01 | 15 | 30 |
| 0.05 | 75 | 150 |
| 0.1 | 150 | 300 |
| 0.3 | 450 | 900 |
| 0.7 | 1050 | 2100 |
| 1 | 1500 | 3000 |

Table 6.1: Genuines training samples

## 6.3 Theoretical example

As a first step in the experimental evaluation of the proposed strategy, we present in this section the results obtained using artificially generated data. This allows the analysis, in a simple and controlled scenario, of how the technique and other *state-of-the-art* strategies perform. And how these are affected with respect to data characteristics and the training of classes being done.

### 6.3.1 Generated data

In order to keep the experiment as simple as possible, we simulate a 2-classifiers fusion scheme. Both classes scores distributions are assumed to be Gaussian giving place to the probability densities 6.5 and 6.6 for the impostors and genuine respectively.

$$p_{H_0}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma_{H_0}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{H_0})^T \Sigma_{H_0}^{-1}(\mathbf{x}-\mu_{H_0})} \tag{6.5}$$

$$p_{H_1}(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma_{H_1}|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{H_1})^T \Sigma_{H_1}^{-1}(\mathbf{x}-\mu_{H_1})} \tag{6.6}$$

The values for the mean and covariance matrix have been set arbitrary for the impostors and genuine classes respectively:

$$\mu_{H_0} = \begin{pmatrix} 0.5 \\ 70 \end{pmatrix}, \qquad \Sigma_{H_0} = \begin{pmatrix} 0.01 & 0 \\ 0 & 25 \end{pmatrix}, \tag{6.7}$$

$$\mu_{H_1} = \begin{pmatrix} 0.7 \\ 80 \end{pmatrix}, \qquad \Sigma_{H_1} = \begin{pmatrix} 0.01 & 0 \\ 0 & 15 \end{pmatrix}. \tag{6.8}$$

A representative set of the samples generated with such distributions is represented in Figure 6.4.

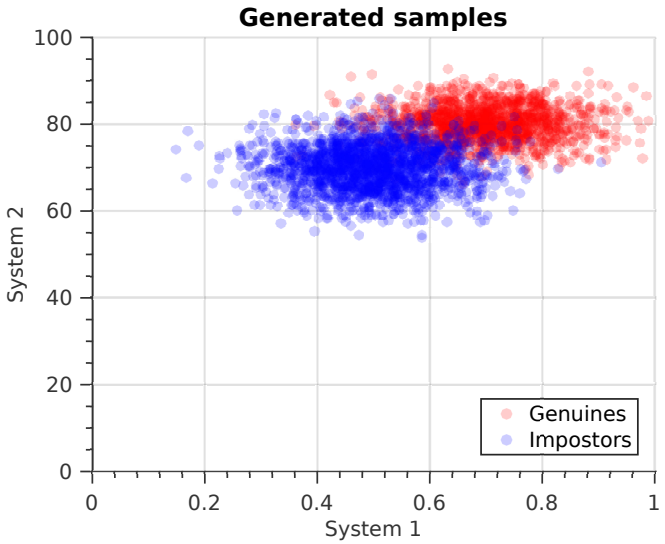



Figure 6.4: Generated data samples

### 6.3.2 Experimental evaluation

In order to perform the experimental evaluation, a dataset of $N = 3000$ identities is generated. Therefore, we have $N$ and $N(N-1)$ matches corresponding to the genuine and impostors classes respectively.
The different presented approaches for verification fusion require the estimation of the classes distributions. Therefore, we execute the evaluation by splitting randomly the available samples in training and testing datasets. As explained in the reference article [12], the *Likelihood-Ratio* approach ensures the best possible fusion performance when the underlying classes distributions $p_{H_0}(\mathbf{x})$ and $p_{H_1}(\mathbf{x})$ are known. But in practice, we only have estimations $\widetilde{p_{H_0}}(\mathbf{x})$ and $\widetilde{p_{H_1}}(\mathbf{x})$ of them, and the performance of the *Likelihood-Ratio* test will depend on the accuracy of such estimations. While the *LR*-based strategy requires modeling both genuine and impostors classes, the *a-contrario* approach only requires the information provided by the impostors' distribution. As mentioned before, this is the great advantage of using *a-contrario* models in biometric applications. In order to obtain genuine representatives for modeling the corresponding distribution, one needs to count

with multiple samples of each person. Meanwhile, for modeling the impostors' distribution one only needs samples from different people. In order to simulate the possible lack of genuine representatives, we perform the benchmarking using a reduced amount of genuine's training samples. We achieve this by using a sample ratio coefficient ($SR$) with which genuine training samples are sampled correspondingly. The distributions are approximated by *Gaussian Mixture Models* ($GMM$), in particular the implementation in [43] was used. The following trained genuine's distributions parameters were experimentally obtained for $SR$ values 0.1 and 1.

