# SECURE MULTIBIOMETRIC SYSTEMS

By
Emanuela Marasco

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
UNIVERSITY OF NAPLES FEDERICO II
VIA CLAUDIO 21, 80125 - NAPLES, ITALY
NOVEMBER 2010

I'm not young enough to know everything.

*Wilde*

# Acknowledgements

*I have been privileged to have Prof. Carlo Sansone as my advisor. His educational guidance, constant suggestions and his attention to details have inspired me to give my best time. I also appreciate very much the important opportunities he created for collaborations with different research groups working on the same topic. This made the Ph.D. experience a memorable one.*

*Special thanks to Dr. Arun Ross, West Virginia University, for allowing an important cultural exchange and providing a very healthy research environment. His lectures on pattern recognition, advanced biometrics, machine learning inspired my research; and several seminars, weekly personal meeting, discussions with his team, international conferences he organized, motivated me and helped my understanding and growth.*

*I am very grateful to Dr. Stephanie Schuckers, Clarkson University, for giving the chance to visit and work in her laboratory to hone my research skills. Thanks to her efforts in ensuring that I spend a productive period.*

*Thanks to Dr. Thrimahos Bourlai for his daily useful talks and suggestions during my staying at WVU. Thanks to Dr. Josef Kittler for his several papers which motivated and inspired me during my PhD studies. Thanks to Dr. Norman Poh for his explanations about data. Thanks to Dr. Gianluca Marcialis for his support and for sharing data. Thanks to Dr. Ayman Abaza for replying to all my questions. Thanks to Simona, Ajita, Raghu, Raghav, Aglika, Neeru, Oghe, Peter, Asem, and Manisha, Citer laboratory WVU; thanks to Flora, Christian, Francesco, Claudio, Carlos, Antonio, Fanny, Francesca, Porfirio, University of Naples; thanks to Owen, Peter, Phiona, Joh, Laura, Clarkson University, for collaborating and for their daily suggestions.*

*I would like to thank my parents, brother, and relatives for their incredible love, prayers, and support; and Ale for his encouragement. I dedicate any progress to them. Finally, a special appreciation to all my colleagues who everyday feel the exigence to know, to understand, to freely think, guided by their passion, even if sometimes this means to accept very precarious conditions. It is to them that I dedicate this thesis to.*

*Naples, Italy* *Emanuela*
*November 28, 2010*

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Although the market for biometric technologies is expanding, the existing biometric systems present still some issues that the research community has to address. In particular, in adverse environmental conditions (e.g., low quality biometric signals), where the error rates increase, it is necessary to create more robust and dependable systems. In the literature on biometrics, the integration of multiple biometric sources has been successfully used to improve the recognition accuracy of the unimodal biometric systems. Multibiometric systems, by exploiting more information, such as different biometric traits, multiple samples, multiple algorithms, make more reliable the biometric authentication. Benefits of multibiometrics depend on the diversity among the component matchers and also, on the competence of each one of them. In non-controlled conditions of data acquisition, there is a degradation of biometric signal quality that often causes a significant deterioration of recognition performance. It is intuitive the concept that, the classifier having the higher quality is more credible than a classifier operating on noisy data. Then, researchers started to propose quality-based fusion schemes, where the quality measures of the samples have been incorporated in the fusion to improve performance. Another promising direction in multibiometrics is to estimate the decision reliability of the component modality matcher

based on the matcher output itself. An interesting open research issue concerns how to estimate the decision reliability and how to exploit this information in a fusion scheme. From a security perspective, a multimodal system appears more protected than its unimodal components, since spoofing two or more modalities is harder than spoofing only one. However, since a multimodal system involves different biometric traits, it offers a higher number of vulnerable points that may be attacked by a hacker who may choice to fake only a subset of them. Recently, researchers investigated if a multimodal system can be deceived by spoofing only a subset but not all the fused modalities. The goal of this thesis is to improve the performance of the existing integration mechanisms in presence of degraded data and their security in presence of spoof attacks. Our contribution concerns three important issues: 1) Reducing verification errors of a fusion scheme at score level based on the statistical Likelihood Ratio test, by adopting a sequential test and, when the number of training samples is limited, a voting strategy. 2) Addressing the problem of identification errors, by setting up a predictor of errors. The proposed predictor exploits ranks and scores generated by the identification operation and can be effectively applied in a multimodal scenario. 3) Improving the security of the existing multibiometric systems against spoof attacks which involve some but not all the fused modalities. Firstly, we showed that in such a real scenario performance of the system dramatically decrease. Then, for the fingerprint modality, we proposed a novel liveness detection algorithm which combines perspiration- and morphology-based static features. Finally, we demonstrated that, by incorporating our algorithm in the fusion scheme, the multimodal system results robust in presence of spoof attacks.

Quelli che s'innamoran di pratica sanza scienza son come il nocchiere, ch'entra in navilio sanza timone o bussola, che mai ha certezza dove si vada.

*Leonardo*

# Chapter 1

# Introduction

## 1.1 Biometric Recognition

The authentication process determines or verifies the identity of an individual. It assumes significant importance in high security applications, such as *logical* access to personal computers, cellular phone, ATMs, or *physical* access to buildings, border crossing [63]. Traditionally, to ensure that only authorized users access to the protected services, *possession-based* (badges) or *knowledge-based* (passwords) solutions have been adopted. However, when a password is divulged to an unauthorized user or a badge is stolen by an impostor, these authentication schemes may be deceived. Vulnerabilities of such schemes are being addressed by the emergence of biometric systems which establish the identity of an individual based on what the person *is*, rather than what the person *carries* or *remembers* [45]. The identity of an individual is encoded by different biometric traits, such as fingerprints, hand geometry, iris, retina, face, hand vein, facial thermograms, signature, voiceprint, gait, palmprint, referred to as biometric *modalities* (see Fig.1.1).

The biometric recognition process involves firstly the acquisition of biometric data and

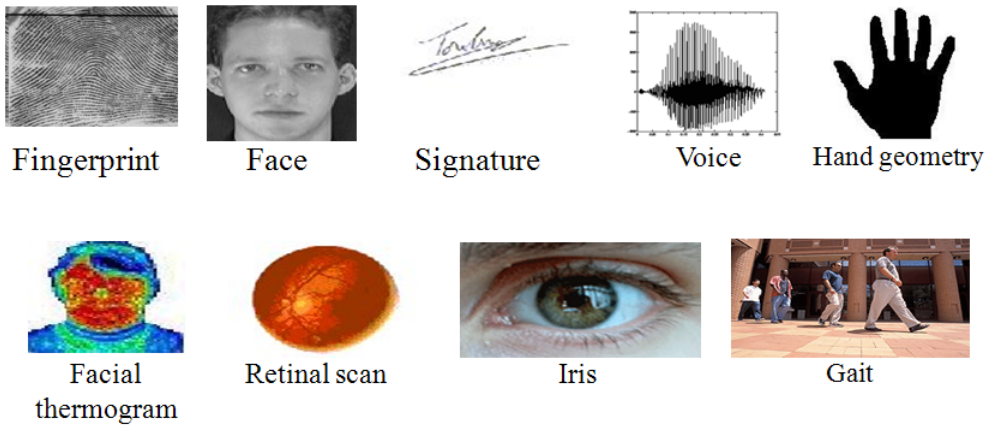


Figure 1.1: Examples of some of the biometric traits used for authenticating an individual.

the extraction of features from the acquired data, then the comparison of these features to the feature set previously stored in the database, referred to as *template*.

A biometric system may operate in the two modes:

- Biometric identity verification
- Biometric identification

In the verification mode, the system has to verify the authenticity of a claimed identity, while in the identification mode it has to assign the correct identity label to one person out of a *watch-list* [32].

A typical biometric system is composed by four main modules:

1. Sensor Module, which defines the interaction of an individual with the system by capturing his biometric data.
2. Feature Extraction Module, which extracts feature values from the acquired data. During enrollment, the extracted feature set, referred to as *template*, is stored in the

database and it represents the identity of a subject.

3. Matching Module, which compares the feature vector extracted from the query to the template. The match score determines the amount of similarity (similarity score) or distance (distance score) between the feature set of enrolled template and the query data. The matching is 1:1 to verify a claimed identity, while it is 1:N to determine an identity.

4. Decision-making Module, which accepts or rejects the user's claimed identity based on the matching score generated in the matching module, in the verification task or declares the user's identity based on the best match score, in the identification task (see Fig.1.5).

### 1.1.1 Performance Evaluation

The feature set extracted from the probe biometric data does not exactly correspond to the template, and subsequently, the matching process is never perfect. This variation may be due to several factors such as *non-controlled* sensing conditions, changes of the biometric characteristic, etc. This aspect clearly impacts on the performance of a biometric system which never achieves a perfect recognition where the accuracy is 100%.

Two types of errors can be made by a biometric verification system:

- *TypeI*, False Rejection. It occurs when an authorized user is wrongly rejected by the system.

- *TypeII*, False Acceptance. It occurs when an impostor is wrongly accepted by the

system. This error is very costly.

Performance is evaluated in terms of False Acceptance Rate (FAR) and False Rejection Rate (FRR), where FAR is usually fixed by the specific application.

The two errors are complementary, trying to lower one of them by varying the threshold, the other error rate automatically increases. Biometric verification looks for the best *trade-off* between these two types of errors. The Equal Error Rate (EER) point is obtained when FAR and FRR coincide. The complete performance curve which represents the full capabilities of the system at different operating points, is given by the Receiver or also Relative Operating Characteristic (ROC) plot, in which FAR is a function of FRR (see Fig.**??**). A common variant of this, is the Detection Error Tradeoff (DET) plot which is obtained using normal deviate scales on both axes. (see Fig.1.4).

Ranking capabilities of an identification system are evaluated using the Cumulative Match Curve (CMC) (see Fig.1.3). While the ROC curve plots the FAR of a 1:1 matcher, CMC represents a measure of 1:N identification system performance [5].

### 1.1.2 Limitations of unibiometrics

Most of the biometric systems, deployed in real world applications requiring a high security level, for authentication rely on the evidence of a single biometric source (e.g. fingerprint, face, voice etc.) [38]. Although if the biometric technique is becoming popular, there are still a variety of vulnerabilities that need to be addressed. Some of the challenges are described below:

*Susceptibility to noise.* Noisy input biometric data may be not accurately matched with the

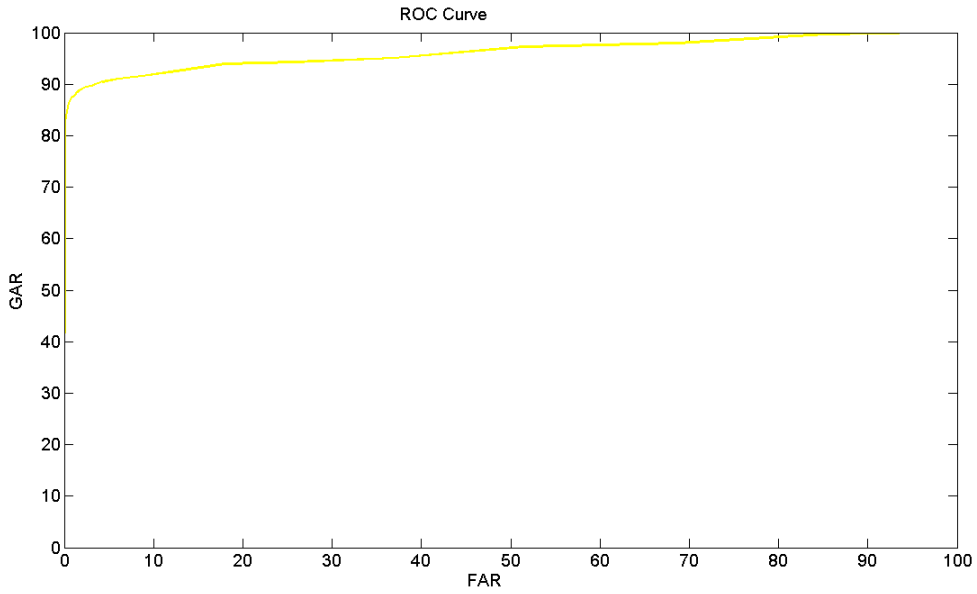


Figure 1.2: ROC curve for a fingerprint modality taken from Nist database.

templates, and subsequently, this may lead to a false rejection.

*Non-universality.* A particular biometric trait may not be possessed by a subset of the users; this may cause an increase of the enrollment failure rate.

*Distinctiveness.* A single biometric trait is expected to vary significantly across different subjects; however, there may be a large similarity among the values of features used to represent that trait [61].

*Intra-class variations.* The matching process may be affected by a significant variation between the biometric data acquired at authentication time and that one used to generate the template.

*Spoofing.* A biometric system may be circumvented by presenting a fake biometric trait to the sensor [23].

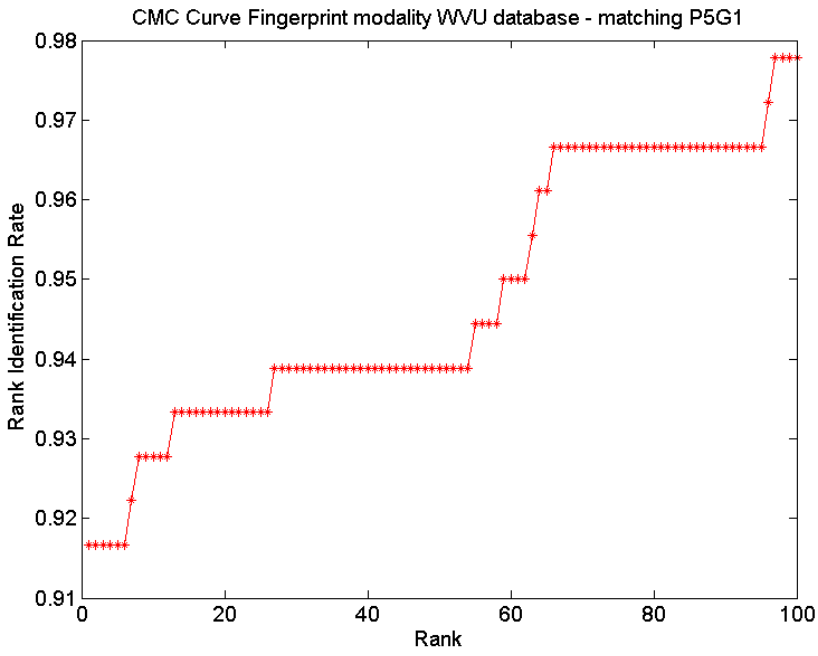


Figure 1.3: CMC curve for a fingerprint modality taken from WVU database. The considered probe (the fourth one) is not very similar to the gallery sample. This impacts the unimodal identification performance.

Examples of application of biometric technologies are showed in Fig.1.6.

## 1.2 Multibiometrics

The latest researches indicate that using a combination of biometric modalities, the human identification is more *reliable* [24]. Several works in the literature on biometrics demonstrate the efficiency of the multimodal fusion to enhance performance and reliability of the automatic recognition [61]. In particular, the work [64] shows the merit of both multimodal and intramodal fusion, and [31] demonstrates the effectiveness of using quality measures in the fusion. Integrating biometric information from multiple sources, multimodal biometric

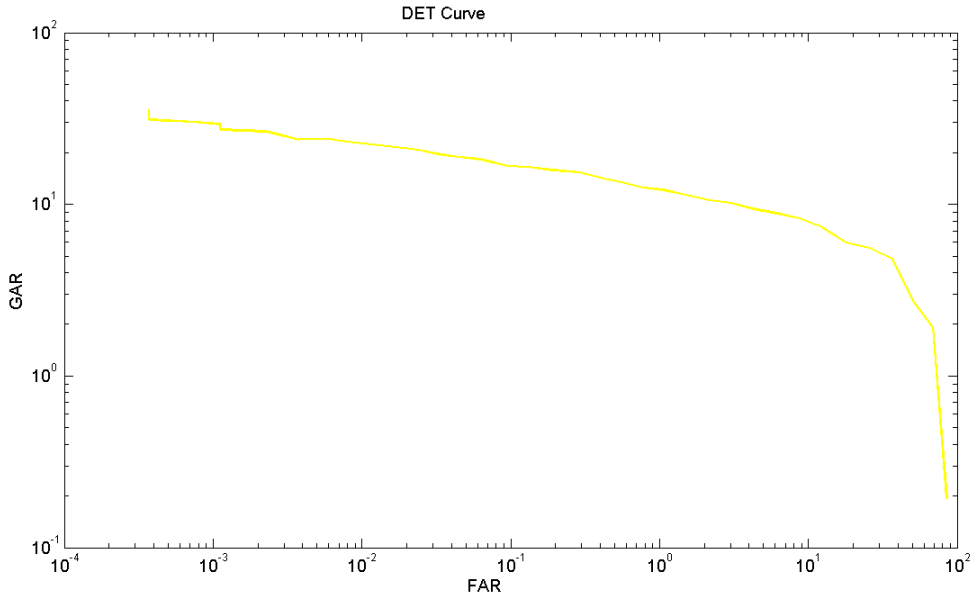


Figure 1.4: DET curve for a fingerprint modality taken from Nist database.

systems are able to improve the authentication performance, increase the population coverage, offer user choice, make biometric authentication systems more reliable and robust to spoofing [29].

However, the benefits of multibiometrics depend on the accuracy, complementarity, reliability and quality measurement of their component biometric experts. Moreover, when designing a multibiometric system, several factors should be considered. These concern the choice and the number of biometric traits, the level of integration and the mechanism adopted to consolidate the information provided by multiple traits.

- Fusion at match score level is usually preferred due to the easy to access and combine the scores presented by different modalities.

- The parallel fusion strategy has been extensively explored, however serial and hybrid

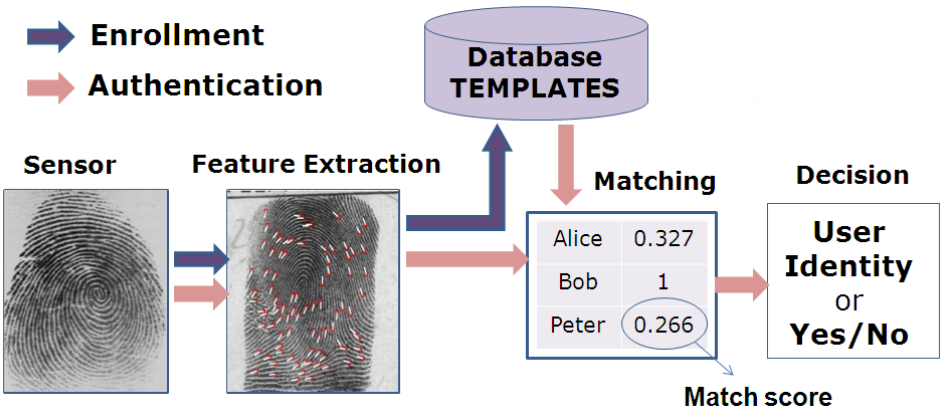


Figure 1.5: The sensor acquires the biometric data of a user from which a representative feature set is extracted. This feature set is matched against the feature set stored in the database of the system. The decision taken by the system is based on the match scores generated during the matching process [21]. In an identification system, these scores are transformed to ranks in order to determine potential matching identities.

architectures present important advantages. In particular, the serial fusion considers the biometric matchers one at a time, and makes a reliable decision by employing few experts and activating the remaining experts only for difficult cases.

In general, it is desirable that a fusion scheme involves statistically independent modality matchers. In a multimodal fusion, the set of expert outputs is expected to be statistically independent, while in intramodal fusion, where the component matchers rely on the same biometric trait, a high dependency is expected among the expert outputs [55].

### 1.2.1 Challenges and Difficult in Multibiometrics

#### Error Rates

Although individual modalities have proven to be reliable in ideal environments, they can be very sensitive to real environmental conditions. In real scenarios, it is difficult to acquire high quality samples, then biometric authentication errors are inevitable [17]. The performance

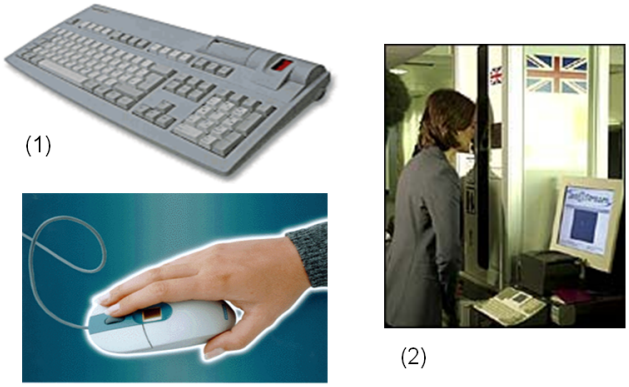


Figure 1.6: 1) Fingerprint sensors installed on a keyboard (the Cherry Biometric Keyboard and on a mouse (the ID Mouse manufactured by Siemens). 2) A border passage system using iris recognition (at London's Heathrow airport).

of several current unimodal systems is reported in Fig.1.7.