### Training results for $SR = 1$ of genuine training samples

In this case, the mixture includes only one Gaussian distribution. Both mean and covariance matrix,$\widetilde{\mu_{H_1,SR=1}}$ and $\widetilde{\Sigma_{H_1,SR=1}}$ are very similar to the ground-truth parameters. This is to be expected considering that all available genuine training samples are used. The corresponding level lines of both the impostors and genuine's distributions are shown in Figure 6.5.

$$\widetilde{\mu_{H_1,SR=1}} = \begin{pmatrix} 0.697 \\ 79.979 \end{pmatrix}$$

$$\widetilde{\Sigma_{H_1,SR=1}} = \begin{pmatrix} 0.009 & 0.015 \\ 0.015 & 16.407 \end{pmatrix}$$

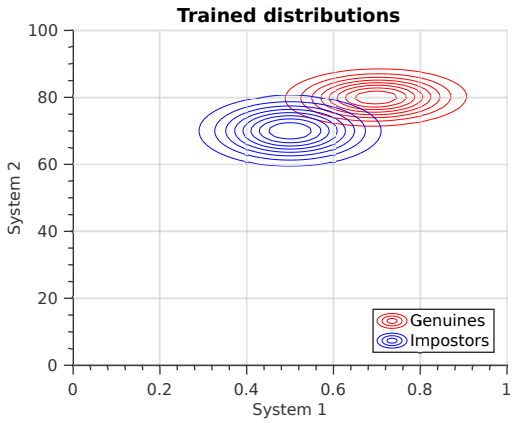



Figure 6.5: Impostors and genuine estimated distributions

### Training results for $SR = 0.1$ of genuine training samples

In this scenario, the obtained $\widetilde{p_{H_1}}(\mathbf{x})$ consist of the mixture of three Gaussians distributions each with its corresponding mean $\widetilde{\mu^i_{H_1,SR=0.1}}$, covariance matrix $\widetilde{\Sigma^i_{H_1,SR=0.1}}$ and weight $\omega^i$ in the mixture for $i = 1 \ldots 3$. The estimation obtained is very inaccurate in approximating the ground-truth distribution from which the data was sampled. This is due to the fact that very few training samples are available. The obtained distribution is shown in Figure 6.6.

$$\widetilde{\mu^1_{H_1,SR=0.1}} = \begin{pmatrix} 0.750 \\ 85.884 \end{pmatrix}$$

$$\widetilde{\mu^2_{H_1,SR=0.1}} = \begin{pmatrix} 0.707 \\ 82.237 \end{pmatrix}$$

$$\widetilde{\mu^3_{H_1,SR=0.1}} = \begin{pmatrix} 0.684 \\ 78.215 \end{pmatrix}$$

$$\widetilde{\Sigma^1_{H_1,SR=0.1}} = \begin{pmatrix} 0.003 & -0.023 \\ -0.023 & 0.340 \end{pmatrix}$$

$$\widetilde{\Sigma^2_{H_1,SR=0.1}} = \begin{pmatrix} 0.008 & -0.087 \\ -0.087 & 1.004 \end{pmatrix}$$

$$\widetilde{\Sigma^3_{H_1,SR=0.1}} = \begin{pmatrix} 0.017 & 0.259 \\ 0.259 & 6.007 \end{pmatrix}$$

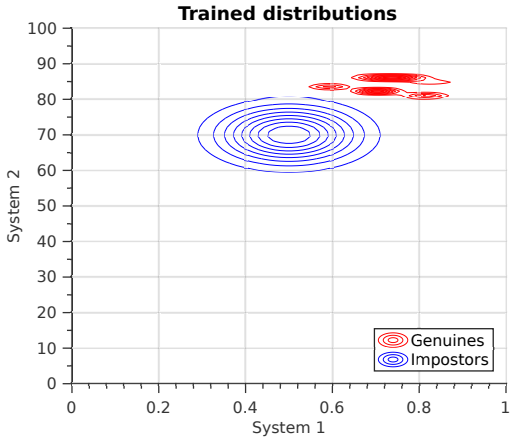



Figure 6.6: Impostors and genuine estimated distributions

$$\omega^1 = 0.415 \quad \omega^2 = 0.404 \quad \omega^3 = 0.180$$

## Testing results

The models trained in each case were used with both the *Likelihood-Ratio* and *a-contrario* approaches on the testing data partition. We also include the obtained performance when the ground-truth genuine's distribution parameters are used. The obtained results are shown in Figures 6.7 and 6.8 for $SR = 1$ and $SR = 0.1$ respectively.

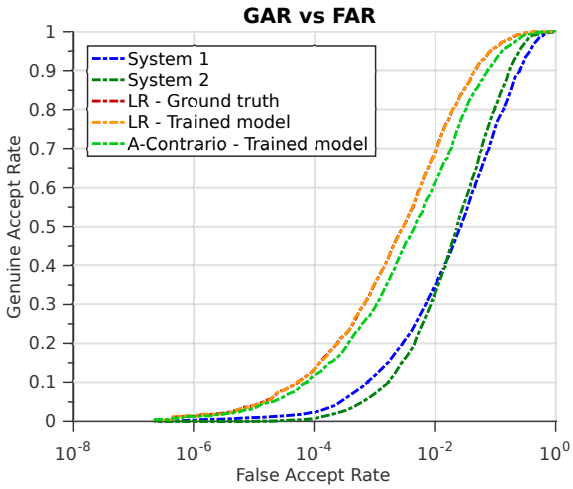



Figure 6.7: Verification fusion performance for $SR = 1$

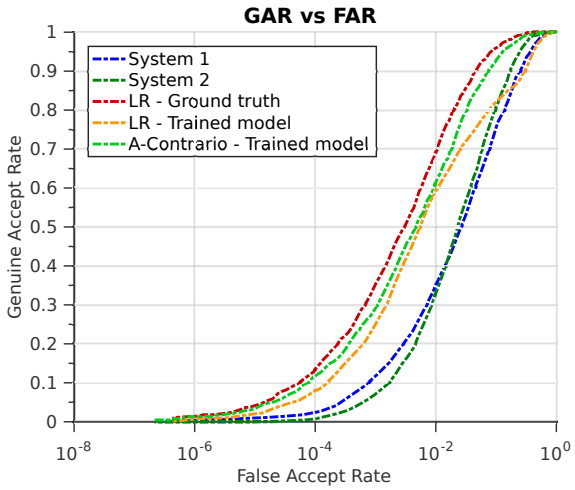



Figure 6.8: Verification fusion performance for $SR = 0.1$

In the first case, 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 obtained with the ground-truth information which is to be expected considering the results of the training shown previously when $SR = 1$. On the other hand, as the sample ratio is decreased, and the accuracy of the probability densities lowered, the obtained performance with the *Likelihood-Ratio* strategy gets worse. In this case, the *a-contrario* approach continues to work equally good as before. These results confirm the claims made by the authors in [12] and show the robustness of the proposed approach with respect to genuine's class available training data.

## 6.4 Experimental evaluation on real datasets

### 6.4.1 Systems individually

The presented *a-contrario* technique has already been tested on systems working individually on Chapter 4. But, the *Likelihood-Ratio* approach has not been tested, there is actually no theoretical limitation on applying this strategy in a *1-D* scenario. The functioning of the algorithm would be the same, with the unique difference that the probability density functions $p_{H_0}(x)$ and $p_{H_1}(x)$ would be one-dimensional and we will have the two of them for each particular system being evaluated. Additionally, they are now evaluated on single values of score/distance $x$ instead of a vector $\mathbf{x}$.
The main difference with the evaluation done using the *a-contrario* approach in this scenario is that the training of the functions cannot be done online as explained in Section 4.1.2. In the former case, as no modeling of the genuine class is necessary, one can take all the available distance/scores values obtained when

comparing a query input against the gallery set and use them to assess if the particular value being evaluated is meaningful/rare under this background information. When the *Likelihood-Ratio* technique is used, this is not possible as strictly one needs to know which value will correspond to the genuine class which is not known before-hand. Additionally, if only the information of one query sample comparison is available, this will not be sufficient to correctly model the genuine distribution. Indeed only one genuine training sample can not characterize correctly such distribution. Following these considerations, in each corresponding evaluated system the available samples were randomly split into training and testing subsets.
Before proceeding with the fusion results, we present as an example the results obtained when the *Likelihood-Ratio* approach is applied individually over each system of the *BSSR1-Face* dataset. In all cases the underlying *pdf* functions were estimated using the *GMM* implementation in [43].