The impact of adverse environmental conditions on the characteristics of the collected biometric data can be quantified by *quality measures*. It is evident that, a degradation in the quality level of the biometric signal input may affect the reliability of the matching process. The performance of the single modality matcher may change as the data quality changes and different modality matchers are sensitive to different aspects of the signal quality. Then, the opinion of a matcher in the decision of the ensemble have to be appropriately weighted, by assigning a higher weight to the matcher with higher quality data. The same observation has to be considered for the reliability, accuracy and competence of each component matcher.

From the viewpoint of a human observer, a sample of good quality may be a fingerprint image with a good contrast and clear ridges. However, if only few minutiae points can be detected, a matcher based on minutiae will be not effective [19], (see 1.9). This can happen, for example, in presence of cuts on a finger, (see Fig.1.8), which alters the ridge structure

| | Test | Test Parameter | FNMR | FMR |
|---|---|---|---|---|
| **Fingerprint** | FVC 2002 | Users mostly in the age group 20-39 | 0.2% | 0.2% |
| **Face** | FRVT 2002 | Enrollment and test images were collected in indoor environment and could be on different days | 10% | 1% |
| **Voice** | NIST 2000 | Text dependent | 10-20% | 2-5% |

Figure 1.7: Unimodal error rates associated with fingerprint, face, and voice biometric systems [25]. FNMR indicates False Non-matched Rate (FRR) while FMR indicates False Matched Rate (FAR).

of the fingerprints resulting in less efficiency of the matching process.

Figure 1.8: A fingerprint image when the presence of minor cuts alters the ridge structure.

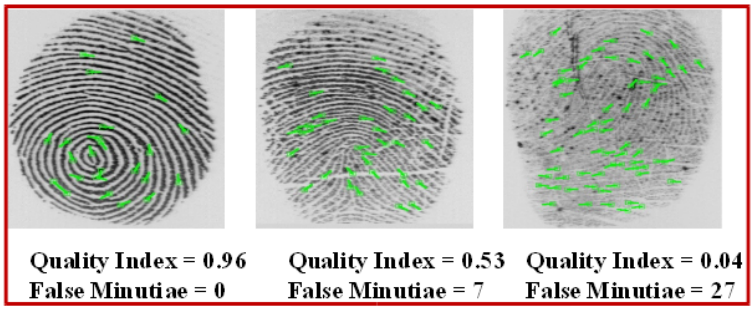


Figure 1.9: Quality measures for fingerprint images input for a minutiae-based matcher.

The quality of collected biometric samples can significantly vary, due to the intrinsic variability of behavioral factors and to the not always well-controlled acquisition conditions. For example, a visible-light face image may change by varying the illumination conditions, facial expressions, makeup, etc. (see Fig.1.10), while a fingerprint image may be affected

by factors like humidity and ambient temperature.

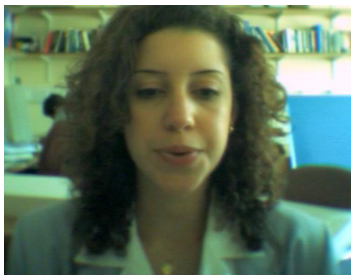

Figure 1.10: An example of a face image acquired in adverse conditions, taken from BANCA database.

### Spoof Attacks

A hacker may gain an unauthorized access by exploring several points of a biometric system. The main vulnerabilities are shown in Fig. 1.11). For example, an impostor may attack the server where the templates are stored by introducing his own template. In this thesis, we focus on attacks at sensor level, where artificially created fingers are presented during authentication.

Previous studies [70] have shown that it is not difficult to create fake fingers using play-doh, gelatin and silicon based on molds of latent fingerprints (see Fig.1.12 and 1.13).

An example of live and spoof fingerprint is shown in Fig.1.14.

The treat of spoofing, where an impostor fakes a biometric trait, has encouraged the use of multimodal biometric systems. However, multimodal systems are not more secure than their unimodal systems alone since the use of multiple modalities offers more vulnerable points to a hacker. The security risk in multimodal systems due to spoof attacks has been evaluated under the assumption that an impostor must fake all the fused modalities to be accepted. However, a malicious user may attack only one or a subset of modalities in the

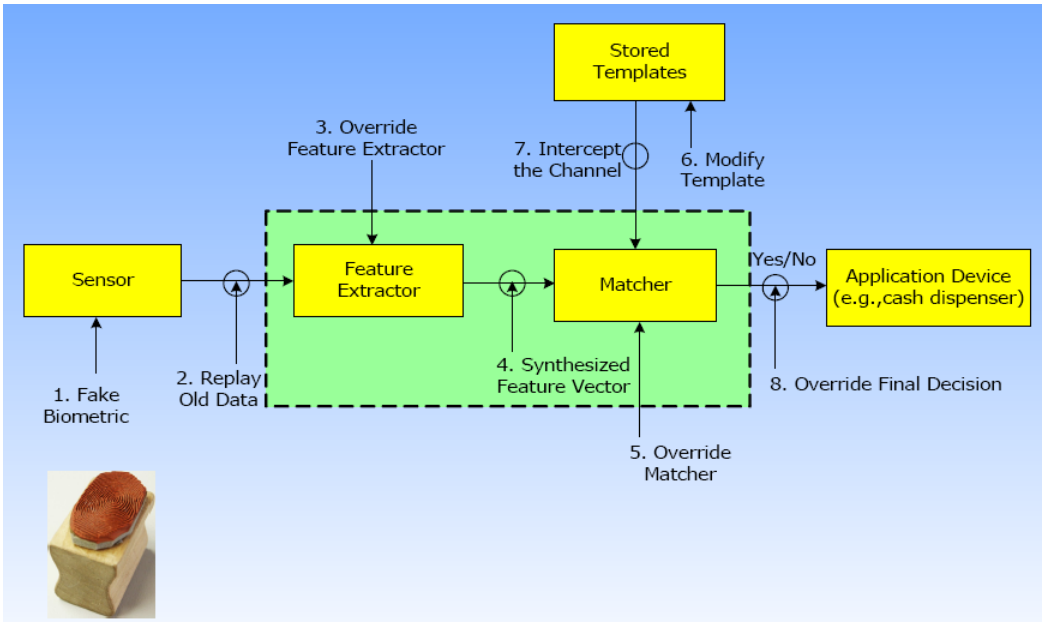


Figure 1.11: Vulnerable points of attacks in a biometric system [12].

system.

**Other Issues**

Multibiometric systems are still rarely used in real applications since combining multiple traits induces some drawbacks as the increase in complexity of the overall system. Moreover, a multibiometric system is expected to have a higher cost, a longer authentication time and a lower user convenience with respect to its unimodal component. Thus, in the evaluation these aspects have to be take into account.

## 1.3 Thesis Contributions

### 1.3.1 Improvement of Performance

*Reducing verification errors* of a score level fusion scheme based on the Likelihood Ratio (LR)-test statistic. Due to the diversity of scenarios, the use of a single rule may be

Figure 1.12: Examples of some inexpensive materials employed for creating artificial fingerprint (Play-Doh and Silicon).

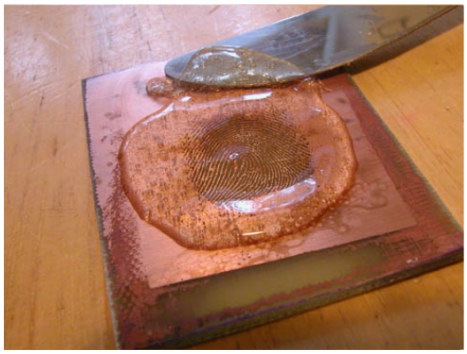

Figure 1.13: Use of gelatin to make a fake fingerprint.

not always efficient, thus we adopted two schemes: *i)* a sequential fusion technique in conjunction with a majority voting strategy to improve the performance of a framework based on LR test; *ii)* a LR-based voting strategy alone, when the number of training samples is limited.

The sequential fusion strategy considers unimodal systems sequentially, so the decision can be made by employing as fewer systems as possible. In this mechanism, the induced cost of the multimodal system increases with its security level, as required by the application. The component systems are sorted in a decreasing order of confidence.

*Addressing the problem of identification errors* by setting up a predictor of errors. The proposed predictor exploits ranks and scores generated by the identification operation and can

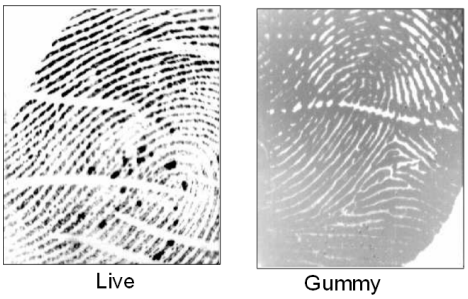


Figure 1.14: An example of live and fake (gummy) fingerprint image.

be effectively applied in a multimodal scenario. The motivation for using reliability information in fusion emerged in relation to multimodal biometric systems, where the modalities performing poorly as a results of degraded quality of biometric information or a low competence of a single matcher should influence the final decision. This suggested a reliability dependent weighting of modalities as solution to the fusion problem. The idea is to combine multiple independent modalities which are not degraded, so the system will offer a more robust authentication in adverse conditions. In a fusion mechanism, it is necessary to take into account the fact that individual decisions depend on the acquisition condition of the data presented to the expert as much as they depend on the discriminating skills of the classifier.

### 1.3.2 Improvement of Security

*Improving the security* of the existing multibiometric systems against spoof attacks. We demonstrated that there is a significant security risk where only a subset of the modalities used in the system are spoofed. We experimentally showed that, in such a real scenario, the performance of the score sum scheme and of the statistic Likelihood Ratio test decreases in presence of spoofing. For the fingerprint modality, we proposed a novel liveness detection

algorithm which combines perspiration- and morphology-based static features. Further, we demonstrated that, by incorporating our algorithm in the fusion scheme, the multimodal system results robust in presence of spoof attacks involving a only subset but not all fused modalities.

The field of combining classifiers is like a teenager: full of energy, enthusiasm, spontaneity, and confusion.

*Ludmila Kuncheva*

# Chapter 2

# Information Fusion in Biometrics

## 2.1 Multiple Biometric Sources of Information

Classifier combination may involve a set of classifiers where all the components use the same representation of the input pattern or each one of them can use its own representation [28]. In the context of biometrics, information fusion concerns the consolidation of evidence provided by multiple biometric sources in order to output a decision [60]. These biometric sources of information may be derived from the same biometric or different biometric traits (see Fig.2.1). In presence of multiple sensors (e.g., capacitive and optical fingerprint sensors), multiple instances (e.g., multiple face images captured under different poses), multiple representations (e.g., texture- and minutiae-based), multiple units (e.g., right eye and left eye), the information is derived from a single biometric modality, while in presence of multiple traits (e.g., iris, face and fingerprint) the information is derived from different biometric modalities [61]. A multimodal system where different traits are fused, is expected to be more robust to noisy data, non-universality, provide higher accuracy and protection against spoof attacks. Exploiting multiple traits can significantly enhance the recognition accuracy

[63]. Further, physically uncorrelated modalities (e.g., fingerprint and iris) are expected to result in a better performance improvement than that achieved by fusing correlated traits (e.g., voice and lip movement).

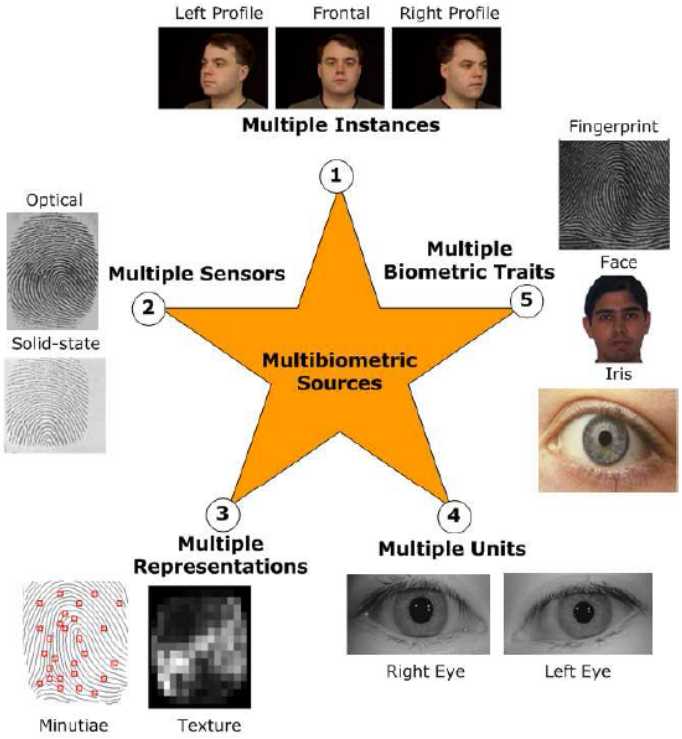


Figure 2.1: Different sources of biometric information which can be fused.

Besides enhancing matching accuracy, there are several advantages of multibiometric systems over traditional unibiometric systems [59].

- Multibiometric systems address the issue of nonuniversality (i.e., limited population coverage) encountered by unibiometric systems. They guarantee a certain degree of flexibility during the user's enrollment since he can use several different traits (e.g., face, voice, fingerprint, iris, hand). Based on the nature of the application and the convenience of the user, only a subset of these traits (e.g., face and voice) is requested

during authentication.

- It makes difficult for an impostor to spoof multiple biometric traits of a legitimately enrolled individual. Furthermore, by asking the user to present a random subset of traits at the point of acquisition, a multibiometric system ensures that the system is interacting with a live user.

- When recognition has to take place in adverse conditions where certain biometric traits cannot be reliably extracted. For example, in the presence of ambient acoustic noise, when an voice characteristics of an individual cannot be accurately measured, then the authentication may be based on the fingerprint.

- Multibiometrics help also in applications where a continuous tracking of an individual is needed, a single trait is not sufficient.

- A multibiometric system may also be viewed as a fault tolerant system which continues to operate even when certain biometric sources become unreliable due to sensor or software malfunctioning. The notion of fault tolerance is especially useful in large-scale authentication systems involving a large number of subjects (such as a border control application), where the distributions of the subjects may overlap.

## 2.2 Different levels to make Fusion

The key to create a secure multimodal biometric system is in how the information from different modalities is fused, (see Fig. 2.2) [61]. The consolidation of biometric information can be performed at various levels: sensor level, feature extraction level, match score level,

rank level (identification operation) and decision level. An additional post-matching fusion level, which we will not be analyzed in this thesis, regards the *dynamic classifier selection* scheme, which chooses the results of the modality matcher with highest probability to output a correct decision about the input pattern [48]. Consolidating data at an early stage of the recognition process involves a higher informative contain concerning the biometric input. Thus, it is potentially able to provide better recognition results, but in practice concatenating data at a level before matching may result difficult or not possible. When the information fusion is performed at sensor level, raw data from different sensors are combined. For example, the fingerprint images taken from different sensors are combined to form a single fingerprint image (fingerprint mosaicking). However, images captured from sensors with a different resolution are not compatible. Fusion at feature level is difficult since the features vectors to be fused may be not compatible (e.g., fingerprint minutiae and eigenface coefficients) and not accessible (feature sets can be proprietary). When the output of each biometric matcher is a subset of possible matches sorted in decreasing order of confidence, the fusion can be done at the rank level. Each possible match is assigned the highest (minimum) rank as computed by different matchers. Ties are broken randomly to arrive at a strict ranking order and the final decision is made based on the combined ranks. Fusion at decision level involves only a limited amount of information since each biometric matcher individually decides about which is the best match based on the biometric input presented to it. Combining match scores provided from different matchers is the most effective fusion strategy because they offer the best trade-off between the informative contain and the ease to implement the fusion.

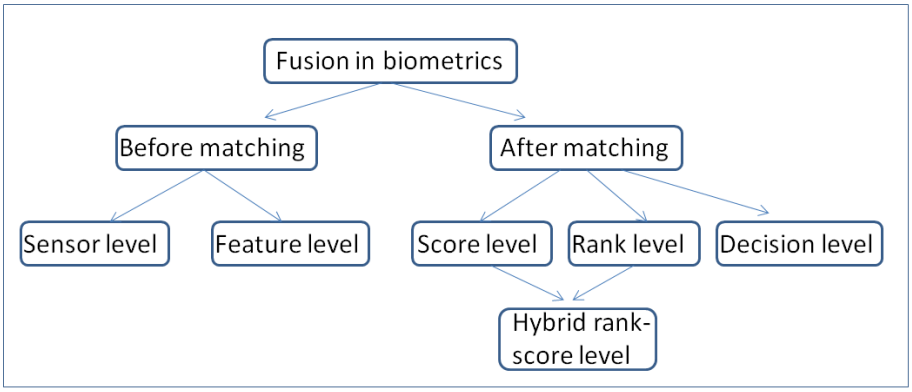


Figure 2.2: Levels of fusion in biometrics.

### 2.2.1 Match Score Information

Match scores are commonly used to consolidate the decisions rendered by multiple biometric classifiers since they are easy to access and to combine. However, the scores output by different biometric matchers may not be homogeneous, can conform to different scales. For example, face matcher may output a distance measure while fingerprint matcher may output a similarity measure. Further, they may follow different statistical distributions [61]. Thus, before integration, match scores must be transformed into a common domain via score normalization. Choosing an effective normalization scheme is a critical part in the design to combine different matchers. It refers to changing the location and scale parameters of the match score distributions outputs of the individual matchers [21] [6]. For a good normalization scheme, the location and scale parameters of match score distributions must be *robust* and *efficient* [48]. Many methods for score normalization have been proposed [22], and fusion rules performance changes by varying the technique. The technique adopted in our fusion framework is the *min-max*, which retains the original distribution of scores except

a scaling factor and transform the scores to a common range of zero to one, based on the minimum and the maximum score values (see Fig. 2.3). Given a set of matching scores $s_k$,

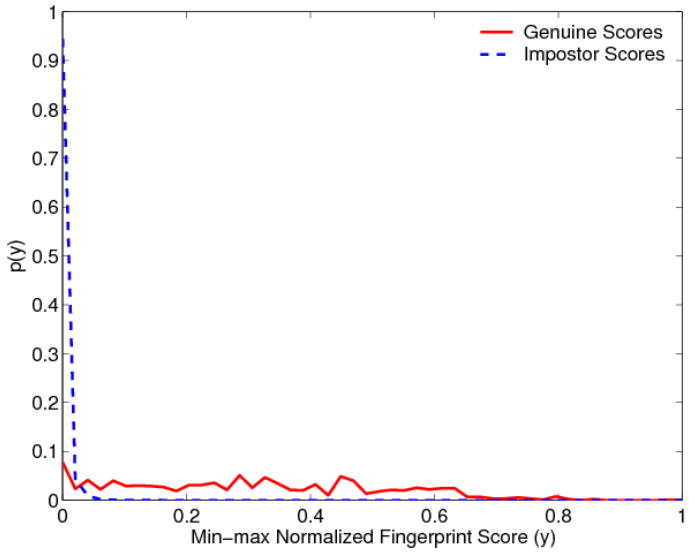


Figure 2.3: Distributions of genuine and impostor match scores after min-max normalization for fingerprint modality [48].

$k = 1 \ldots K$, the normalized scores are given by (2.1).

$$s_k = \frac{s_k - min}{max - min} \tag{2.1}$$

The most commonly used score normalization technique is the *z-score*, which exploits the average score and the score variations of each matcher. The normalized scores are given by (2.2).

$$s_k = \frac{s_k - \mu}{\sigma} \tag{2.2}$$

A technique which presents high efficiency and robustness is the *tanh*. The normalized scores are given by (2.3).

$$s_k = \frac{1}{2}\{\tanh\,(0.01(\frac{s_k - \mu_{GH}}{\sigma_{GH}})) + 1\} \tag{2.3}$$

where $\mu_{GH}$ and $\sigma_{GH}$ are the mean and standard deviation of the genuine score distribution as given by Hampel.