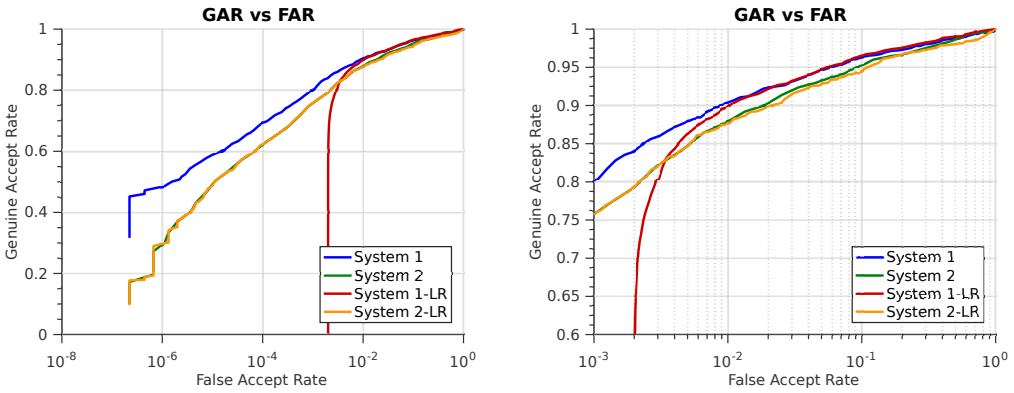



Figure 6.9: *Likelihood-Ratio* verification performance in *BSSR1-Face* dataset. 6.9a *GAR* vs *FAR* plot. 6.9b *GAR* vs *FAR* , zoomed plot for $FAR = [10^{-3}, 1]$ range.

Several conclusions could be drawn from the results depicted in Figures 6.9a and 6.9b:

- **A marginal improvement is obtained with the *Likelihood-Ratio* approach**
  While there is an improvement when *LR* is used in both face recognition systems, this improvement is marginal with respect to the performance obtained by each system without applying *LR*. This can be explained by two reasons. First, no additional information is included in the classification by means of using the *LR* approach as it happens when one is fusing multiple systems. Secondly, as explained in [9] the distance/score returned by a biometric system for a particular match between samples is proportional to the likelihood of that comparison to belong to the impostors/genuine classes. The algorithms included in a system are developed and trained in order to have a relation between the output value and corresponding class as accurate as possible. By applying the *LR* strategy, we are mapping the obtained

output scores into likelihoods for each class, using the estimation of them done on the training data. If the classification being used is already good, there is no guarantee that this remapping will introduce any improvement.

- **The application of LR may introduce some practical problems**
  From both figures it is clear that there is a problem with the obtained metrics for *System 1* when $LR$ is applied. There exists an abrupt decrease in the $GAR$ index and a minimum value of $FAR = 0.002$ is obtained. This does not occur when the performance of the system is assessed without $LR$. This happens due to the fact that a considerable number of impostors samples in the testing set (9044 in total) obtain a null probability of belonging to the impostors class. Therefore, they achieve an infinite $LR$ ratio. This is a practical consequence of the training of the *pdfs* being used. If no training distances are lower (or near) than the query one a null probability associated with the impostors class could be obtained. While this is not a theoretical problem of the approach, it is a practical issue that should be addressed when statistical strategies as this one are applied.

### 6.4.2 Fusion results

#### Face datasets

The results obtained in the *BSSR1-Face* datasets for the different used sample ratios of the used training genuines samples are shown on Figures 6.10 and 6.11

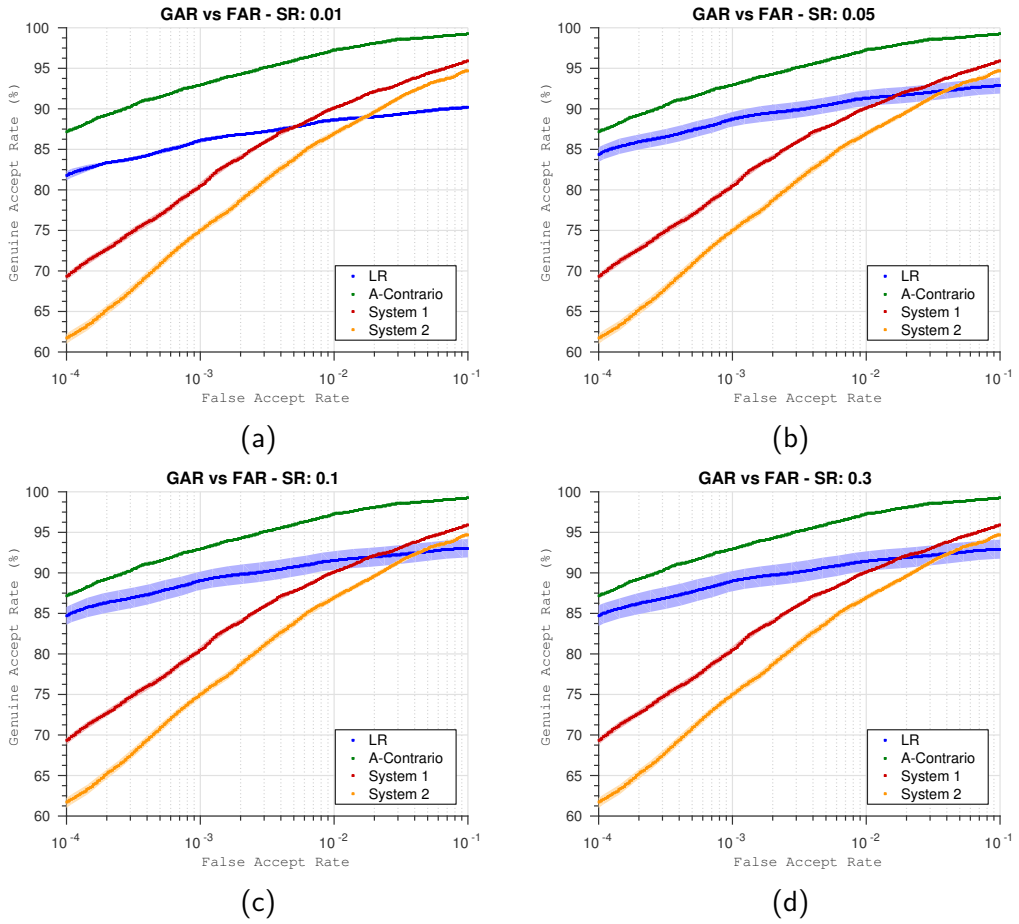
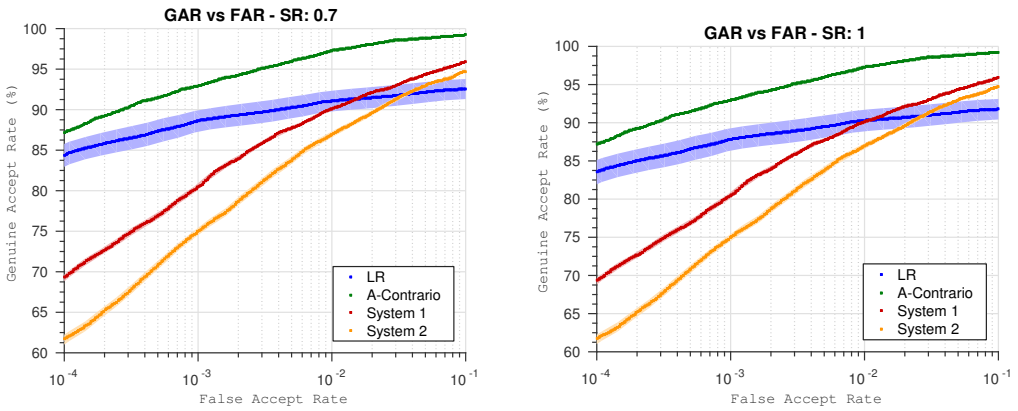



Figure 6.10: Fusion results in *BSSR1-Face* database for different sample ratios of genuine's training samples.6.10a $SR = 0.01$, 6.10b $SR = 0.05$, 6.10c $SR = 0.1$, 6.10d $SR = 0.3$.

Several conclusions from the obtained results are presented below.