*Median* normalization is a robust technique which is not sensitive to the points composing the tails of the score distributions. The normalized scores are given by (2.8).

$$s_k = \frac{s_k - median(s_k)}{MAD} \tag{2.4}$$

where $MAD = median(|s_k - median(s_k)|)$ This scheme does not retain the input distributions (see Fig. 2.4). The main drawback appears in presence of score distributions which are not Gaussian, since median cannot be accurately estimated. When the scores of different

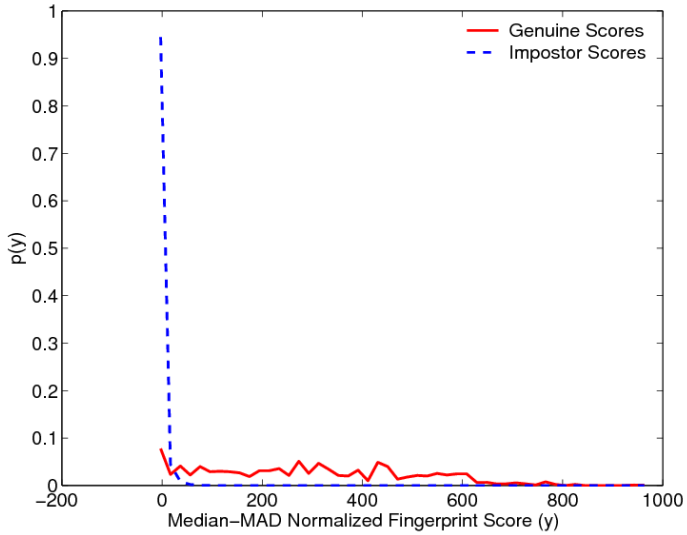


Figure 2.4: Distributions of genuine and impostor match scores after median-MAD normalization for fingerprint modality [48].

matchers are on a logarithmic scale, applying *decimal scale* can be useful. The scheme is the following (2.5), where $n = \log 10 maxsk$

$$s_k = \frac{s_k}{10^n} \tag{2.5}$$

### 2.2.2 Rank Information

At the rank level, each biometric matcher orders the candidate identities in the gallery according to their similarities to the given probe and transforms this ordering into a set of $N$ integer values or *ranks*. A fusion scheme at this level consolidates the rankings provided by multiple biometric matchers in order to obtain a consensus rank for each identity in the gallery [61]. If we consider an input image having low quality, the genuine score as well as the impostor scores are likely to be low [44] [66]. The use of such a score (for a genuine user) during the fusion process may confuse a fusion algorithm. The rank, on the other hand, is a relatively stable statistic and does not require normalization; combining this rank with other ranks (for the genuine class) in a judicious manner can result in a correct classification.

### 2.2.3 Hybrid Rank-Score Information

The use of both ranks and match scores [6] is expected to be more reliable and has been demonstrated to increase the recognition accuracy of a multibiometric system [49]. For a given probe image, a $N \times C$ score matrix $S = [s_{ik}]$ can be generated where $s_{ik}$ represents the similarity score computed by the $k^{th}$ modality matcher $C_k$ after comparing the probe against the $i^{th}$ entry in the gallery database, $i = 1 \ldots N$ and $k = 1 \ldots C$. For each modality, the corresponding scores can be sorted in decreasing order. So a $N \times C$ rank matrix $R = [r_{ik}]$ can be generated where $r_{ik}$ is the rank assigned to the $i^{th}$ identity in the database by the matcher $C_k$. Thus, the output of each matcher, $C_k$, can be viewed as a two-tupled entry $(s_{ik}, r_{ik})$, $i = 1 \ldots N$, (see Fig.2.5).

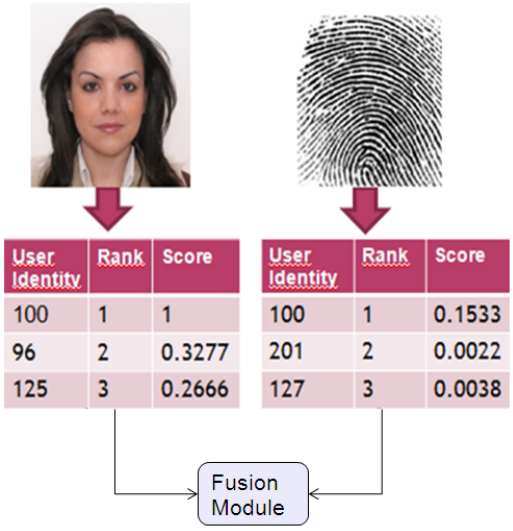


Figure 2.5: Fusing face and fingerprint biometric systems at hybrid rank-score level.

So, the information presented by multiple traits may be consolidated at various levels of recognition process. At feature extraction level, a new feature set is produced by fusing the features sets of multiple modalities, and this new feature set is used in the matching module. At match score level, the scores produced by multiple matchers are integrated, while at decision level the decisions made by the individual systems are combined. The integration at feature extraction level is expected to perform better, but the feature space of different biometric traits may not be compatible and most commercial systems do not provide access to information at this level. So, researchers found at score level a good compromise between the ease in realizing the fusion and the information content.

## 2.3 Post-matching Fusion Approaches

### 2.3.1 Fusion Approaches at match score-level

Fusion at match score level concerns combining the match scores generated by multiple classifiers in order to make a decision about the identity of the subject. In literature, the fusion at score level is performed by employing different approaches [46] based on different models [63].

1. *Classifier-based schemes.* The model is a classifier which is trained using a feature vector composed by the scores output by the matchers to be fused [37]. This is accurate to correctly discriminate between genuine and impostor classes, regardless of the non-homogeneity of the score, but it typically requires a large training set. In particular, the case when the scores output by different matchers are conflicting, in absence of sufficient training samples may be not well represented in the training data, resulting in incorrect decision. Wang *et al.* used Fisher's discriminant analysis and a neural network classifier with radial basis function employing a two-dimensional feature vector composed by iris and face scores [77]. Ross and Jain used linear discriminant classifiers and Decision Tree to combine fingerprint, face and hand-geometry scores [61]. A Support Vector Machine was used to combine face and speech scores by Sanderson [65].

2. *Transformation-based schemes.* In situations where it is not possible to acquire a large number of labeled multibiometric data in an operational environment, it may be convenient to directly combine the match scores without interpreting them in

a probabilistic framework [62]. The match scores provided by different matchers are firstly transformed into a common domain (*score normalization*), then they are combined using a simple fusion rule. This approach is quite complex since it implicates a wide experimental analysis to choose the best normalization scheme and combination weights for the specific dataset of interest. The model is based on a normalization function. The operators which are commonly used in the literature are *min*, *max*, *median*,*weighted sum* and *weighted product*, defined by (2.6), (2.7), (2.8), (2.9) and (2.10).

$$s_{min} = \min_k s_k \tag{2.6}$$

$$s_{max} = \max_k s_k \tag{2.7}$$

$$s_{median} = median_k s_k \tag{2.8}$$

$$s_{sum} = \sum_{k=1}^{K} w_k s_k \tag{2.9}$$

$$s_{prod} = \prod_{k=1}^{K} s_k^{wk} \tag{2.10}$$

where $w_k$ are parameters that need to be estimated. The simple *sum* operator (or *mean*) is a special case of *weighted sum* with $w = \frac{1}{N}$, while the *product* operator is a special case of *weighted product* with $w = 1$. The operators which do not contain parameters to be tuned, are known as *fixed* combiners [53]. Based on experimental results, researchers agree that *fixed* rules usually perform well for ensemble of classifiers having similar performance, while *trained* rules handle better matchers having different accuracy. Thus, when fusing different modalities, individual matchers often

exhibit different performance, then for this problem *trained* rules should perform better than *fixed* rules [58]. It has been shown that, the simple sum rule gives very good accuracy in combining multiple biometric systems [58].

3. *Density-based schemes.* The model is built by estimating density functions for the genuine and impostor score distributions [74]. The match scores are considered as random variables, whose class conditional densities are not *a priori* known [13]. So, this approach requires an explicit estimation of density functions from the training data [63]. A recent method belonging to this category is the score fusion framework based on the Likelihood Ratio test, proposed by Nandakumar et al. in [46]. It models the scores of a biometric matcher by a mixture of Gaussians and perform a statistical test to discriminate between genuine and impostor classes. This framework produces high recognition rates at a chosen operating point (in terms of False Acceptance Rate), without the need of parameter tuning by the system designer once the method for score density estimation has been defined. Optimal performance, in fact, can be achieved when it is possible to perform accurate estimations of the genuine and impostor score densities. The Gaussian Mixture Model (GMM) lets to obtain reliable estimations of the distributions, even if the amount of data needed for it increases as the number of considered biometrics increases.

   Let $\mathbf{s} = [s_1, s_2, ..., s_K]$ denote the scores emitted by multiple matchers, with $s_k$ representing the match score of the $k_{th}$ matcher, $k = 1, ..., K$. Adopting the Bayesian decision rule, the probability of error can be minimized.

Assign $\mathbf{s}$ to genuine if $P(genuine|\mathbf{s})$ is greater or equal to $P(impostor|\mathbf{s})$. The *a posterior probability* $P(genuine|\mathbf{s})$ can be derived from the class-conditional density functions $P(\mathbf{s}|genuine)$ using the Bayes formula [62]:

$P(genuine|\mathbf{s}) = P(\mathbf{s}|genuine)P(genuine)/P(\mathbf{s})$

Moreover, as noted by the authors in [46], the performance of their method can be improved by using a suitable *quality measure* together with each score. Most of the available biometric systems, however, do not provides such measures. The main drawback of a likelihood ratio fusion rule is that performance can be affected by inaccurate estimations of the density functions.

Due to the diversity of scenarios encountered in the datasets, training and using a single fusion rule on the entire dataset may not be appropriate. Recently [74], the idea of dynamically selecting biometric fusion algorithms has been adopted.

### 2.3.2 Fusion Approaches at Rank-Level

For systems operating in identification mode, rank level fusion is a viable option. It provides a richer information into the decision-making process compared to the decision level, without requiring a normalization phase before combining [1]. Let $K$ be the number of matchers to be fused and $N$ the number of enrolled users. Let $r_{ij}$ be the rank assigned to the $j^{th}$ user enrolled in the database by the $i^{th}$ matcher, $i = 1 \dots K$, and $i = 1 \dots N$, then $R_{ij}$.

*Highest rank scheme.* For each subject, the combined rank is given by the lowest rank (2.11). This rank fusion technique presents the advantage of utilizing the strength of each

matcher.

$$R_i = \min_{k=1}^{K} r_{ik}, \qquad i = 1, 2, ...N \tag{2.11}$$

*Borda Count scheme.* For each subject, the combined rank is given by the sum of the ranks assigned by the individual matchers (2.12). Such a rule presents the advantage of taking into account the variability of the single matcher outputs. Its drawbacks lie in the assumptions that, the matchers are statistically independent and they perform equally well. This makes the Borda Count method highly vulnerable to the effect of weak classifiers.

$$R_i = \sum_{k=1}^{K} r_{ik}, \qquad i = 1, 2, ...N \tag{2.12}$$

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

$$R_i = \sum_{k=1}^{K} w_k r_{ik}, \qquad i = 1, 2, ...N \tag{2.13}$$

The weight $w_k$, $i = 1 \dots K$, (see equation (2.13)), is determined through a training phase by logistic regression. This method is useful when the different biometric matchers have significant differences in their accuracies [63].

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

### 2.3.3 Fusion Approaches at Hybrid Rank-Score Level

An interesting technique for the integration of multiple classifiers at an hybrid rank-score level is introduced using a HybridBF network. In [6] Falavigna and Brunelli showed that, a system based on the integration of a visual and an acoustic subsystems achieves superior performance compared to that of its components. The proposed approach reconstructs a mapping from the set of scores and the corresponding ranks into a set 0,1. The matching of the probe against each gallery identity is mapped to 1, if it corresponds to the correct label, to 0 otherwise. The reconstruction of the mapping is formulated as a learning task problem, where the training set is composed by non-matched and matched inputs and the system will appropriately classify unseen data. This method has some drawbacks. Firstly, a network-based framework requires a large amount of training examples to tune the free parameters involved. Secondly, it requires the availability of all classifiers. Finally, when a new user is added, the network has to be trained again.

Recently, Nandakumar et al. [49] proposed a scheme that utilizes both ranks and scores to perform fusion in identification systems. They defined a hybrid rank-score fusion rule based on a combination of score and rank statistics, defined as indicated in the equation (2.14).

Assign query to identity $I_n$ if

$$R_n \geq R_i, \quad i = 1, 2, ...N \tag{2.14}$$

where the combined score and rank statistic is defined by equation (2.15).

$$R_i(\mathbf{S},\mathbf{R}) = P(I_i|\mathbf{S})r_i, \quad i = 1, 2, ...N \tag{2.15}$$

in which $P(I_i|\mathbf{S})$ is the posterior probability that $I_i$ is the true identity given the score matrix $\mathbf{S}$, and $r_i$ is given by the equation (2.16) under the assumption that the matchers are independent.

$$r_i = \prod_{k=1}^{K} P_k r_{ik} \qquad i = 1, 2, ...N \tag{2.16}$$

This approach, however, requires an explicit estimation of the genuine and impostor distributions, and a large dataset is required to accurately estimate the score distributions.

**Summary**

Multibiometric systems consolidate the evidence provided by multiple sources of biometric information, and subsequently, they are able to improve recognition performance compared to its unimodal components. In order to maximize the benefits of multimodal biometric systems, an effective fusion scheme is needed to consolidate the information provided by different modalities. Among the possible integration levels, fusion at match score level is the most commonly used, since scores are easy to access and to combine. However, they are not homogeneous, then an efficient normalization phase is required before fusion. In this chapter, we reported various normalization and integration schemes which have been proposed in the biometric literature for multimodal biometric systems designing.

The state of mind which enables a man to do work of this kind ...; the daily effort comes from no deliberate intention or program, but straight from the heart.

*Einstein*

# Chapter 3

# Multibiometric Verification Scenario

The goal of a multi-modal systems is to alleviate limitations of mono-modal systems, in particular to reduce decision errors. Among the existing approaches for combining several biometric traits, the fusion of match scores has been widely adopted. Recently, a scheme using the Likelihood Ratio (LR) Test has been proposed. In such approach, the distributions of genuine and impostor scores are modeled as a finite mixture of gaussians that can be accurately estimated only in presence of a huge training set.

In this chapter, we proposed a solution to reduce some limitations of the existing *density-based* approaches; in particular, we presented two novel score fusion strategies based on the Likelihood Ratio test. We propose both a sequential test and a voting strategy. By using them, on one hand we tried to implicitly use the quality information embedded into the scores. On the other hand, we obtained a system that demonstrated to be more robust than the original one with respect to the lack of data for training [40]. Our case study concerns the combination of face and fingerprint recognition systems at score level, as shown in

Fig.3.1.

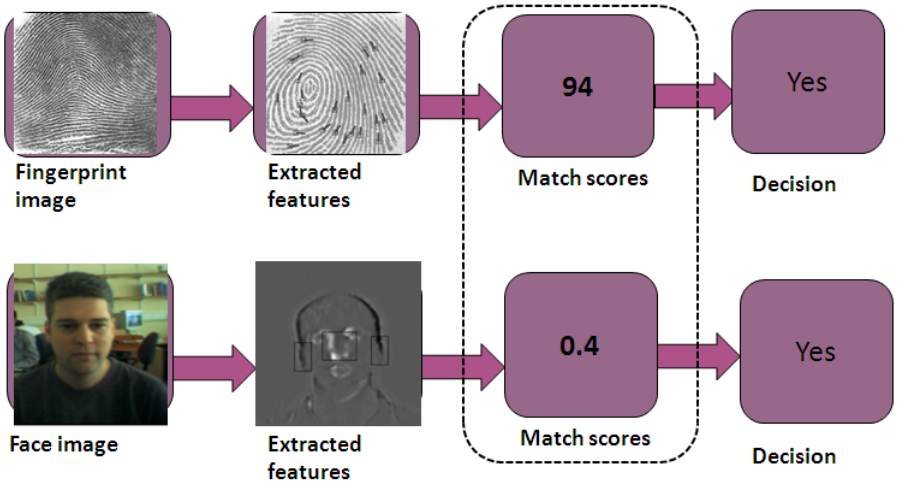


Figure 3.1: Combination of face and fingerprint modalities.

## 3.1 The Likelihood Ratio Test

Nandakumar and Chen [46] formulate the problem of Identity Verification in terms of hypothesis testing: let $\Psi$ denote a statistical test for deciding if the hypothesis $H$:{*the score vector* $\boldsymbol{s}$ *belongs to the Genuine class*} has been correctly formulated. The choice is based on the value of the observed match score and it lies between only two decisions: accepting $H$ or rejecting it. As it is known [18], different tests should be compared with respect to the concepts of *size* and *power*, that are respectively the probability of accepting $H$ when it is false (also called *False Accept Rate* - FAR) and the probability of accepting $H$ when it is true (also called *Genuine Accept Rate* - GAR) [18]. In the context of *prudential decision making* [36], the NP lemma [18] recognizes that, in choosing between a hypothesis $H$ and an alternative, the test based on the Likelihood Ratio is the best because it maximizes the

*power* for a fixed *size* [18]. Let

$$LR(\mathbf{s}) = \frac{f_{gen}(\mathbf{s})}{f_{imp}(\mathbf{s})} \tag{3.1}$$

be the *Likelihood Ratio* ($LR$), that is the probability of the observed outcome under $H$ divided by the probability of assuming its alternative. As stated by the Neyman and Pearson theorem [18], the framework proposed by Nandakumar and Jain ensures that the most powerful test is the one, say $\Psi(\mathbf{s})$, that satisfies the equations (1) for some $\eta$

$$\Psi(\mathbf{s}) = \begin{cases} 1, & \text{when } LR(\mathbf{s}) \geq \eta \\ \\ 0, & \text{when } LR(\mathbf{s}) < \eta \end{cases} \tag{3.2}$$

where $\mathbf{s} = [s_1, s_2, ...s_K]$ is an observed set of $K$ match scores that is assigned to the genuine class if $LR(\mathbf{s})$ is greater than a fixed threshold $\eta$, with $\eta \geq 0$.

### 3.1.1 The Estimation of Match Score Densities

As it is known in biometric literature [63], it is hard to choose a specific parametric form for approximating the density of genuine and impostor match scores, because the match distributions have a large tail, discrete components and not only one mode.

Given a training set, density estimation can be done by employing parametric or non-parametric techniques [4]. The non-parametric techniques do not assume any form of the density function and are completely data-driven; on the contrary, parametric techniques assume that the form of the density function is known (e.g., Gaussian) and estimate its parameters from the training data. The power of this scheme resides in its generality [14]: exactly the same procedure can be used also if the known functions are a mixture of Gaussians. In [46] the authors have proved the effectiveness of the GMM for modeling score

distributions and of the likelihood ratio fusion test in achieving high recognition rates when densities estimations are based on GMM [46].

Let $\mathbf{s} = [s_1, s_2, ...s_K]$ denote the score vector of $K$ different biometric matchers, where $s_j$ is the random variable representing the match score provided by the $j^{th}$ matcher, with $j = 1, 2, ..., K$. Let $f_{gen}(textitbfs)$ and $f_{imp}(\mathbf{s})$ denote the conditional joint density of the score vector $\mathbf{s}$ given respectively the genuine and impostor class. The estimates of $f_{gen}(\mathbf{s})$ and $f_{imp}(\mathbf{s})$ are obtained as a mixture of Gaussians:

$$\hat{f}_{gen}(s) = \sum_{j=1}^{M_{gen}} p_{gen,j} \Phi^K(s; \mu_{gen,j}, \Sigma_{gen,j}) \tag{3.3}$$

$$\hat{f}_{imp}(s) = \sum_{j=1}^{M_{imp}} p_{imp,j} \Phi^K(s; \mu_{imp,j}, \Sigma_{imp,j}) \tag{3.4}$$

where $\Phi^K(s; \mu; \Sigma) = (2\pi)^{-K/2} |\Sigma|^{-1/2} exp(-\frac{1}{2}(s-\mu)^T \Sigma^{-1}(s-\mu))$ denotes the Gaussian density with mean $\mu$ and covariance matrix $\Sigma$, and $M_{gen}$ ($M_{imp}$) represents the number of mixture components. Mixture parameters can be approximated by employing the fitting procedure of Figuereido and Jain [15], that uses EM algorithm and Minimum Message Length (MML) criterion. It also estimates the optimal number of Gaussians and is able to treat discrete values by modeling them as a mixture with a very small variance represented as a regularization factor added to the diagonal of the covariance matrix.

Fusion based on GMM estimations achieves high performance [46], but there is an important drawback. In practice, one has to determine reliable models for estimations of

genuine and impostor match score densities from the available score to be used for training. In absence of a large database, it is hard to obtain an accurate model, and this limitation is particularly true for multibiometric systems, as the number of considered biometrics increases.

## 3.2 The Proposed Approach

As said in the introduction, the quality of the acquired biometric data affects the efficiency of a matching process [47]. When the samples presented to a matcher are of poor quality, it cannot reliably distinguish between genuine and impostor users. For example, some true minutiae may not be detected in noisy fingerprint images, and missing minutiae may lead to errors. Moreover, as stated in the previous Section, when several biometrics are available, a not huge dataset could be not sufficient for having a proper density estimate by means of the GMM. So, we propose two approaches for improving the performance of the standard $LR$ test.