- **Both fusion approaches reach greater performance than the one obtained using the systems individually.**
  Besides being the expected result, this is a good validation of the presented strategies. Both of them exploit the statistical information in the combination of both results and achieve greater performance than the one obtained with each system separately. The only exception is obtained with the $LR$ approach for $FAR$ values greater than $10^{-2}$. This should not be a concern as using as a working point one with a $FAR$ above this value is not common in practice.




(a) (b)

Figure 6.11: Fusion results in *BSSR1-Face* database for different sample ratios of genuine's training samples.6.11a $SR = 0.7$, 6.11b $SR = 1$.

- **The *LR* approach is very robust, even when few genuine training samples are available**
  The obtained results in the ideal scenario in which all the available genuine training samples are used ($SR = 1$) is maintained until very few training samples are available. The decrease in performance just becomes notorious when $SR = 0.01$, corresponding to the use of 15 of the original 1500 training samples. And even in this case, the performance of the method is still better than the one obtained by the systems individually for a great range of $FAR$ values.

- **The *a-contrario* approach performance presents almost no variations across the experiments.**
  It can be seen that the confidence interval for the *a-contrario* strategy is very narrow. This indicates that the obtained results remain stable on the 20 experiments performed. This could be explained by the fact that the used background model is trained over the impostors' samples. As the number of these samples is considerably more than the number of genuine training samples, the trained model remains stable despite changes of the impostors' data of each particular experiment. The $LR$ approach suffers from these variations in a greater manner due to its dependence on the genuine class training samples in addition to the impostors' ones for the estimation of both classes associated *pdfs*.

# Chapter 7

# Conclusions and future work

## 7.1 Conclusions

Several conclusions on different subjects could be drawn from the present work. First, it is remarkable how important the obtained distances (or scores) of a biometric system are. In these lies the complete information that allows for a correct classification when it is correctly analyzed and interpreted. Normally, the introduced advances in the biometrics field are targeted at the feature extraction stage in order to obtain better representations that would benefit the separation between impostors and genuine comparisons. It is true that these improvements will finally have an impact on the obtained matching scores. After these outputs are obtained usually just a threshold is used for classification without too much effort in their analysis. In this work, we show how the statistical information contained on these scores could be exploited by means of simple strategies. This becomes even more important when we consider the great use of biometrics in non-academic environments. In these scenarios normally closed systems that work as a black-boxes are used. In these cases, the only available information to its operators is the obtained output scores. Then, it becomes crucial to exploit as much as possible this information.
Second, the adaptation of the *a-contrario* framework in the different biometrics applications analyzed in this thesis shows how well-known problems could benefit from a paradigm shift. Both verification and identification problems are normally tackled by controlling the two type of errors that can occur in the classification of a match between samples. The false positives and false negatives are obtained when a match is considered genuine while it is an impostor and vice-versa, respectively. Usually, both impostors and genuine classes are modeled in order to control at the same time these two errors categories. While different statistical approaches to the classification problem use effectively the impostors' data as *SVM* and *Likelihood-Ratio* [12], they still highly rely also on the genuine's information. By using an *a-contrario* based strategy we show how the different biometrics use-cases could benefit and gain robustness even in the extreme case where no genuine's information is available.

Third, from the results obtained in this thesis, it is possible to confirm the importance of fusion in the biometrics field. This is not a novel conclusion and several works have already reported the increase in performance when multiple information sources are used together. But, this work helps to visualize how large performance improvements can be obtained by using simple techniques. The fusion schemes introduced in [9] do only require applying simple operators to the scores of the systems being combined. The other approaches used in the work need to model the impostors and genuine classes but are, nevertheless, very straightforward to apply.
Finally, in this work, we explore the importance of implementing response reliability techniques in biometric identification systems. As introduced in Section 5.1, this type of control is not usually implemented in practice. And it is highly important as identification systems are more and more used automatically. As explained previously a global performance measure obtained in the development of the system does not give a clear insight of the correctness of the identifications being done for each particular query individual. In this work, we presented a different strategy to tackle this problem and showed how an *a-contrario* based approach could provide a good solution.