### 3.2.1 LR-based Majority Voting

An analysis of how the exclusion of some biometric modalities affects the GMM estimate: this approach (hereinafter denoted as *voting LR*) can be associated to the attempt of implicitly individuating degraded quality samples, when the quality measures are not available. In practice, given a $K$-dimensional score vector, we estimate the $K$ conditional class joint densities of $K$-1 scores, by using a GMM technique. Then, we fixed for each of the $K$ estimates a threshold $\eta$ on the training set that gives rise to a FAR equal to 0%. When we have to judge a new sample, $K$, $LR$ tests are made on the $K$ densities and if at least one of

the $LR$ tests recognizes the sample as genuine it is declared as genuine by the system. The ratio of this procedure lies in the fact that we want to detect if a particular score, say $s_i$, coming from a genuine sample, could be affected by a low quality. In this case, it can be expected that all the score vectors including $s_i$ it will result in a low $LR$ value, giving rise to a false rejection. Only the $K$-1 dimensional score vector that do not include $s_i$ could have a $LR$ value able to overcome the threshold. So, if at least one test is passed, the sample with a single modality affected by low quality can be correctly recognized. The choice of fixing $\eta$ on the training set so as to obtain a FAR equal to 0%, is motivated by the need of having a system characterized by a FAR as low as possible. Since this approach uses only $K$-1 dimensional score vectors, it should be also more robust to the lack of training data. See Fig.3.2.

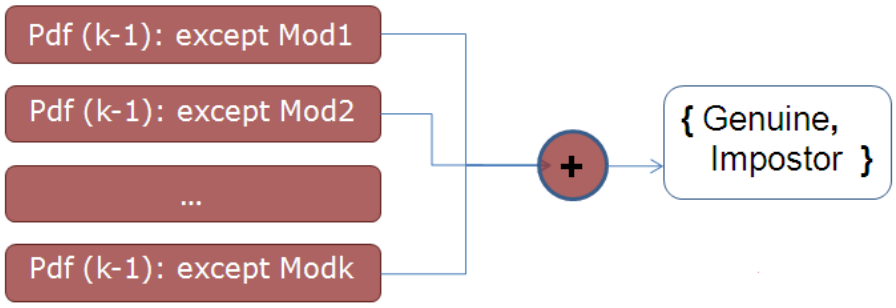


Figure 3.2: The input (biometric, claimed Id) is classified as genuine if at least one of the k LR test outputs genuine.

### 3.2.2 LR-based Sequential Approach

A sequential likelihood ratio test (hereinafter denoted as *Sequential LR*) that introduces the option of suspending the judgment if the hypothesis is accepted or rejected with a not sufficient degree of confidence. This is a sort of sequential probability test (as stated in [76]

by Wald) that use additional data for taking the final decision, when it is not possible to make a decision with a sufficient reliability by only using the initial observation. In this case $LR(\mathbf{s})$ is first compared with two different thresholds, say $A_k$ and $B_k$:

in equation (4.2)

$$\Psi(\mathbf{s}) = \begin{cases} 1, & \text{when } LR(\mathbf{s}) > A_k \\ Suspension & \text{when } B_k \leq LR(\mathbf{s}) \leq A_k \\ 0, & \text{when } LR(\mathbf{s}) < B_k \end{cases} \tag{3.5}$$

The thresholds $A_k$ and $B_k$ should be chosen so as to draw an uncertainty region around the value of the threshold $\eta$ given by the standard LR test. In practice, a fraction $\nu$ of this threshold can be chosen, so as $B_k = (1-\nu)\cdot\eta$ and $A_k = (1+\nu)\cdot\eta$. If $LR(s) > A_k$, to turn the decision to advantage the genuine class, while if $LR(s) < B_k$, to turn the decision to advantage the impostor class. In the case of *suspension*, i.e., when $B_k \leq LR(s) \leq A_k$, the test procedure does not make any decision but activates a further step. The suspension of the judgment is motivated by the fact that samples that are quite near to the threshold could be misclassified due to the presence of one biometric trait acquired with a low quality. So, as a second step we propose to adopt the same approach presented in the previous case. In other words, $K$ tests are made on score vectors of $K$-1 dimensions and the hypothesis is refused only if it is refused by all the $K$ voting components. See Fig.3.3.

### 3.2.3 Experiments

A brief description of the two modalities used in our experiments id given below.

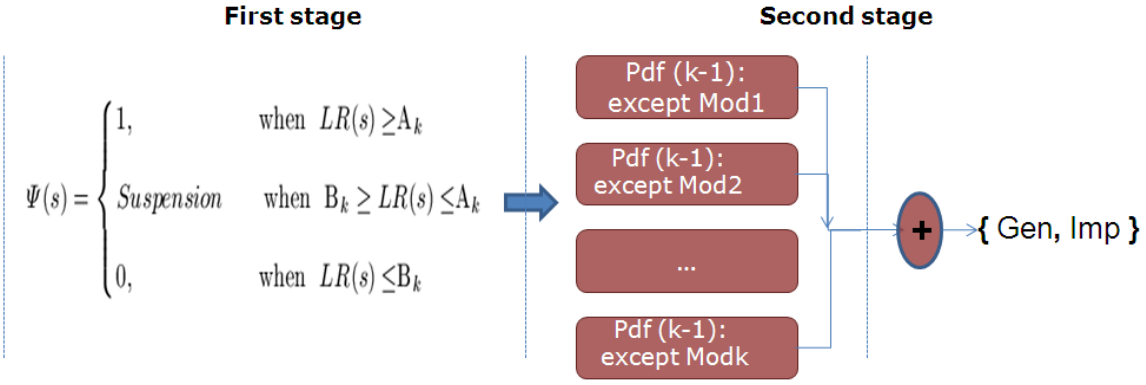


Figure 3.3: The samples classified with low confidence by the LR Standard-based rule are classified a second time by an additional LR voting-based stage.

### Fingerprint Verification

A fingerprint is a pattern of ridges and valleys located on the tip of a finger. Digital images of these patterns are provided by compact sensors (see Fig.3.4). Typically, the features extracted from a fingerprint image are the so called *minutiae points*, which correspond to the position and orientation of ridges endings and bifurcations (see Fig.3.5). The match score is obtained after comparing the set of minutiae extracted from the user's print with those composing the template [61].

Figure 3.4: The optical scanner Fx2000 Biometrika.

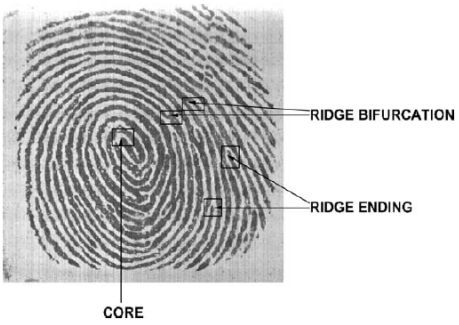


Figure 3.5: Minutiae points extracted from a fingerprint image. They correspond to the position and orientation of ridge endings or bifurcations [61].

### Face Verification

Given a face image, the problem is to verify one or more persons in the scene. This involves a matching between the feature set extracted from a face image and the template stored in the database. A face detection process usually locates the face before the feature extraction. In a controlled environment, enrolled and query images are taken in an uniform background with identical poses and lighting conditions. In uncontrolled environments, factors as different poses, scales, orientations and illuminance conditions, make this process difficult. Moreover, occlusions, facial expressions or emotions, presence of components (e.g., glasses), represent the most challenging problems in face recognition (see Fig.3.6).

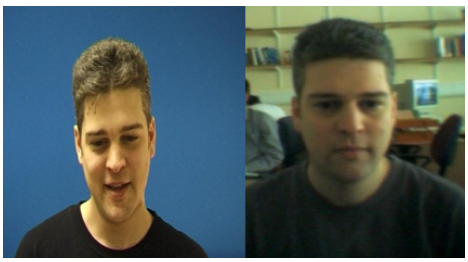

Figure 3.6: Two face images taken from the Banca database. On the left the acquisition of the subject has been performed under controlled conditions, while on the right under uncontrolled conditions.

### Datasets

The performances of our approaches are evaluated on two databases. The first one is a public domain database, namely, NIST-BSSR1 (Biometric Scores Set - Release 1). The BSSR1 is a *true* multimodal database i.e., the face and the fingerprint images coming from the same person at the same time. We performed experiments by employing the first partition made up of face and fingerprint scores belonging to a set of 517 people. For each individual, it is available a score coming from the comparison of two right index fingerprint, a score obtained by comparing impressions of two left index fingerprint, and two scores (from two different matchers, say $C$ and $G$) that are the outputs of the matching between two frontal faces. So, in this case the match score for each modality indicates a *distance*. Then, our first dataset consists in an unbalanced population composed by 517 genuine and 266,772 (517*516) impostor users.

The second database is a subset of the BioSecure multimodal database. This database contains 51 subjects in the Development Set (training) and 156 different subjects in the Evaluation Set (testing). For each subject, four biometric samples are available over two sessions: session 1 and session 2. The first sample of the first session was used to compose the gallery database while the second sample of the first session and the two samples of the second session were used as probes ($P_1$, $P_2$, $P_3$). For the purpose of this study, we used the face and three fingerprint modalities, denoted as *fnf*, *fo1*, *fo2* and *fo3* [54]. The details about the number of match scores per person are reported in Tables 4.2 and 4.3.

Table 3.1: The Biosecure DS2 database: Development Set

| Biometric | Subjects | Samples | Scores |
|---|---|---|---|
| Face | 51 | 4 per subject | Gen $204 \times 3$<br>Imp $51 \times 50 \times 16$ |
| Fingerprint | 51 | 4 per subject | Gen $(204 \times 3) \times 3$<br>Imp $(51 \times 50 \times 16) \times 3$ |

Table 3.2: The Biosecure DS2 database: Evaluation Set

| Biometric | Subjects | Samples | Scores |
|---|---|---|---|
| Face | 156 | 4 per subject | Gen $624 \times 3$<br>Imp $156 \times 155 \times 16$ |
| Fingerprint | 156 | 4 per subject | Gen $(624 \times 3) \times 3$<br>Imp $(156 \times 155 \times 16) \times 3$ |

**Evaluation Procedure**

We have performed a first experiment in which the training set is composed by half of the genuine and half of the impostor randomly selected from the dataset. The rest of the data are used as test set. The second experiment was directed to analyze how the reduction of the available scores for training affects the accuracy of the densities model. So, we performed another test in which the training set is halved with respect to the previous case, while the size of the test set remains unchanged. Both of these training-test partitioning have been randomly repeated 10 times and we report the average performance over the 10 runs.

The current test procedure gives a specific rule for making one of the following decisions: (1)accept the hypothesis to being tested, (2) to reject it, (3) to continue the experiment by making an additional observation by performing an appropriate voting combination on the score On the basis of the pure likelihood ratio test, one of the three decisions above is

made. If the third decision is made, we accept the hypothesis if is accepted by at least one of the voting

**Experimental Results**

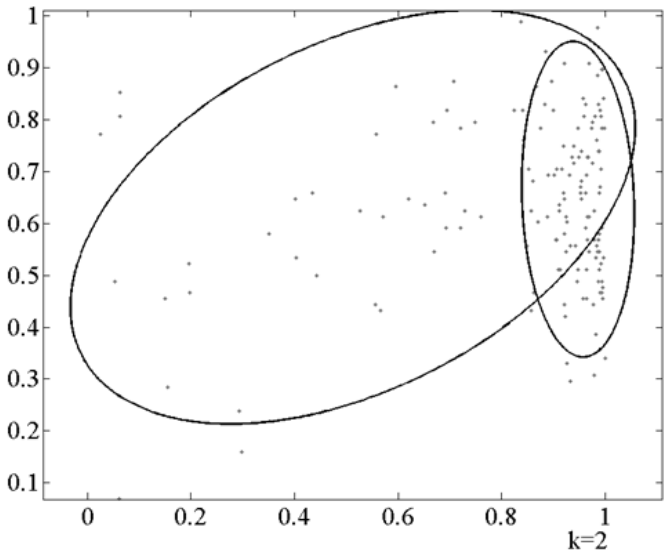


Figure 3.7: Fitting a gaussian mixture: the solid ellipses are level-curves of each component estimate (Biosecure database).

Tables 1 and 2 report the result of the two proposed approaches compared with the standard LR test. Moreover, we also report the $K$-1 dimensional score vector that allowed us to obtain the best results when used alone (in particular this score vector was composed by the outputs of the two fingerprint matchers and of the *Face G* matcher). Three values of $\nu$ have been considered, namely 0.2, 0.25 and 0.30.

Our system was designed for reducing to zero the number of accepted impostors. So, in order to have a fair comparison, the chosen operating point for each run of the standard LR test was obtained by fixing the FAR equal to 0% on the test set. The obtained threshold $\eta$

Table 3.3: Test set results with a training set of equal size (on Nist database)

| | LR | LR on (LfInd, RxInd,FaceG) | Voting LR | Serial LR $\nu = 0.2$ | Serial LR $\nu = 0.25$ | Serial LR $\nu = 0.30$ |
|---|---|---|---|---|---|---|
| **FAR** | 0.0% | 0.0% | 0.0% | 0.0% | 0.0% | 0.000003% |
| **GAR** | 95.60% | 93.26% | 97.77% | 98.22% | 98.22% | 98.30% |

Table 3.4: Test set results with a training set of halved size (on Nist database)

| | LR | LR on (LfInd, RxInd,FaceG) | Voting LR | Serial LR $\nu = 0.2$ | Serial LR $\nu = 0.25$ | Serial LR $\nu = 0.30$ |
|---|---|---|---|---|---|---|
| **FAR** | 0.0% | 0.0003% | 0.0% | 0.000009% | 0.000011% | 0.000011% |
| **GAR** | 81.24% | 95.35% | 98.30% | 88.09% | 88.09% | 88.01% |

is also used in the first step of the *sequential LR* approach.

From the previous tables it is evident that the *sequential LR* always improves the GAR obtainable with a standard LR, since its second stage is able to reduce misclassification of genuine samples with respect to the *pure* likelihood ratio, for those samples classified with a low degree of confidence.

Another interesting results is that the *voting LR* approach seems to be more robust with respect to the lack of training data. When only 25% of the data are used for training, in fact, it is able to significantly improve the GAR with respect to the standard LR approach. In

Table 3.5: Average number of suspended patterns on Nist database

| Training set | Serial LR $\nu = 0.2$ | Serial LR $\nu = 0.25$ | Serial LR $\nu = 0.30$ |
|---|---|---|---|
| 50% | 2.8 | 3.6 | 4.2 |
| 25% | 2.2 | 2.2 | 2.4 |

this case, sequential LR is instead only able to slightly improve the LR performance in terms of GAR, but it also introduces few false accepted samples. On the contrary, when sufficient data for densities estimation are available, sequential LR achieves the best performance. All summarizing, it is worth noting that in both experiments the proposed approaches outperformed the standard LR test when a system at FAR=0% have to be realized.

Finally, is is interesting to consider the score distributions reported in Figures 1 and 2, where the joint distributions of *Left Index*, *Face C* and *Face G* and of *Face C* and *Face G* only are respectively shown. As it is evident (see also the considerations made by [72] on this problem), the use of only two modalities significantly reduces the possibility of distinguish between genuine users and impostors. This is why we did not propose to further iterate the sequential test by considering, for example, also the joint densities of all the possible score pairs.

As the Table 1 shows, at fixed FAR=0%, on the partitioning is 50%-50%, the GAR of the *pure* likelihood ratio is 95.6%, while the GAR of the Sequential test is 98.2%. Moreover, we observe that the implicit use of the quality measures in the fusion scheme improves the accuracy, also when is not available a large data for training, this results in saving of about 50% in the number of observations. In fact, by using the 50% of the data for the training, the GAR of the voting strategy alone is 97.8%, by using the 25% of the data for the training its GAR is 95.6%, then the system is robust to the lack of scores. Finally, the likelihood ratio where the density function is estimated by excluding the Face C modality, the GAR is 96.6% on the partitioning 50%-50%, 95.4% on the partitioning 25%-50%.

Table 3.6: Test set results with a training set of equal size on Biosecure dataset. ($fnf1$: face modality; $fo2$, $fo3$: fingerprint modalities).

| | LR | LR on (fnf1, fo2,fo3) | Voting LR | Serial LR $\nu = 0.2$ | Serial LR $\nu = 0.25$ | Serial LR $\nu = 0.30$ |
|---|---|---|---|---|---|---|
| **FAR** | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| **GAR** | 88.29% | 95.71% | 97.33% | 88.48% | 88.55% | 88.57% |

Table 3.7: Test set results with a training set of halved size on Biosecure dataset. ($fnf1$: face modality; $fo2$, $fo3$: fingerprint modalities).

| | LR | LR on ($fnf1$, $fo2$, $fo3$) | Voting LR | Serial LR $\nu = 0.2$ | Serial LR $\nu = 0.25$ | Serial LR $\nu = 0.30$ |
|---|---|---|---|---|---|---|
| **FAR** | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% |
| **GAR** | 88.01% | 96.58% | 98.29% | 88.14% | 88.14% | 88.29% |

Table 3.8: Average number of suspended patterns (Biosecure database)

| Training set | Serial LR |
|---|---|
| 50% | 1.5 |
| 25% | 1.5 |

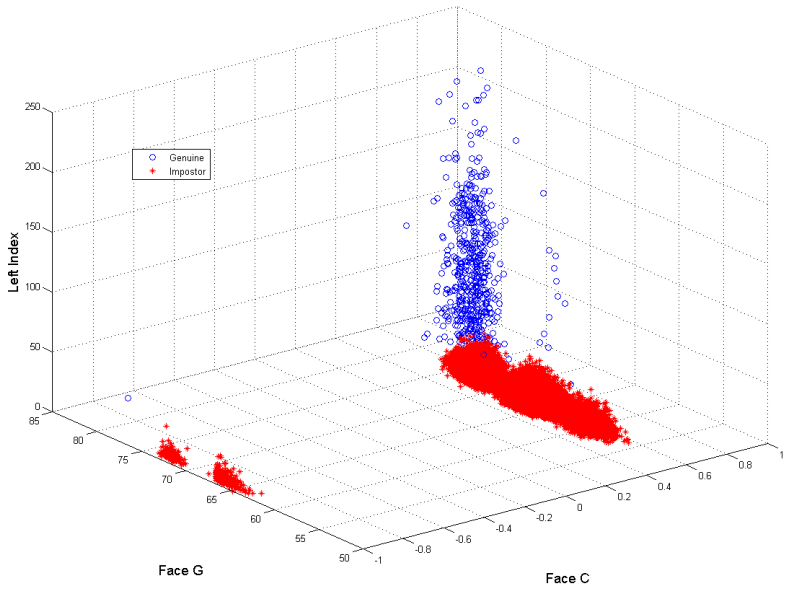


Figure 3.8: Score distribution of Left Index, Face C and Face G from NIST-BSSR1

**Summary**

In this chapter, we have proposed two Likelihood Ratio (LR)-based approaches for combining $K$ biometric modality matchers, in order to minimize the number of false accepted users. We demonstrated that, when the density functions of the standard LR can not be accurately estimated, a voting strategy, involving K density estimations of K-1 modalities, is able to effectively improve the performance of the multimodal system. We demonstrated also that, when the density functions of the standard LR can be accurately estimated, an additional stage, based on the previous voting strategy, can reduce the number of misclassified samples belonging to an uncertainty region, resulting in very good GAR.

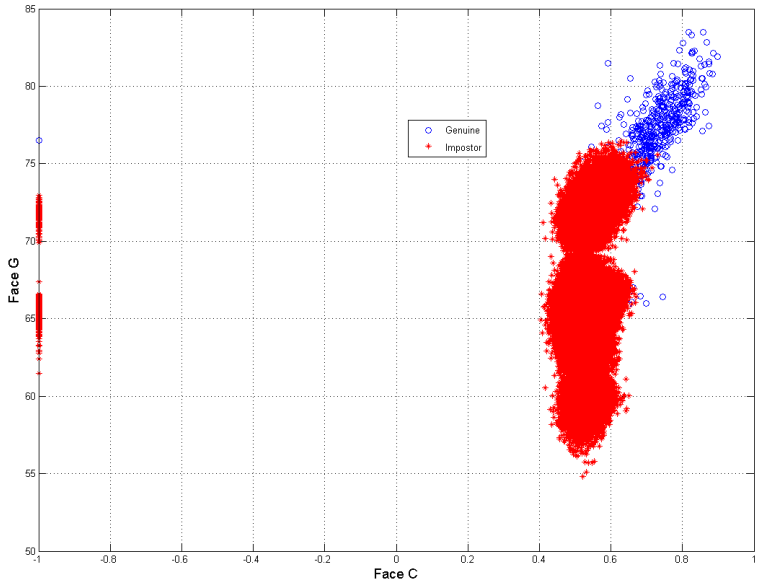


Figure 3.9: Score distribution of Face C and Face G from NIST-BSSR1

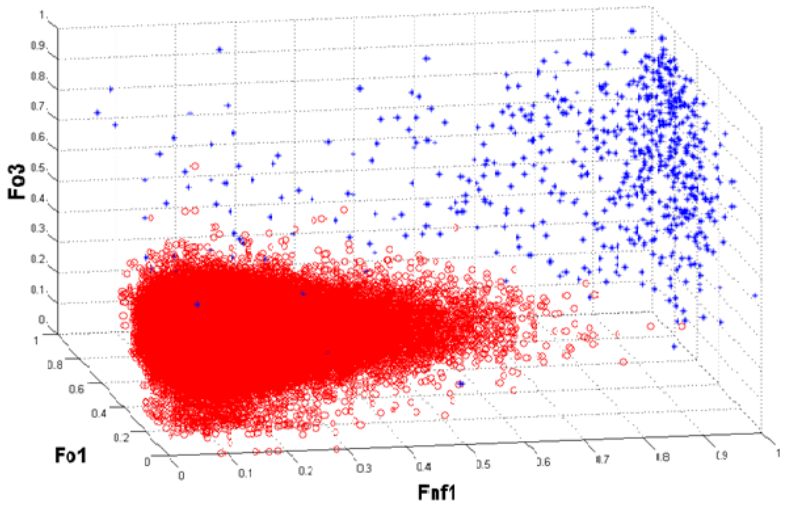


Figure 3.10: Score distribution of $fnf1$ (face), $fo1$ and $fo3$ (fingerprints) from Biosecure database. The red points represent impostor scores while the blue points represent the genuine scores.