## 7.2 Future work

Different lines of future work could be followed. First, as in every work in the biometrics field, the experimental setup in this thesis could be extended by considering larger datasets with higher dimensions. By using more samples, both in gallery and query datasets, the performance of the different presented approaches could be better understood. In theory, all the presented strategies should scale well when the data size is incremented, but this should be verified experimentally. Regarding the use of higher dimensional data, in this work we explore the fusion in a two-dimensional scenario. Higher dimensional data should improve results as more information of the comparison between two biometric samples would be available. But, in order to use such information, the presented *a-contrario* fusion approach should be adapted. In particular, the frontiers introduced in Section 6.1 should be updated accordingly to the dimension of the data being used.
Second, in Section 4.1 different alternatives for estimating the background model used in the *a-contrario* application to the biometric verification are presented. Recapitulating briefly: they are divided between *pre-computed* and *computed in classification time* strategies. In the experimental evaluation, it is concluded that the only one providing improvements with respect to the usual approach is the one modeling the background model in classification time. But this increases the classification time and computation requirements. Each input must be compared, not only against the declared identity corresponding sample but also to all the other ones in the gallery in order to estimate the background model. This does not represent an issue in the identification process where each input sample is compared against all enrolled samples anyways, but it is an extra load in the verification

scenario. Different alternatives could be explored in order to alleviate this extra work. For instance, one could use a sample ratio of extra comparisons realized for estimating the background model. Comparing the input against a subset of the enrolled identities one could obtain an approximation of the model less accurate but, perhaps, sufficient for applying the *a-contrario* strategy. Another possible improvement would lie in the clusterization of the distances obtained when the input is compared against all gallery representatives. In this way, the model could be estimated more efficiently by using a subset of the comparisons representing the resulting clusters.
Third, in Section 6.1 several alternatives for the definition of the criteria used in a biometric fusion context were explored. In this thesis we only evaluate the *sum* rule considering that it is the one recommended in the literature [9]. The evaluation of the remaining rules remains as future work.
Finally, another possible way for extending this work lies in the exploration of alternative outliers detection approaches. These could be used to assess when a particular match between two biometric samples should be considered meaningful. For instance, *One-Class SVM* [49] has been successfully used for outliers detection.

This page intentionally left blank

# Bibliography


[1] 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*, 2014, pp. 1–8.

[2] "How facial recognition works in xbox kinect." [Online]. Available: https://www.wired.com/2010/11/how-facial-recognition-works-in-xbox-kinect/

[3] "Iphone x facial recognition: hands-on with face id, the biggest feature in apple's new phone." [Online]. Available: http://www.independent.co.uk/life-style/gadgets-and-tech/features/iphone-x-facial-recognition-faceid-hands-on-feature-review-apple-new-phone-a8028851.html

[4] "Amazon's Alexa can now recognize different voices and give personalized responses." [Online]. Available: https://www.theverge.com/circuitbreaker/2017/10/11/16460120/amazon-echo-multi-user-voice-new-feature

[5] "Using ePassport gates at airport border control." [Online]. Available: https://www.nidirect.gov.uk/articles/using-epassport-gates-airport-border-control

[6] "Estrenaron las cámaras de reconocimiento facial en el Estadio Centenario." [Online]. Available: http://www.subrayado.com.uy/Site/noticia/65597/estrenan-hoy-las-camaras-de-reconocimiento-facial-en-el-estadio-centenario

[7] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*, ser. SpringerLink : B{ü}cher. Springer US, 2011. [Online]. Available: https://books.google.fr/books?id=ZPt2xrZFtzkC

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

[9] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

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

[11] A. K. Jain, K. Nandakumar, and A. A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognition*, vol. 38, pp. 2270–2285, 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320305000592

[12] 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*, vol. 30, no. 2, pp. 342–347, 2008.

[13] 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*, vol. 5, 2004.

[14] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems." *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 450–455, mar 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15747798

[15] 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*, vol. 5, no. 3, pp. 470–479, 2010.

[16] A. Desolneux, L. Moisan, and J. M. Morel, *From Gestalt Theory to Image Analysis: A Probabilistic Approach*, 1st ed. Springer Publishing Company, Incorporated, 2007.

[17] W. Metzger, *Gesetze des Sehens*, ser. Senckenberg-Buch ; 53 [i.e. 54]. Kramer, 1975. [Online]. Available: https://books.google.com.uy/books?id=jOY3AQAAIAAJ

[18] A. Desolneux, L. Moisan, and J. M. Morel, "Maximal meaningful events and applications to image analysis," *The Annals of Statistics*, vol. 31, no. 6, pp. 1822–1851, 2003. [Online]. Available: http://projecteuclid.org/euclid.aos/1074290328

[19] A. Desolneux, L. Moisah, and J. M. Morel, "A grouping principle and four applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 4, pp. 508–513, apr 2003. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1190576

[20] R. Grompone von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall, "LSD A Fast Line Segment Detector with a False Detection Control," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 722–732, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=4731268

[21] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J. M. Morel, "An A Contrario Decision Method for Shape Element Recognition," *International Journal of*

*Computer Vision*, vol. 69, no. 3, pp. 295–315, apr 2006. [Online]. Available: http://link.springer.com/10.1007/s11263-006-7546-0

[22] J. Rabin, J. Delon, and Y. Gousseau, "A contrario matching of SIFT-like descriptors," *2008 19th International Conference on Pattern Recognition*, pp. 1–4, dec 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4761371

[23] W. Feller, *An introduction to probability theory and its applications.* John Wiley & Sons, 2008, vol. 2.

[24] A. Desolneux, L. Moisan, and J. M. Morel, "Meaningful alignments," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 7–23, 2000.

[25] R. Grompone von Gioi, J. Jakubowicz, J. M. Morel, and G. Randall, "On Straight Line Segment Detection," *Journal of Mathematical Imaging and Vision*, vol. 32, pp. 313–347, 2008.

[26] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses.* Springer Science & Business Media, 2006.

[27] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[28] C. E. Bonferroni, *Teoria statistica delle classi e calcolo delle probabilita.* Libreria internazionale Seeber, 1936.

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

[30] C. Watson, "Nist Special Database 14, Fingerprint Database." *US National Institute of Standards and Technology*, 1993.

[31] R. Cappelli, M. Ferrara, and D. Maltoni, "[2010][12]Minutia Cylinder-Code A New Representation and Matching Technique for Fingerprint Recognition.pdf," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2128–2141, 2010.

[32] J. N. Bradley, C. M. Brislawn, and T. Hopper, "FBI wavelet/scalar quantization standard for gray-scale fingerprint image compression," *Visual Information Processing*, vol. 2, pp. 293–304, 1993. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1012727

[33] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *2003 IEEE International SOI Conference. Proceedings (Cat. No.03CH37443)*, oct 2003, pp. 44–.

[34] A. K. Jain and S. Z. Li, *Handbook of Face Recognition.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[35] M. De Marsico, M. Nappi, D. Riccio, and H. Wechsler, "Robust Face Recognition for Uncontrolled Pose and Illumination Changes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 43, no. 1, pp. 149–163, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6196234

[36] M. De Marsico, M. Nappi, D. Riccio, and G. Tortora, "NABS: Novel Approaches for Biometric Systems," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, no. 4, pp. 481–493, jul 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5551235

[37] N. Poh and S. Bengio, "Improving fusion with margin-derived confidence in biometric authentication tasks," *Audio- and Video-Based Biometric Person Authentication*, pp. 474–483, 2005. [Online]. Available: http://link.springer.com/chapter/10.1007/11527923{_}49

[38] L. D. Martino, A. Fernandez, R. Grompone 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)*, sep 2016, pp. 1–8.

[39] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. A. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," *National Institut of Standards and Technology Gaithersburg*, pp. 1–4, 1998.

[40] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26.

[41] G. McLachlan and D. Peel, *Finite mixture models*. John Wiley & Sons, 2004.

[42] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, "Kernel density estimation via diffusion," *Ann. Statist.*, vol. 38, no. 5, pp. 2916–2957, 2010. [Online]. Available: http://dx.doi.org/10.1214/10-AOS799

[43] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, 2002.

[44] E. Tabassi, C. Wilson, and C. Watson, "Nist fingerprint image quality," *NIST Res. Rep. NISTIR7151*, pp. 34–36, 2004.

[45] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[46] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[47] Biometric System Laboratory, "Fingerprint Verification Competition," p. 1, 2013. [Online]. Available: https://biolab.csr.unibo.it/fvcongoing/UI/Form/Home.aspx

[48] 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," University of Massachusetts, Amherst, Tech. Rep., 2007.

[49] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001. [Online]. Available: https://doi.org/10.1162/089976601750264965

This page intentionally left blank

# List of Tables

This page intentionally left blank

# List of Figures

Esta es la última página.
Compilado el Friday 29th December, 2017.
http://iie.fing.edu.uy/