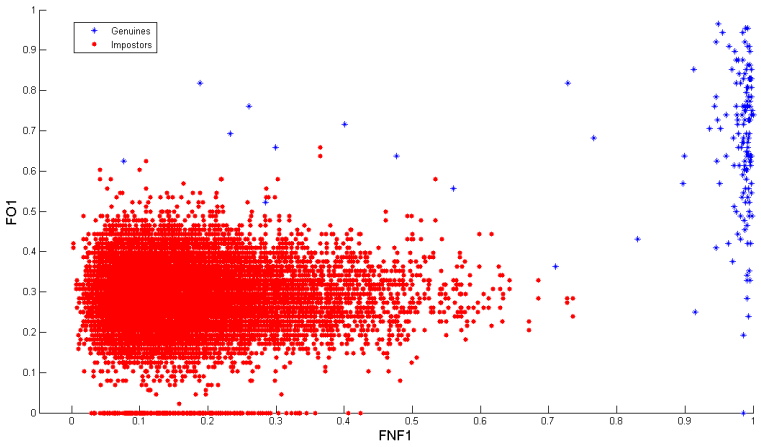


Figure 3.11: Score distribution of $fnf1$ (face), $fo1$ (fingerprint) from Biosecure database.

Scientists are no better than anyone else at forecasting the future. In fact, their predictions are usually wildly inaccurate.

*Robert Wiston*

# Chapter 4

# Multibiometric Identification Scenario

The goal of a biometric *identification* system is to determine the identity of the input biometric data. In such a system, the input probe (e.g., a face image) is compared against a labeled gallery data (e.g., face images in a watch-list) resulting in a set of ranked scores pertaining to the different identities in the gallery database. The identity corresponding to the best score is then typically associated with that of the probe, see Fig.4.1.

In adverse environmental conditions (i.e. illumination changes in a face image) the performance of the unimodal systems may be not efficient [34]. Moreover, in large-scale identification systems, the feature space of the identities in the gallery may significantly overlap resulting in the degradation of identification accuracy. Further, in real scenarios the input data is often noisy, and the similarity between the probe and the associated gallery data is substantially reduced thereby impacting overall recognition accuracy.

This chapter concerns itself with the possibility of automatically determining if the decision rendered by a biometric identification system is correct or not. Our aim is to

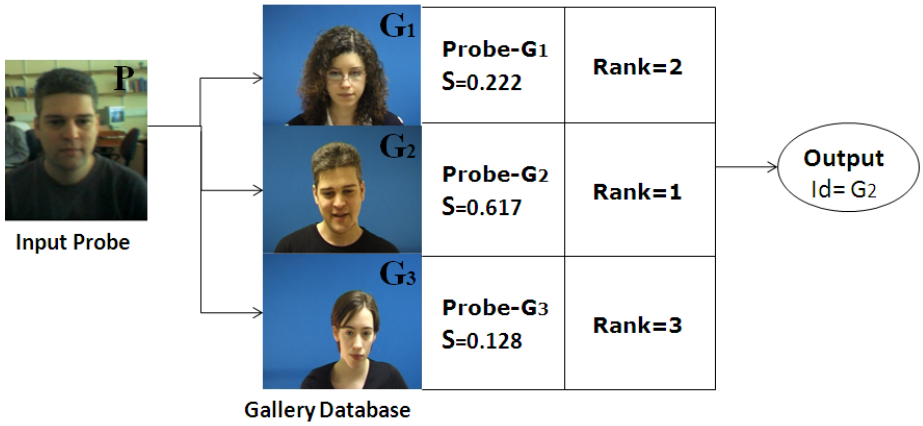


Figure 4.1: The vector of features extracted from the probe image is compared against all the templates stored in the database. A set of scores is generated and sorted. The one rank value is assigned to the high similarity match score and the corresponding identity is chosen as output of the system. The face images have been taken from the BANCA database.

predict identification errors and improve the recognition accuracy of the biometric system. Our method utilizes the rank and score information generated by the identification operation in order to validate the output. Further, we demonstrate that the proposed predictor can be effectively applied in multimodal scenarios. Experiments performed on two multimodal databases show the effectiveness of our framework in improving identification performance of biometric systems. Finally, we investigate the question of whether it is possible to improve the performance of the identification system by using the non-matched scores, referred to as *neighbors* of the rank one identity. Our case study concerns the combination of face and fingerprint recognition systems at hybrid rank-score level, as shown in Fig.4.2.

## 4.1 Predicting Identification Errors

This section focuses on reducing identification errors of multibiometric systems by involving in the fusion scheme only outputs which are not degraded [30]. In real applications it is

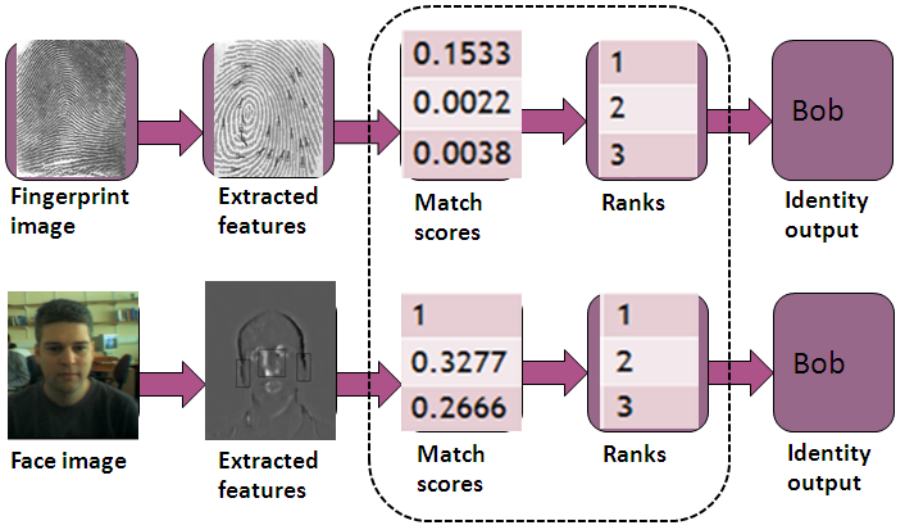


Figure 4.2: Combination of face and fingerprint modalities.

necessary that the unimodal system produces an estimate of decision reliability. This information corresponds to the conditional probability that the decision made by the unimodal system, given the available evidence $E$, is correct [33]. Estimating the level of trust in the correctness of the decision can offer a feedback which may aid to appropriately weight unimodal results in a fusion scheme. An instance of error, corresponding to a False Rejection, occurs when a legitimate user is not able to achieve a good enough similarity score to match its entry in the gallery. A more dangerous error, corresponding to a False Acceptance, occurs when an impostor has achieved a match within the gallery [68].

In the recent literature, quality and non-quality based approaches for biometric system failure prediction have been proposed. Based on the correlation between quality and recognition performance, quality was considered a good indicator in many studies in which quality was promoted as a predictor of failure. Traditional biometric evaluation relied on a

notion of quality associated to the raw image quality to determine the performance of the system [67]. A failure condition corresponds to the presence of a low quality image input.

This notion of failure prediction is limiting since in presence of a low quality image the system has to acquire another image sample. Such a scheme cannot be applied without a real-time indication about the quality of the data, as it usually happens with multimodal biometrics. Moreover, this notion of quality assessment was confuted in [68] by showing cases in which, given a subject, *poor* quality images produce better matching scores than *high* quality images.

An interesting alternative to quality analysis was presented by Scheirer and Boult in [67], in which they proposed the idea of post-recognition failure predictor. Such a failure predictor is able to learn when a system fails and when it succeeds, and to predict which input is more likely to fail. Based on the decisions made from a classification system, they defined two types of error, i.e. a Failure Prediction False Accept Rate (FPFAR) and a a Failure Prediction False Reject Rate (FPFRR) and the Failure Prediction Receiver Operator Characteristic (FPROC). This prediction analysis has been shown to be effective for single modalities and able to enhance the overall performance when exploited in fusion schemes [68].

They introduced failure prediction features derived from similarity scores and designed to capture distributional information that is not represented from just a raw score. They extracted the differences between scores and the DCT coefficients after transforming the top $n$ scores. The failure prediction analysis of their system predicts individual modality failures and drives the fusion weighting them. In the work, the authors have presented a multi-modal

recognition system integrating fusion-based failure prediction. The proposed multi-stage architecture presents the fusion module at the highest level to integrate the results of failure prediction across modalities. Four different fusion techniques were proposed to improve failure prediction, three of them are able to improve failure prediction, and subsequently, the recognition system performance. The described approach consists of a fusion via prediction. According to the proposed multi-modal failure prediction, if one modality has failed, it is possible to fuse information from another one that has succeeded; this lets to achieve good recognition performance. They firstly evaluated the performance of four fusion techniques for failure prediction, then they evaluated the failure prediction fusion-based recognition system. The usage of DCT transform made quite complex the feature extraction, and subsequently, the time needed for training the system expensive. Our approach simplifies this aspect by exploiting rank and score information for predicting errors and, subsequently, improving the performance of the biometric system. In the proposed methodology, the probability that the output decision is reliable is estimated by a pattern classifier referred to as a *predictor*. Its role is to detect potentially erroneous decisions. Further, we propose three fusion mechanisms based on the trained predictor that can extend the benefits of the proposed scheme into a multimodal scenario. In particular, a predictor-based voting strategy, a predictor-based serial fusion scheme and a predictor-based Borda Count method are presented and compared against other common approaches to *rank-level* fusion.

The idea of marginalizing potentially incorrect decisions in a pattern recognition system was used by Chow [8] to define an optimum rejection rule. In the pattern recognition and machine learning literature, several techniques have been proposed to predict the reliability

of a classification decision rendered by a pattern recognition system (e.g., [10]). However, such methods have been sparingly used in the biometric literature until recently. Kryszczuk et al. [34] presented a method in which classifier decisions and the corresponding reliability information are combined to predict and correct verification decisions. Kryszczuk et al. [33] later proposed a framework for probabilistic error rectification based on credence estimation which was used to eliminate unreliable verification decisions.

### 4.1.1 Analysis Ratio-based

As stated earlier, a generic identification system compares the input biometric data to all the known identities stored in the database and outputs a set of similarity scores. The scores are then sorted in decreasing order to form a ranking list in which the lowest rank is assigned to the highest similarity [1]. Let $\mathbf{G} = [G_1, G_2, ...G_N]$ be the *gallery* set, composed by $N$ biometric samples belonging to $N$ different subjects. Let $\mathbf{P} = [P_1, P_2, ...P_M]$ be the *probe* set, composed by $M$ *unknown* samples belonging to subjects that are presumed to be in the *gallery*. Given a single probe image, $N$ comparisons of that probe against the gallery are performed and $N$ similarity scores are generated [5].

The present study is based on computing the ratio of scores corresponding to rank 1 and the other ranks. The vector of these ratios is treated like a feature vector and used for training a pattern classifier. Such a classifier is used to learn the relationship between the ratios and the *posterior* probabilities of the *correct* and *error* classes. Here, the term "correct class" is used to indicate that the rank-1 identity is indeed the correct identity of the probe; the term "error class" is used to indicate that the rank-1 identity does not

correspond to the correct identity of the probe. Thus, the classifier (*predictor*) is used to learn the decision boundary between the correct identification region and the erroneous one [39].

For a given input probe, let $\rho_j$ denote the ratio of the rank-1 score to that of the rank-j score. Thus, the vector $(\rho_2, \rho_3, \ldots \rho_{d+1})^t$, $d \in \{1 \ldots N-1\}$, is used as input to the classifier. Typically, the rank-1 similarity score is expected to be significantly higher than the other scores (for a genuine match at rank-1). However, there are situations when the rank-1 score may be comparable to that of other scores associated with the nearby ranks thus suggesting the *possibility* of an error. In this work, we confirm this notion and, further, exploit it to improve recognition accuracy.

---

**Algorithm for training the unimodal predictor using ratios**
Let $\mathbf{G} = [G_1, G_2, ...G_N]$ be the *gallery* set.
Let $\mathbf{P} = [P_1, P_2, ...P_M]$ be the *probe* set.

1. For each probe, generate $N$ similarity match scores $s_i$, $i = 1 \ldots N$ by comparing that probe against the gallery.
2. Sort the match scores in decreasing order.
3. Based on the previous sorted match scores, assign a rank $R_i$ to each enrolled identity.
4. Compute the ratio $\rho_j$ between the score corresponding to rank-1 and the score corresponding to rank-j.
5. Label the ratio score vector as *correct* if rank 1 is assigned to the correct identity by the unimodal matcher; otherwise label it as an *error*.
6. Use the labeled ratio score vectors as feature vectors to train a supervised classifier.

---

Fig. 4.3 shows the architecture of the proposed approach.

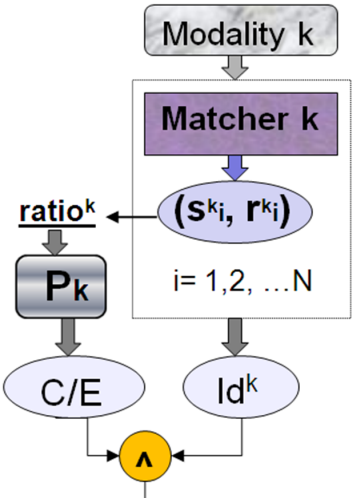


Figure 4.3: Error Prediction in a unimodal identification system. Here, $s_i^k$ and $r_i^k$ denote the score and rank, respectively, assigned to the $i^{th}$ identity in the gallery by the $k^{th}$ matcher; $P_k$ denotes the classifier used to predict if the rank-1 identification is correct (C) or not (E) based on the vector of score ratios (ratio$^k$). The output of the matcher, $Id^k$, is accepted or rejected based on the predictor.

### 4.1.2 Differences-based Analysis

The present study is based on computing the difference of scores corresponding to rank 1 and the other ranks. The vector of these differences is treated like a feature vector and used for training a pattern classifier. Such a classifier is used to learn the relationship between the ratios and the *posterior* probabilities of the *correct* and *error* classes. Here, the term "correct class" is used to indicate that the rank-1 identity is indeed the correct identity of the probe; the term "error class" is used to indicate that the rank-1 identity does not correspond to the correct identity of the probe. Thus, the classifier ( *predictor*) is used to learn the decision boundary between the correct identification region and the erroneous one.

For a given input probe, let $\delta_{ij}$ denote the difference of the rank-i score to that of the

rank-j score. Thus, the vector $(\delta_2, \delta_3, \ldots \delta_{d+1})^t$, $d \in \{1 \ldots N-1\}$, is used as input to the classifier. The distribution of the differences between scores in terms of ranks gives the information about the direction of the largest variance for each modality. Typically, the rank-1 similarity score is expected to be significantly higher than the other scores (for a genuine match at rank-1). However, there are situations when the rank-1 score may be comparable to that of other scores associated with the nearby ranks thus suggesting the *possibility* of an error. In this work, we confirm this notion and, further, exploit it to improve recognition accuracy. The difference which presents the highest variability can be projected in the space of two modalities to analyze the separation between the classes *error* and *correct*.

---

**Algorithm for training the unimodal predictor using differences**
Let $\mathbf{G} = [G_1, G_2, ...G_N]$ be the *gallery* set.
Let $\mathbf{P} = [P_1, P_2, ...P_M]$ be the *probe* set.

1. For each probe, generate $N$ similarity match scores $s_i$, $i = 1 \ldots N$ by comparing that probe against the gallery.
2. Sort the match scores in decreasing order.
3. Based on the previous sorted match scores, assign a rank $R_i$ to each enrolled identity.
4. Compute the differences $\delta_{ij}$ between the score corresponding to rank-i and the score corresponding to rank-j.
5. Label the difference score vector as *correct* if rank 1 is assigned to the correct identity by the unimodal matcher; otherwise label it as an *error*.
6. Use the labeled difference score vectors as feature vectors to train a supervised classifier.

---

## 4.2 A Predictor-based Framework

In a multibiometric identification system, the output of $K$ different biometric modality matchers $C_1, C_2, ...C_K$ have to be consolidated. The information observed at the score level can be represented as a $N \times K$ matrix $\mathbf{S} = [s_n^k]$, where $s_n^k$ represents the match score output when the probe image is compared against the $n^{th}$ gallery image, using the $k^{th}$ classifier, $k = 1, ..., K$; $n = 1, ..., N$. This score matrix can be converted to a rank matrix $\mathbf{R} = [r_n^k]$ where $r_n^k$ represents the rank of the $n^{th}$ gallery image with respect to the probe as assessed by the $k^{th}$ modality matcher.

### 4.2.1 Predictor-based Majority Voting

In the majority voting scheme, the outputs of the $K$ classifiers are examined and the most commonly occurring output is selected as the final output. Thus, for a given probe, $K$ unimodal matchers are employed and the winner is the identity to which the majority of matchers have assigned a rank value equal to one. The majority vote will result in an ensemble decision [35]:

$$\arg \max_{i=1...N} \sum_{k=1}^{K} d_{ik} \cdot v_k \tag{4.1}$$

where the binary variable $d_{ik}$ is 1 if the $k^{th}$ matcher outputs identity $i$ in rank-1, and the binary variable $v_k$ is 1 if the identification is deemed to be *correct* by the $k^{th}$ predictor. Fig. 4.4 presents this scheme. The majority vote scheme will assign an identity to the probe only if the output of at least $\lfloor \frac{1}{2} \Sigma_{k=1}^{K} v_k \rfloor + 1$ unimodal systems correspond to the same identity and are deemed to be correct by $v_k$. For example, suppose there are 5 unimodal systems. Assume that for the identity corresponding to the true user (say, 'Bob'), 3 out of 5 systems

output the identity 'Alice', while the other two output 'Bob'. Suppose that 2 out of the 3 predictors state that the output 'Alice' is in error, while the others indicate that their respective outputs were correct, then the final output of the predictor-based multimodal system will be 'Bob'. If a majority is not possible, then the proposed mechanism attempts to use the rank-1 accuracy of individual classifiers to make the decision. According to this design, when the unimodal outputs are $K$ different identities, the output from the overall system will correspond to the identity output from the unimodal system with the highest accuracy (as assessed using training data before deployment of individual matchers). Those contributions considered as errors by the predictor module are excluded from the final decision.

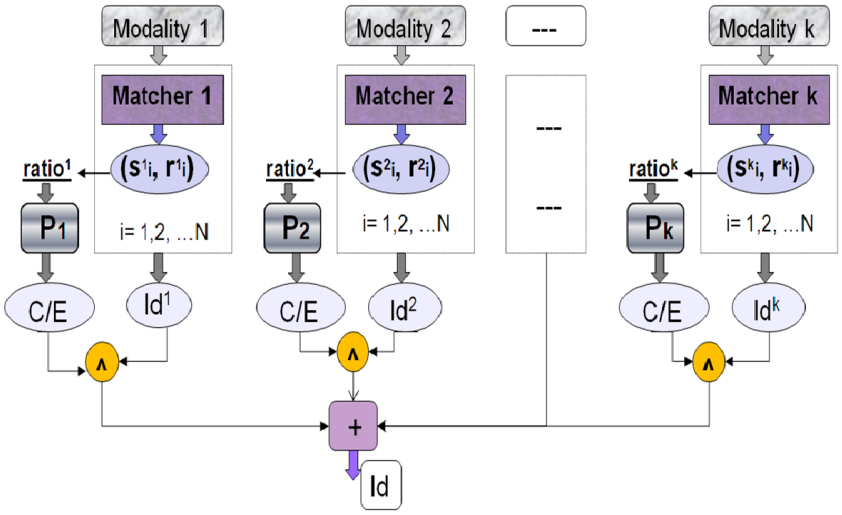


Figure 4.4: Predictor-based Majority Voting.

### 4.2.2 Predictor-based Serial Scheme

In the serial scheme, the decisional process is split into two successive stages [43]. The subject to be authenticated submits the first biometric modality to the system which is processed and matched against all the templates present in the gallery. If the resulting identity is labeled to be correct by the predictor module, the input biometric trait is associated to the current identity, otherwise the system suspends the decision and an additional processing stage is performed. In the second stage, *K-1* additional biometric modalities are automatically requested and a voting strategy involving *K-1* unimodal matchers is adopted in the second stage. The described predictor-based serial combination framework is shown in the Fig. 4.5. It can be formulated as follows:

$$Id_m = \begin{cases} Id_u, & \text{if} \quad v_u = 1 \\ \arg\max_{i=1\ldots N} \sum_{k=1}^{K-1} d_{ik} \cdot v_k & \text{if} \quad v_u = 0 \end{cases} \tag{4.2}$$

where $Id_m$ is the output of the multimodal system and $Id_u$ is the output of the unimodal system at the first stage.

### 4.2.3 Predictor-based Borda Count

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

$$R_i = \sum_{k=1}^{K} w_k \, r_{ik}, \qquad i = 1, 2, \ldots N \tag{4.3}$$

This method assigns a higher weight to the ranks provided by the more accurate matcher. Therefore, it is useful when different biometric matchers exhibit significant differences in

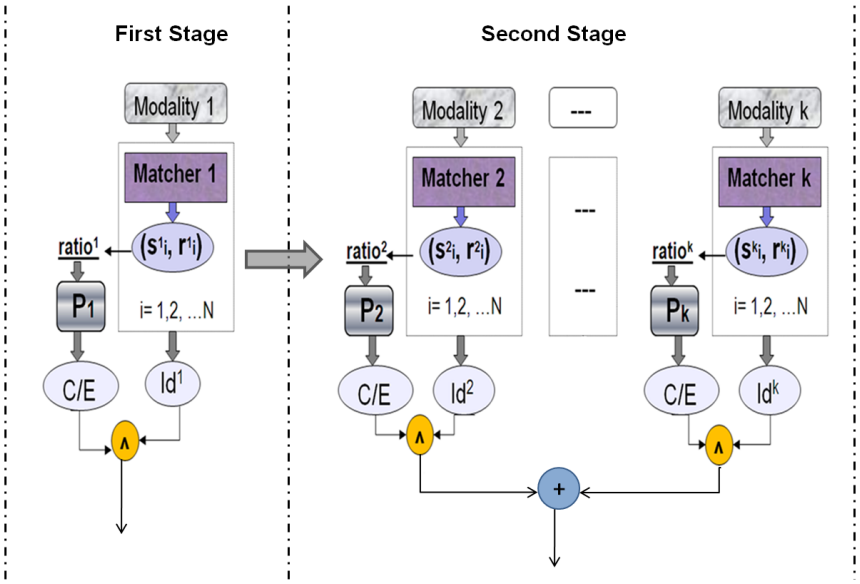


Figure 4.5: Predictor-based Serial Fusion: the first stage is based on the unimodal system and the error predictor for this modality while the second stage consists of a predictor-based majority voting scheme which uses K-1 modalities.

their accuracies. A training phase has to be performed to determine the weights. In the proposed predictor-based fusion scheme, the unimodal outputs labeled as errors by the predictor have to be excluded from the sum in the equation above which determines the fused rank for each identity. This can be achieved by computing the weight $w_k$ as the ratio between the number of correct identifications detected by the predictor and the total number of test probes. follows:

$$w_k = \frac{v_{ik}}{\sum_{k=1}^{K} v_{ik}} \qquad i = 1, 2, ...N \tag{4.4}$$

The weight factor based on the predictor reduces the effect of inaccurate decisions provided by potentially incorrect matchers. As an example, consider the fusion of 5 modality matchers. Assume that for the identity corresponding to the true identity (say, $i = 1$), 4 out of the

Table 4.1: WVU Multimodal Biometric Database

| Biometric | Subjects | Samples | Scores |
|---|---|---|---|
| Face | 240 | 5 per subject | Gen $1200 \times 4$<br>Imp $240 \times 239 \times 25$ |
| Fingerprint | 240 | 5 per finger | Gen $(1200 \times 4) \times 4$<br>Imp $(240 \times 239 \times 25) \times 4$ |

5 matchers result in rank 1, while the $5^{th}$ results in rank 5. If the outputs of the predictors are 1 for the first 4 modalities and 0 for the last one, the final rank will be 4.

### 4.2.4 Performance Evaluation

#### Datasets

In the present thesis, we have considered a multimodal identification system that integrates fingerprint and face experts. The performance of the proposed strategy was evaluated on two databases. The first is the West Virginia University (WVU) multimodal biometric database. A subset of this database pertaining to the fingerprint (left thumb [FL1], right thumb [FR1], left index [FL2], right index [FR2]) and face modalities of 240 subjects was used in our experiments. Five samples per subject for each modality were available. Table 4.1 provides the details of the database. For the *face* modality, frontal images were collected in a controlled scenario. For the *fingerprint* modality, images were collected using an optical biometric scanner, without explicitly controlling the quality [11]. The entire dataset was divided into five sets: the first sample of each identity was used to compose the *gallery* and the remaining four samples of each identity were used as *probes* ($P_1$, $P_2$, $P_3$, $P_4$). The VeriFinger software was used for generating the fingerprint scores and the VeriLook software was used for generating the face scores.

Table 4.2: The Biosecure database: Development Set

| Biometric | Subjects | Samples | Scores |
|---|---|---|---|
| Face | 51 | 4 per subject | Gen $204 \times 3$<br>Imp $51 \times 50 \times 16$ |
| Fingerprint | 51 | 4 per subject | Gen $(204 \times 3) \times 3$<br>Imp $(51 \times 50 \times 16) \times 3$ |

Table 4.3: The Biosecure database: Evaluation Set

| Biometric | Subjects | Samples | Scores |
|---|---|---|---|
| Face | 156 | 4 per subject | Gen $624 \times 3$<br>Imp $156 \times 155 \times 16$ |
| Fingerprint | 156 | 4 per subject | Gen $(624 \times 3) \times 3$<br>Imp $(156 \times 155 \times 16) \times 3$ |

The second database is a subset of the BioSecure multimodal database. This database contains 51 subjects in the Development Set (training) and 156 different subjects in the Evaluation Set (testing). For each subject, four biometric samples are available over two sessions: session 1 and session 2. The first sample of each subject in the first session was used to compose the gallery database while the second sample of the first session and the two samples of the second session were used as probes ($P_1$, $P_2$, $P_3$). For the purpose of this study, we used the face and three fingerprint modalities, denoted as *fnf*, *fo1*, *fo2* and *fo3*, respectively [54]. The details about the number of match scores per person are reported in Tables 4.2 and 4.3.

**Evaluation Procedure**

First, we performed a preliminary analysis to understand the distribution of the ratios between scores as a function of the ranks (i.e., the $\rho_k$'s) for the correct and error classes.

This was used to determine the dimension of the vector of ratios (i.e., $d$) that is suitable for error prediction. The number $d$ was empirically derived for each modality in the individual databases considered in this work. Next, the proposed algorithm was evaluated on the two databases. Since the number of identification errors made by some of the biometric matchers is low, the negative class cannot be efficiently represented. This affects the training of the predictor. In order to maximize the amount of available data, the training and testing was performed by adopting the *leave-one-out* strategy. The classifier was trained by using the samples provided by all but one of the identities in the gallery and its performance was tested on the excluded identity [6].

**Results**

As Fig. 4.6 and Fig. 4.7 show, in the space of ratios, the distributions of the misclassified identities are reasonably separated from those that were correctly recognized.

Fig. 4.11 and Fig. 4.10 show that, for both modalities face and fingerprint, the difference between the rank-1 score and rank-2 score, here referred to *r1r2*, presents the highest variance.

We have also plotted the difference *r1r2* in the space of face and fingerprint modalities, as shown in Fig. 4.12, and the correct identifications are well separated from the errors, as assessed by the modality matchers.

The classification was accomplished using three different classifiers: a *Support Vector Machine (SVM)*, a *Decision Tree* and a *Bayesian* classifier. Since the SVM classifier gave the best results on both databases, only its performance is being reported in this chapter.

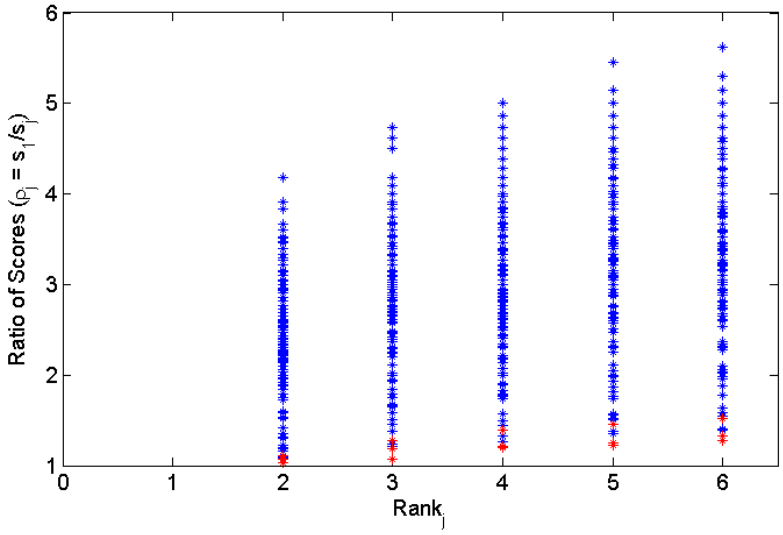


Figure 4.6: The distribution of the ratios between scores in terms of ranks of all the users in the WVU database for the face modality, where the gallery set is composed by the first sample of each subject and the probe set by the fifth sample. Red points represent rank-1 misclassifications.

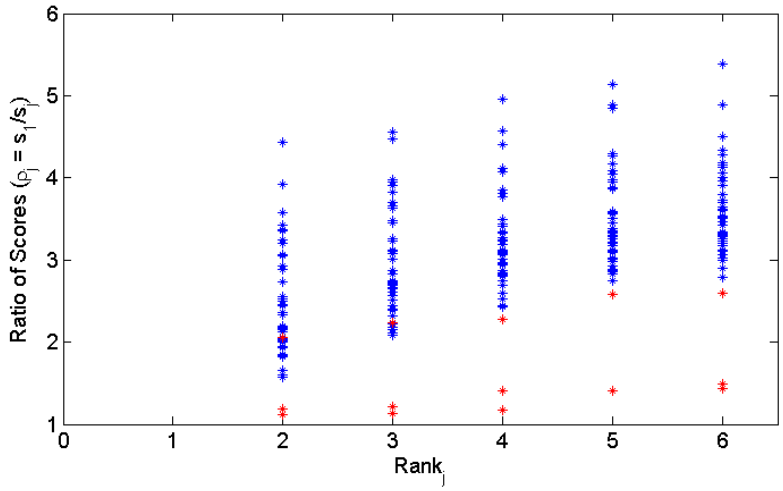


Figure 4.7: The distribution of the ratios between scores in terms of ranks of all the users of the Development Set in the Biosecure database for the face modality, where the gallery set is composed by the first sample of each subject and the probe set by the second sample. Red points represent rank-1 misclassifications.

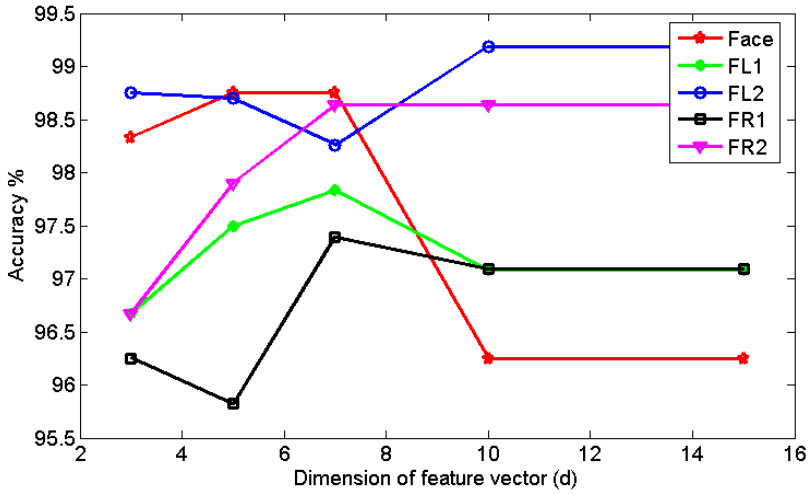


Figure 4.8: Performance of the prediction scheme using a Support Vector Machine trained on the WVU data.

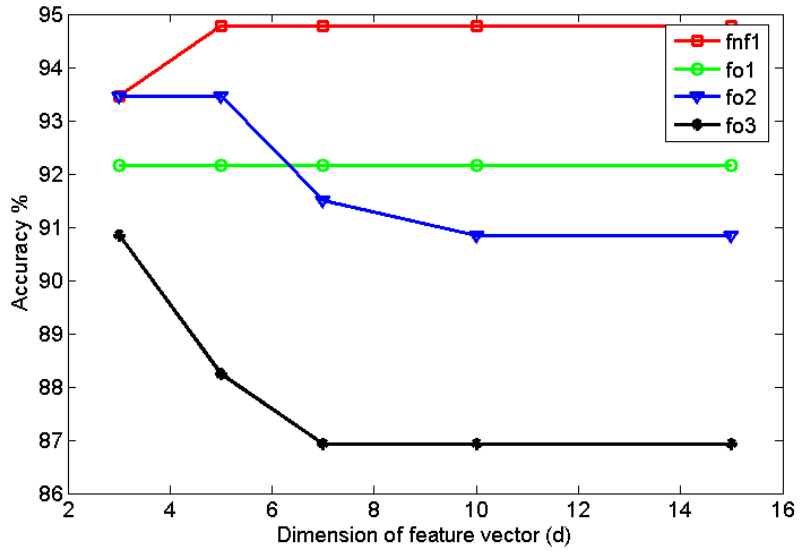


Figure 4.9: Performance of the prediction scheme using a Support Vector Machine trained on the Biosecure data.

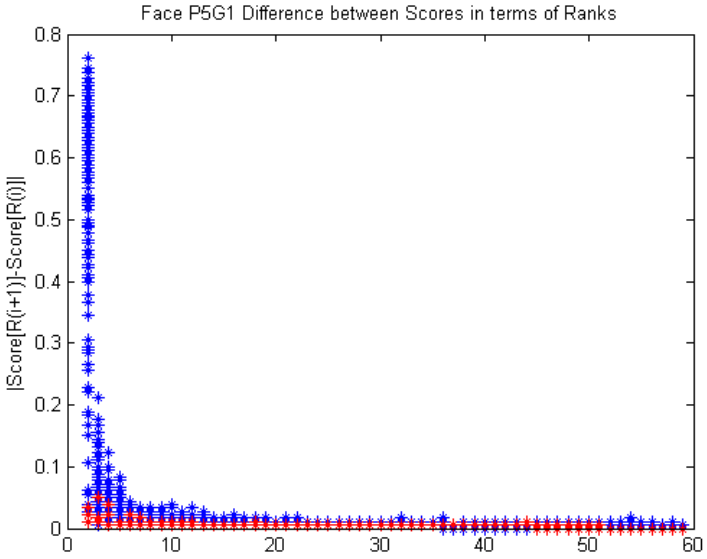


Figure 4.10: The distribution of the differences between scores in terms of ranks of all the users in the WVU database for the face modality, where the gallery set is composed by the first sample of each subject and the probe set by the fifth sample. Red points represent rank-1 misclassifications.

Further, the classification performance was observed as a function of $d$, i.e., the number of ratios used to construct the feature vector. The face modality in the WVU database required $d = 5$; the FL1, FR1 and FR2 modalities required $d = 7$ and the fingerprint FL2 modality required $d = 10$ (see Figure 4.8). For the Biosecure dataset, all the 3 fingerprint modalities required $d = 3$ while the face required $d = 5$ (see Figure 4.9).

Tables 4.4, 4.5, 4.7 and 4.8 compare the results of the proposed scheme against other schemes. We compared the performance of our methods against the *Highest Rank* and *Borda Count* approaches [63] as well as the *pure* Majority Voting Scheme in which the predictor for each modality was not used (ties were broken randomly). From these tables it is evident that the *predictor-based majority voting* which uses the predictor for each modality, outperformed

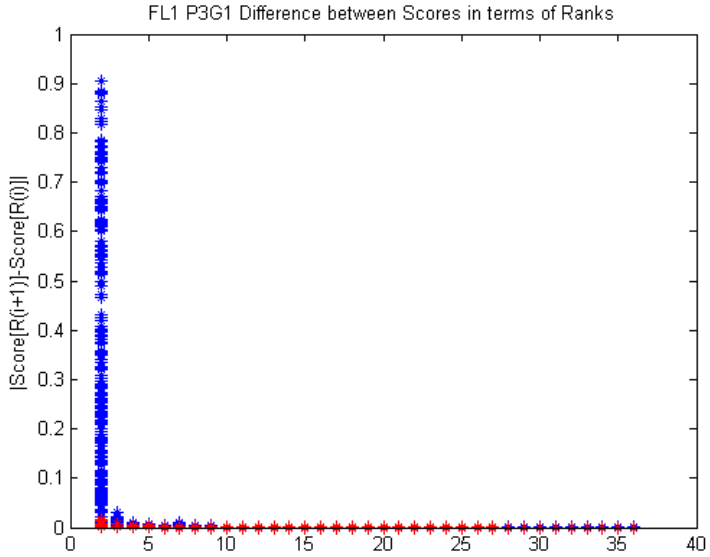


Figure 4.11: The distribution of the differences between scores in terms of ranks of all the users in the WVU database for the fingerprint modality, where the gallery set is composed by the first sample of each subject and the probe set by the third sample. Red points represent rank-1 misclassifications.

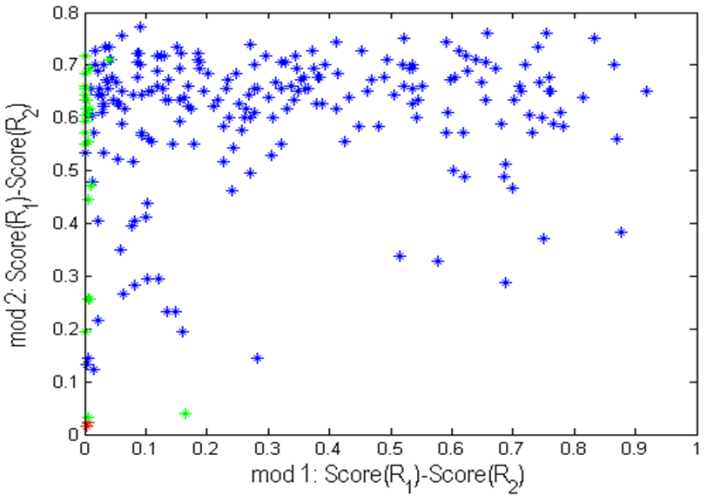


Figure 4.12: The distribution of the differences between scores in terms of ranks of all the users in the WVU database for the face and fingerprint modalities, where blue points represent a correct identification as assessed by both modalities, while green and red points represent cases in which the unimodal labels about a potential error are contrasting.

Table 4.4: Performance of the traditional fusion schemes on the four probe sets in the WVU database.

| Probe | Highest Rank | Borda Count | Pure Majority Voting |
|---|---|---|---|
| P1 | 91.67% | 97.22% | 100.00% |
| P2 | 88.33% | 95.56% | 99.44% |
| P3 | 90.56% | 96.11% | 97.78% |
| P4 | 93.33% | 96.67% | 99.44% |
| *Avg* | *90.97%* | *96.39%* | *99.17%* |

Table 4.5: Performance of the predictor-based fusion schemes on the four probe sets in the WVU database, where the predictor was training using ratio score vectors

| Probe | Predictor-based Majority Voting | Predictor-based Serial | Predictor-based Borda Count |
|---|---|---|---|
| P1 | 100.00% | 100.00% | 97.22% |
| P2 | 100.00% | 99.44% | 96.11% |
| P3 | 100.00% | 99.44% | 96.11% |
| P4 | 100.00% | 98.89% | 97.22% |
| *Avg* | *100.00%* | *99.44%* | *96.67%* |

the other traditional approaches. Moreover, the serial scheme also improved the correct identification rate since, in the second stage, it is able to handle those cases that are classified as *errors* in the first stage. We also observed that the improvement in performance was especially significant in the case of the BioSecure Database where traditional rank-level fusion schemes did not perform very well.

Table 4.6 reports results of two predictor-based fusion scheme, where the predictor has been trained by using the difference score vectors. From these tables it is evident that the *predictor-based majority voting* which uses the predictor for each modality, outperformed the other traditional approaches. Moreover, the serial scheme also improved the correct

Table 4.6: Performance of the predictor-based fusion schemes on the four probe sets in the WVU database, where the predictor was training by using difference score vector

| Probe | Predictor-based Majority Voting | Predictor-based Serial |
|---|---|---|
| P1 | 100.00% | 100.00% |
| P2 | 100.00% | 99.58% |
| P3 | 100.00% | 100.00% |
| P4 | 100.00% | 99.58% |
| *Avg* | *100.00%* | *99.79%* |

Table 4.7: Performance of the traditional fusion schemes on the three probe sets in the Biosecure database

| Probe | Highest Rank | Borda Count | Pure Majority Voting |
|---|---|---|---|
| $P_1$ | 87.18% | 96.15% | 89.74% |
| $P_2$ | 78.85% | 88.46% | 83.97% |
| $P_3$ | 74.36% | 92.31% | 84.62% |
| *Avg* | *80.13%* | *92.31%* | *86.11%* |

identification rate since, in the second stage, it is able to handle those cases that are classified as *errors* in the first stage.

### 4.2.5 Cross-Validation Evaluation

The training and testing of the error prediction scheme was also performed by adopting the *cross validation* strategy to maximize the amount of available data during the training phase. The classifier was trained over 5 iterations by using the samples provided by the 25% of the identities in the gallery and its performance was tested on the excluded identities [6].

The classification was accomplished using three different classifiers: a *Support Vector Machine (SVM)*, a *Decision Tree* and a *Bayesian* classifier. Since the Decision Tree classifier

Table 4.8: Performance of the predictor-based fusion schemes on the three probe sets in the Biosecure database, where the predictor was training using ratio score vectors

| Probe | Predictor-based Majority Voting | Predictor-based Borda Count | Predictor-based Serial |
|---|---|---|---|
| $P_1$ | 100.00% | 96.15% | 100.00% |
| $P_2$ | 94.23% | 89.10% | 94.87% |
| $P_3$ | 97.44% | 92.31% | 94.87% |
| *Avg* | *97.22%* | *92.52%* | *96.58%* |

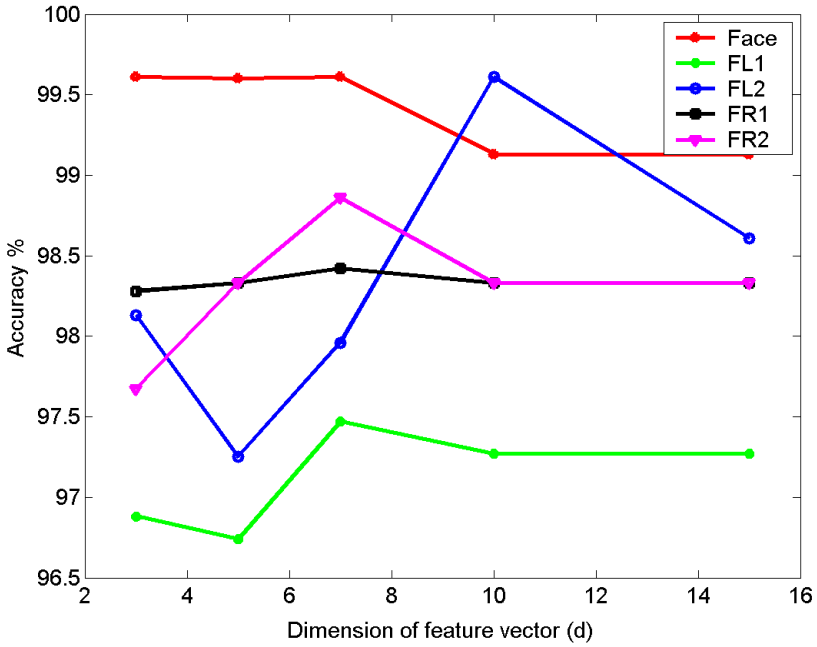


Figure 4.13: Performance of the prediction scheme using a Decision Tree trained on the WVU data, where the predictor was training using ratio score vectors.

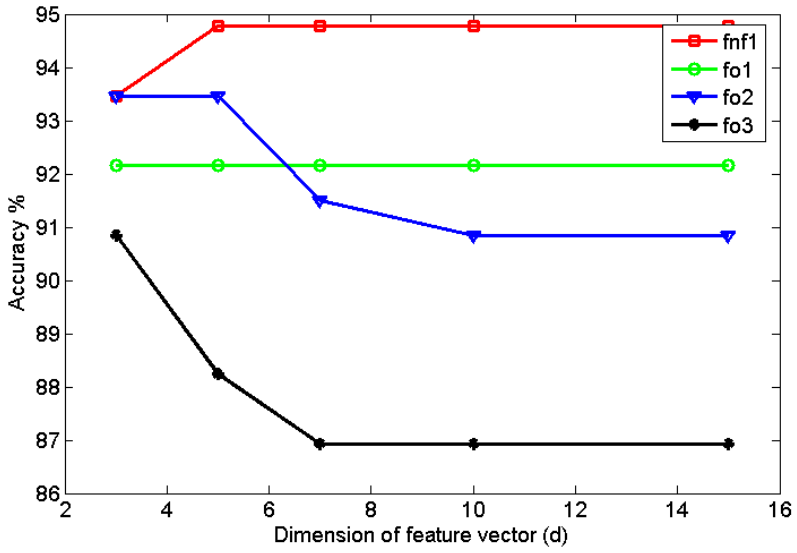


Figure 4.14: Performance of the prediction scheme using a Support Vector Machine trained on the Biosecure data, where the predictor was training using ratio score vectors.

gave the best results on WVU database and the SVM classifier gave the best results on Biosecure database, only their performances are being reported in this chapter. Further, the classification performance was also observed as a function of $d$, i.e., the number of ratios used to compose the feature vector. In fact, an instance of error may occur at prediction level too, since an error can be made by the predictor itself. The face modality in the WVU database required $d = 5$; the FL1, FR1 and FR2 modalities required $d = 7$ and the fingerprint FL2 modality required $d = 10$ (see Figure 4.13). For the Biosecure dataset, all the 3 fingerprint modalities required $d = 3$ while the face required $d = 5$ (see Figure 4.14).

Tables 4.9, 4.10, 4.11 and 4.12 compare the results of the proposed scheme against other schemes. We compared the performance of our methods against the *Highest Rank* and *Borda Count* approaches [63] as well as the *pure* Majority Voting Scheme in which the

Table 4.9: Performance of traditional fusion schemes on the four probe sets in the WVU database. The accuracy has been evaluated by 5-fold cross validation and the classification rates have been averaged.

| Probe | Highest Rank | Borda Count | Pure Majority Voting |
|---|---|---|---|
| P1 | 93.89% | 92.89% | 99.33% |
| P2 | 91.78% | 91.67% | 98.89% |
| P3 | 91.67% | 90.78% | 98.11% |
| P4 | 91.67% | 90.78% | 98.11% |
| *Avg* | *92.25%* | *91.53%* | *98.61%* |

Table 4.10: Performance of the predictor-based fusion schemes on the four probe sets in the WVU database. The accuracy has been evaluated by 5-fold cross validation and the classification rates have been averaged.

| Probe | Highest Rank | Borda Count | Pure Majority Voting |
|---|---|---|---|
| P1 | 99.66% | 96.56% | 92.89% |
| P2 | 98.77% | 93.67% | 91.45% |
| P3 | 99.44% | 92.22% | 90.11% |
| P4 | 99.33% | 93.67% | 91.89% |
| *Avg* | *99.30%* | *94.03%* | *91.59%* |

Table 4.11: Performance of the traditional fusion schemes on the three probe sets in the Biosecure database

| Probe | Highest Rank | Borda Count | Pure Majority Voting |
|---|---|---|---|
| $P_1$ | 87.18% | 96.15% | 89.74% |
| $P_2$ | 78.85% | 88.46% | 83.97% |
| $P_3$ | 74.36% | 92.31% | 84.62% |
| *Avg* | *80.13%* | *92.31%* | *86.11%* |

Table 4.12: Performance of the predictor-based fusion schemes on the three probe sets in the Biosecure database

| Probe | Predictor-based Majority Voting | Predictor-based Borda Count | Predictor-based Sequential |
|---|---|---|---|
| $P_1$ | 90.38% | 90.38% | 95.51% |
| $P_2$ | 87.18% | 87.18% | 89.10% |
| $P_3$ | 91.67% | 91.67% | 92.31% |
| *Avg* | *89,74%* | *89,74%* | *92,32%* |

predictor for each modality was not used (ties were broken randomly). From these tables it is evident that the *weighted majority voting* which uses the predictor for each modality, outperformed the other traditional approaches. The training has been affected by a lack of examples belonging to the negative class.

## 4.3 Graph-based Framework for Personal Identification Fusion at Rank-Score Level

In this section, we investigate the question of whether it is possible to improve the performance of the identification system by using the non-matched scores. The idea is to incorporate the similarities of the query with that of its neighbors in order to have more information to be fused. Biometric identification techniques typically base the decision only on the match score representing the similarity between the identity query and the template of each gallery identity stored in the database. Traditional fusion methods derive the combined score by taking only the match scores related to a particular subject (the identity query). The proposed framework attempts to use additional information when computing the integrated score for each person. In particular, the combination functions at rank level

usually consider only the rank one output from each biometric matcher to compute the integrated rank of person *i-th*. This kind of fusion is called *local*. Conversely, the distribution of the *i-th* query identity in a *global method* is modeled by considering the subset of the enrolled persons similar to the query. The proposed approach belongs to this last category and it uses a subset of *non-matching* templates in the database, referred as *cohorts* [73]. When dealing with a large number of classes, as in the case with biometric person identification systems, they tend to overlap. For most biometrics to find a good model for representing a universal background class is an interesting challenge. The *complement* class for the query identity is given by those models as impostors, which have good resemblance with the model of the subject to which the system has assigned rank value one. In our problem, good impostors are represented by the identities in top of the candidate list, in fact they are expected to be more similar to the identity at rank one.

### 4.3.1 Cohort Analysis in Biometrics

The strategy of looking beyond the similarity of the query with only that of the claimed identity was already proposed by Bolle *et al.* in [16] in the biometric verification scenario. In their work, the matching scores of the other people are used into the decision making. In identification mode, the cohort information is associated to the neighbors in the candidate list of the genuine identity. This additional information can be exploited in a fusion scheme, when computing the integrated rank for each person.

### 4.3.2 Our approach

The proposed framework is based on the idea that the distances between the query and its neighbors may help to reduce the error rate. A crucial step of the proposed strategy consists in *Cohort* selection. For each enrolled template, we identify its *cohorts* based on a ranking criterion. The proposed combination approach attempts to extend the traditional methods by using the match scores corresponding to a subset of all people.

### 4.3.3 Graph Theory for Modeling

For most biometrics, to find a good model representing a universal background class is an interesting challenge. In order to improve the recognition accuracy, the match scores and the information about their relative ranking are treated as two different pieces of the evidence [49]. This means that, we consider the output of each unimodal system as a list of candidates with the confidence measure associated to each item. The top of such as list is model through a graph. It is composed by two levels: a root node representing the genuine identity and its neighbor nodes representing the impostors that are the most similar to the identity having rank one [27], (see Fig.4.15).

**Summary**

In this chapter, we presented a methodology in which both ranks and scores have been used to improve the identification accuracy of multimodal biometric systems. For each modality, ranks and scores have been used to design a pattern classifier (*predictor*) which is able to estimate the decision reliability of the corresponding modality matcher in order to detect identification errors. This information has been introduced in novel fusion schemes. The

| | | |
|---|---|---|
| **0.88365** | **1** | **31** |
| **0.82627** | 2 | 83 |
| **0.76004** | 3 | 45 |
| **0.75968** | 4 | 38 |
| **0.75356** | 5 | 138 |
| **0.73294** | 6 | 29 |
| **0.72623** | 7 | 79 |
| **0.72376** | 8 | 71 |
| **0.72068** | 9 | 127 |
| **0.71764** | 10 | 208 |

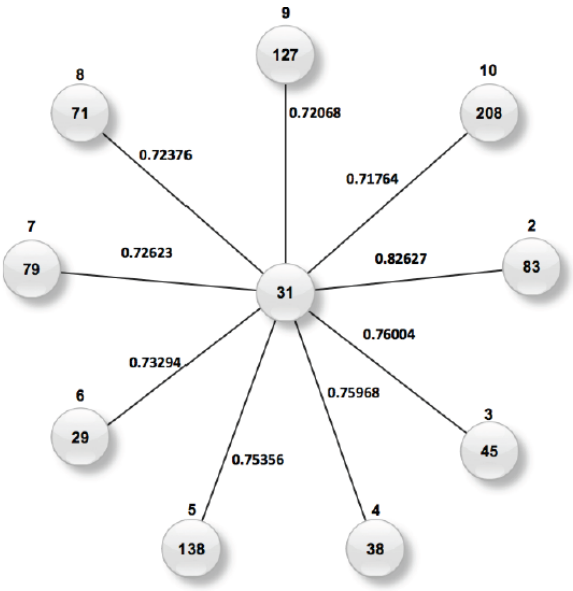


Figure 4.15: The two-levels graph represents the top 10 of the candidate list.

proposed predictor-based techniques performed better than the commonly used rank-level fusion mechanisms. In particular, the predictor-based majority voting resulted in the best accuracy by achieving an average recognition rate of 100% on the WVU dataset and 97.22% on the Biosecure dataset. The improvement in performance is especially significant in the BioSecure database. Since the predictor is based on a training phase, it generalizes very well across identities. Consequently, the predictor does not have to be retrained when a new individual is added to the database. Experiments are underway to determine the robustness of the scheme to variations in quality on the input data. It must be stated that the simple sum of scores results in good identification performance on the database used in our experiments; however, we can consider the methodology proposed in this chapter as a promising approach for using both ranks and scores in a systematic way to predict identification errors in biometric systems.

A genius is eternal patience.

*Michelangelo*

# Chapter 5

# Robustness to Spoof Attacks

A biological measurement can be qualified as a biometric and then used in a recognition process, only if it satisfy basic requisite like universality, permanence, distinctiveness, circumvention. The last property concerns the possibility of a non-client being falsely accepted, typically by spoofing the biometric trait [25]. Previous works have shown that it is possible to spoof a variety of fingerprint technologies through relatively simple techniques. They use molds of fingers made with materials as *Silicon*, *Play-Doh*, *Clay* and *Gelatin* (gummy finger). In 2002, Matsumoto *et al.* [70] conducted experimental *spoofing* research by creating gummy fingers to attack fingerprint verification systems. They have reported a vulnerability evaluation of 68%-100% for cooperative users and 67% for not-cooperative users (when data were extracted from latent fingerprints).

The main focus of this chapter concerns the security risk in multimodal biometric systems due to spoof attacks. We have analyzed the performance of the most efficient multibiometric systems in presence of spoofing and our experiments show that the probability of deceiving a multibiometric system is high even if only one modality is spoofed Then,

we proposed a novel liveness detection algorithm for the fingerprint modality, which combines static features based on the skin perspiration phenomenon and on the morphologic properties of the fingerprint [69]. The experiments were carried out by adopting standard databases taken from the Liveness Detection Competition 2009 (LivDet09) in which *Biometrika*, *CrossMatch* and *Identix* sensors were used [42]. Further, we presented a novel study focused on how the performance of the liveness detection algorithms changes when fake fingers are produced by employing materials that are different with respect to those adopted for training. Finally, we tested this our algorithm in a fusion scheme.

## 5.1 Analysis of the Robustness of Multimodal Biometric Systems against Spoof Attacks

From a security perspective, a multimodal system appears more protected than its unimodal components, since spoofing two or more modalities is harder than spoofing only one [63]. However, since a multimodal system involves different biometric traits, it offers a higher number of vulnerable points that may be attacked and a hacker may fake only a subset of them. There is indeed a *trade-off* between the number of fused biometric traits and the offered security level. Recently, researchers investigated if a multimodal system can be deceived by spoofing only a subset of the fused modalities [57]. Rodrigues *et al.* proposed a method which considers as measure of security also the information pertaining the ease to spoof each biometric in order to weight the contribution provided by the single modality to the multimodal system. The idea is that, if a high quality sample gives a low match score, the probability of success for a spoof attack is high. This work has been extended in [26],

by exploring the multimodal vulnerability and strategies for fusion in a scenario in which partial spoofing has occurred. In this section, we also looked at the cases where some but not all modalities are spoofed. The experiments were conducted by employing the scores sum rule on two multimodal databases composed by face and fingerprint.

### 5.1.1 Experimental Analysis

#### Datasets

The performance of the considered strategy was evaluated on two multimodal databases.

The first is NIST-BSSR1 (Biometric Scores Set - Release 1). It is a *true* multimodal database i.e., the face and the fingerprint images coming from the same person at the same time. Our experiments were carried out by employing the first partition made up of face and fingerprint scores belonging to a set of 517 people. For each individual, it is available a score coming from the comparison of two right index fingerprints, a score obtained by comparing impressions of two left index fingerprints, and two scores (from two different matchers, say $C$ and $G$) that are the outputs of the matching between two frontal faces. The match score for each modality indicates a *distance*. Our dataset consists in an unbalanced population composed by 517 genuine and 266,772 (517*516) impostor match scores.

The second database is a subset of the BioSecure multimodal database. This database contains 51 subjects in the Development Set (training) and 156 different subjects in the Evaluation Set (testing). For each subject, four biometric samples are available over two sessions: session 1 and session 2. The first sample of each subject in the first session was used to compose the gallery database while the second sample of the first session and the two samples of the second session were used as probes ($P_1$, $P_2$, $P_3$). For the purpose of

this study, we have employed one face and three fingerprint modalities, denoted as *fnf*, *fo1*, *fo2* and *fo3*, respectively [54]. The scores used in our experiments are the output of the matching between the first available sample and the second one for each subject. Our second dataset consists in an unbalanced population composed by 516 genuine and 24,180 (156*155) impostor match scores.

**Experimental Procedure**

According to the assumption that live-spoof match scores would be similarly distributed as live-live match scores, the simulation of an unimodal spoof attack has been realized by substituting a genuine match score in place of an impostor match score. Given the availability of four modalities, we have firstly analyzed a multi-biometric system which exploits four modalities without spoofing simulation, then the cases where one, two, three and all the modalities have been spoofed. Fusing the match scores from multiple sources, such as from face and different instances of fingerprints, the resulting system should achieve a higher recognition accuracy [26]. The current system has been designed by computing the FRR and FAR/SFAR at different threshold levels and plotting them in a DET curve on a log scale. As common practice, the operating point of the system corresponds to the point where the FRR value is very close to the FAR (ERR Equal Error Rate) on the curve representing the no spoofing simulation scenario. This sets a common threshold level at which the additional curves representing scenarios where spoof attacks have been simulated can be compared to that one where spoof attacks are absent. Fig. 5.1 and 5.2 show three groups of DET curves based on SFAR, for one modality spoofed, two modalities spoofed

and three modalities spoofed.

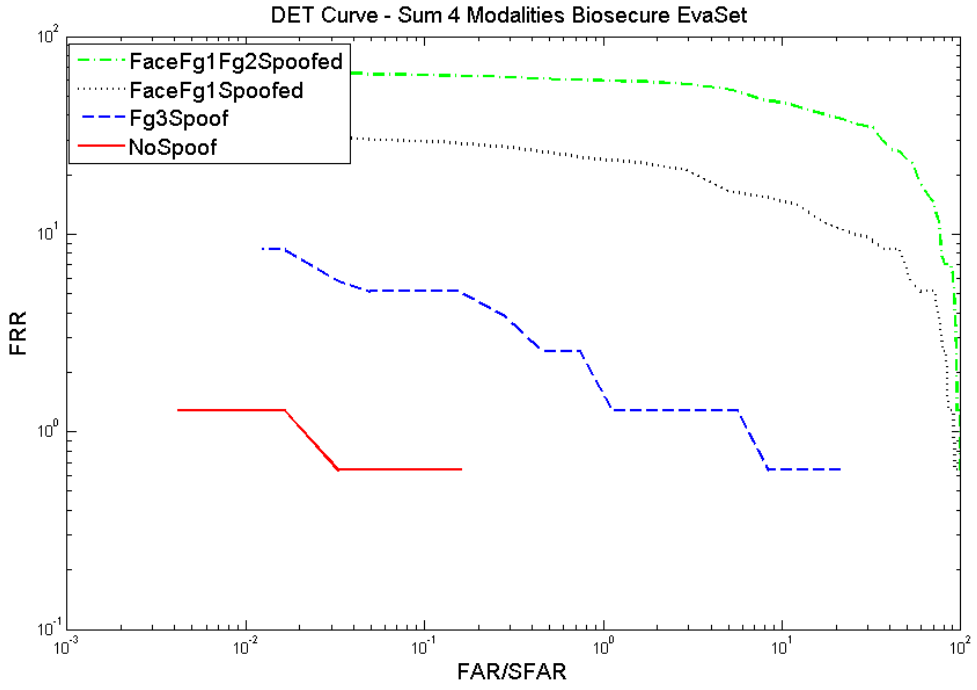


Figure 5.1: DET plot for a multi-modal system which exploits four modalities taken from Biosecure database. The dark black line indicates the performance of the traditional fusion scheme based on the sum rule with trade-off between FAR and FRR.

Further, a two modality system has been designed by using face and fingerprint scores. The related DET curves are shown in Fig. 5.3 and 5.4.

**Discussion**

The results of Figure 1, concerning the Biosecure data, show an ERR (FAR/FRR) of 0.64%. For this operating point, when one of three fingerprint modalities is spoofed, referred to as *fo1*, *fo2* and *fo3*, the average SFAR is respectively of 6.29%, 8.92% and 8.45%, with an associated FRR of 0.64%. When the face modality is spoofed, the SFAR jumps up to an average of 77.67%, since that modality presents the highest recognition accuracy in an

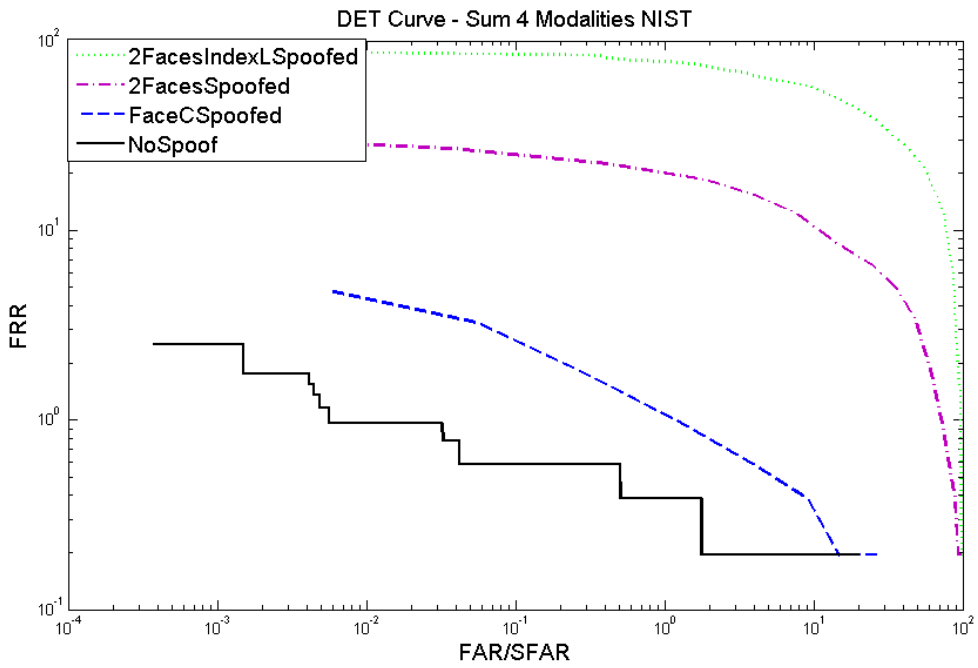


Figure 5.2: DET plot for a multi-modal system which exploits four modalities taken from Nist database.

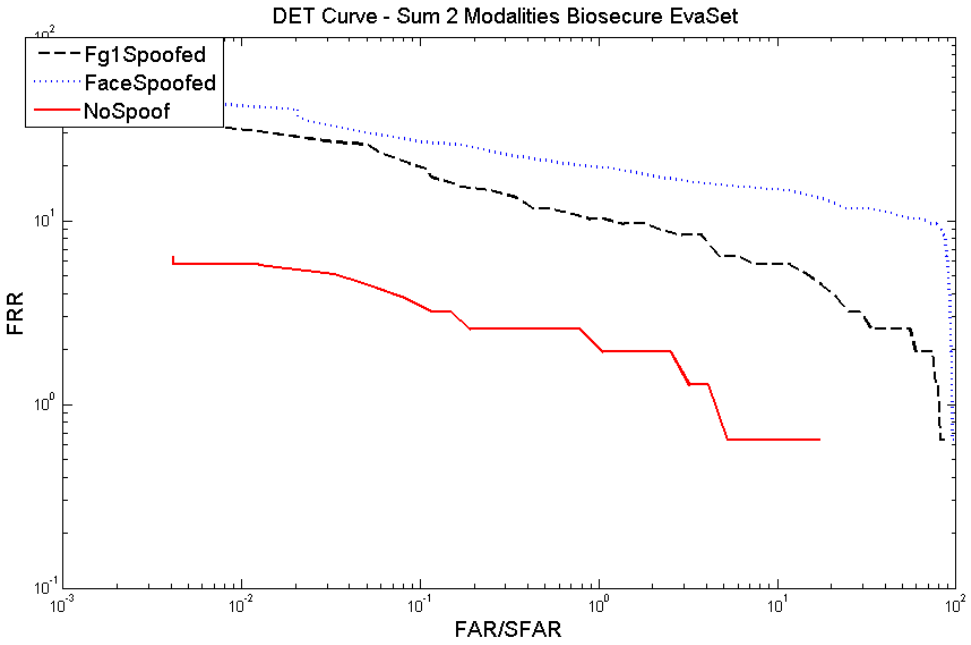


Figure 5.3: DET plot for a multi-modal system which exploits two modalities taken from Biosecure database. The dark black line indicates the performance of the traditional fusion scheme based on the sum rule with trade-off between FAR and FRR.

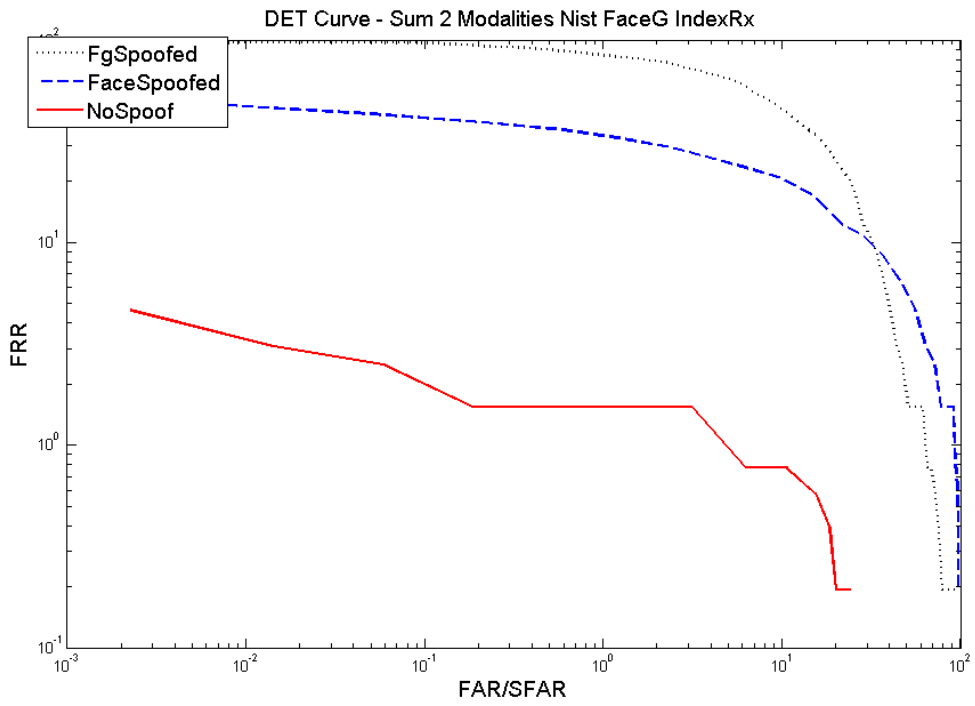


Figure 5.4: DET plot for a multi-modal system which exploits two modalities taken from Nist database.

unimodal scenario. When two of four modalities are spoofed, SFAR achieves 91.89%, while when three of four modalities are spoofed, SFAR jumps up to 98.66%.

The results of Figure 2, concerning the Nist data, show an ERR (FAR/FRR) of 0.58%. For this operating point, when one of four modalities is spoofed, the average SFAR is 4.04%, with an associated FRR of 0.58%. When two modalities are spoofed, SFAR jumps up to 81.40% and to 97.10% when three modalities are spoofed.

The two modality system presented analogous performance, as shown in Figures 3 and 4. On Biosecure database, with an FRR of 1.93%, when one modality is spoofed SFAR becomes 74.74% averaged over the two modalities. On Nist database, with an FRR of 1.54%, when one modality is spoofed SFAR becomes 63.26% averaged over the two modalities.

### 5.1.2 Likelihood Ratio Test

The Likelihood Ratio (LR) between the genuine and impostor distribution is known to be the optimal fusion method which minimizes the probability of error. We obtained a representative estimation of both distributions using training data taken from Biosecure and Nist databases using a Gaussian mixture model. The training process was carried out employing only *non-spoofed* impostor scores, while the testing scenario involved the case in which only a subset of the fused biometric modality was spoofed (see Fig.5.5 and Fig.5.6).

### 5.1.3 Identification Scenario

It was interesting to simulate a spoof attack to a biometric identification system, where in case of an identification error, the score at rank1 was substituted with the score corresponding to the true identity of the considered matching (see Fig.5.7).

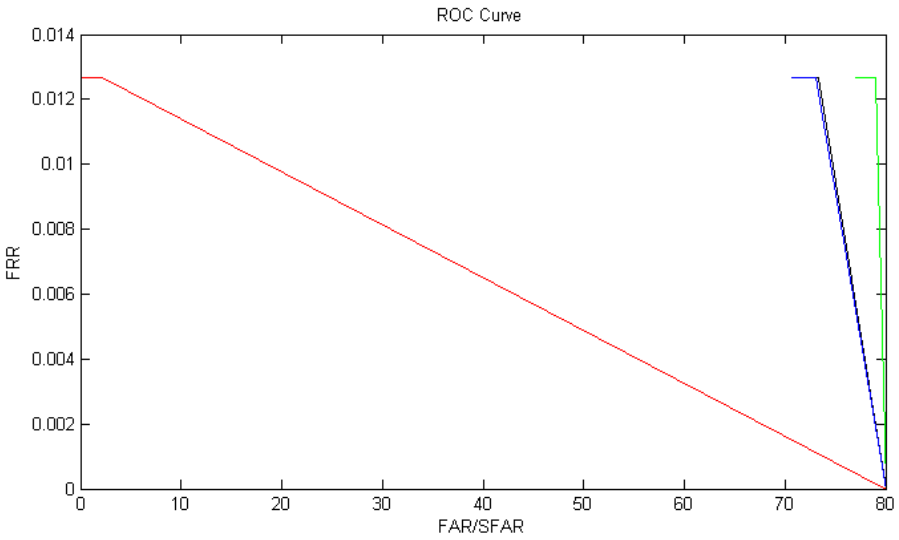


Figure 5.5: Performance of the Likelihood Ratio Test based on joint density distributions of two fingerprint modalities and two face modalities taken from the Biosecure database.

### 5.1.4 Discussion

In this section, we analyzed the security of the existing multibiometric systems a subset of the fused modalities is successfully spoofed. The experiments showed a significant vulnerability of the existing fusion scheme in presence of attacks where not all modalities are spoofed. Our idea is to detect spoof attack to the single component matcher before fusion. This concerns the incorporation of a spoofing detection algorithm in a fusion scheme in order to achieve an increase of the multimodal performance in the described real scenario. Thus, we explored the topic concerning the detection of *vitality* in fingerprint images.

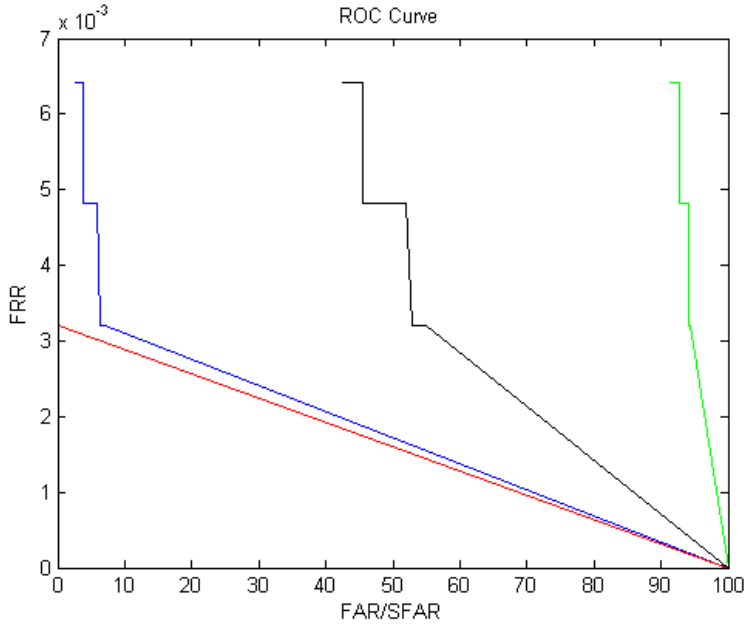


Figure 5.6: Performance of the Likelihood Ratio Test based on joint density distributions of two fingerprint modalities and two face modalities taken from the Nist database.

## 5.2 Combining Morphology- and Perspiration-based Features for Liveness Detection in Fingerprint Scanners

Fingerprint scanners are the most widely adopted for personal identification. However, the security of a fingerprint-based identification system is compromised in presence of fake biometric data. In fact, it is possible to deceive automatic fingerprint identification systems by presenting a well-duplicated synthetic finger. Artificial fingers created from fingerprints of enrolled users used to attempt to gain unauthorized access are called *spoofs*[51]. This kind of attack at sensor level can occur when people wish to disguise their own identity or when a person wants to gain privileges of an authorized person. To minimize sensor vulnerability, different approaches have been proposed. As an efficient mean to circumvent

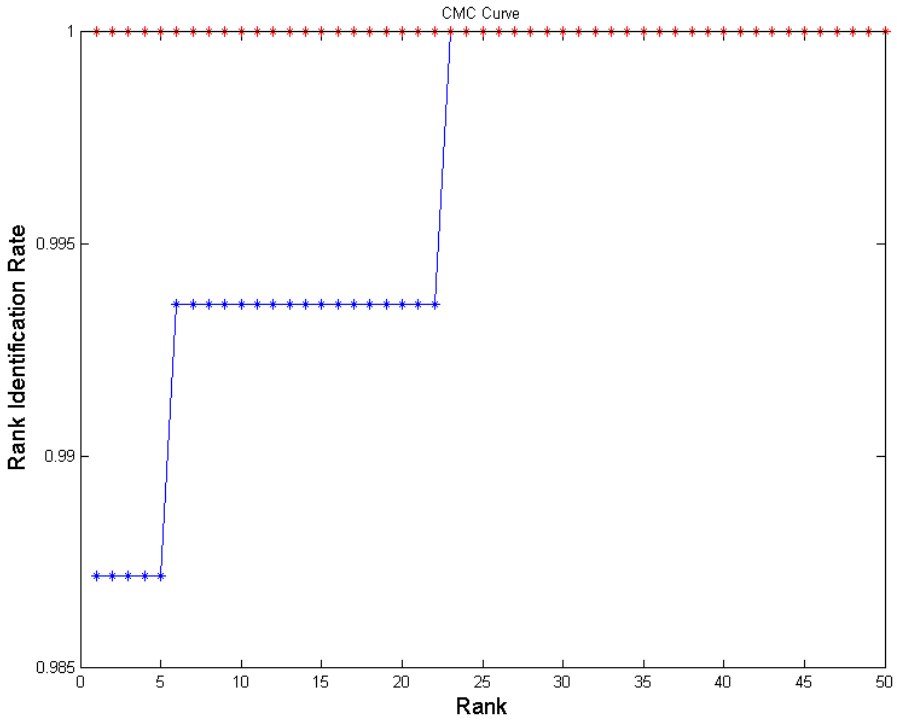


Figure 5.7: Performance of the score sum involving two fingerprint modalities and two face modalities taken from the Biosecure database, in identification operation.

attacks that use spoof fingers, *liveness detection* has been suggested. In the context of biometrics, *liveness detection* means the capability for the system to detect if the biometric sample presented is really from a live finger tip or not. Liveness methods may belong to two main categories, see Fig.5.8.

The first one exploits characteristics as the temperature of the finger, the electrical conductivity of the skin and the pulse oximetry. They can be detected by using additional hardware in conjunction with the biometric sensor. This makes costly the device. The second category performs an extra process of the biometric sample in order to detect the vitality information directly from the fingerprint images. In this chapter, we focus on this second

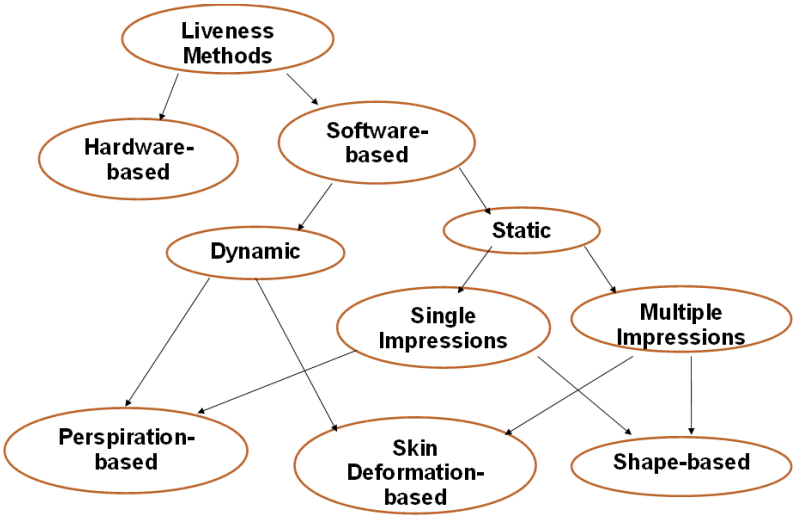


Figure 5.8: An example of live and fake(gummy) fingerprint image.

category of approaches, known as *software-based* [69]. The existing software-based solutions may include *dynamic* or *static* methods [7]. Static characteristics (as temperature, conductivity) and dynamic behaviors (skin deformation, perspiration) of live finger tips have been extensively studied in fingerprint liveness detection research. In particular, morphology- and perspiration-based characteristics have been typically exploited separately. Since both features provide discriminant information about live and fake fingers, it is reasonable to investigate also their joint contribution.

### 5.2.1 Dynamic approaches

Dynamic features derive from the analysis of multiple frames of the same finger. A typical *dynamic* property of a live finger is the perspiration phenomenon that starts from the pores and evolves in time across the ridges, see Figure 5.9. This distinctive spatial moisture pattern can be detected by observing multiple fingerprint images acquired in two appropriate different times. An interesting method based on perspiration changes in live fingers was

presented in [3]. In this method, the changing perspiration pattern is isolated through a wavelet analysis of the entire fingerprint image. For an image processing algorithm, to quantify the sweating pattern is challenging. Since this pattern is a physiological phenomenon, it is variable across subjects. Further, it presents a certain sensitivity to the environment, the pressure of the finger, the time interval and the initial moisture content of the skin [56]. Its effectiveness requires an efficient extraction of the evolving pattern from images.

Figure 5.9: The image shows a macro photography of a live fingerprint.

### 5.2.2 Static approaches

Static features can be extracted from a single fingerprint impression or as difference between different impressions. Generally, static measurements may be altered by factors as the pressure of the finger on the scanner surface. According to the taxonomy proposed in [52], features extracted by different impressions can be skin deformation-based or morphology-based, while features extracted by a single impression can be perspiration-based or morphology-based. Morphology-based features give a general description of the fingerprint pattern using its geometrical properties. Those based on the perspiration phenomenon quantify perspiration patterns along ridges in live subjects. Elastic deformations

due to the contact, the pressure and the rotation of the fingertip on the plane surface of the sensor, are more evident in fake fingerprints made using artificial materials than in live fingerprints. Deformation-based methods detect liveness by comparing these distortions through static features [78]. The elastic behavior of live and fake fingers has been analyzed by extracting a specific set of *minutiae points*, see Figure 5.10. The second type of static features using multiple impressions relies on a morphologic investigation which exploits the thickness of the ridges that is modified after producing the fingerprint replica.

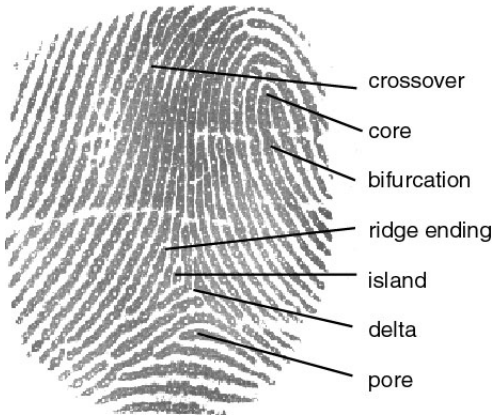


Figure 5.10: The image shows the discontinuities that interrupt the flow of ridges which are the basis for most fingerprint authentication methods. Minutiae are the points at which a ridge stops, and bifurcations are the points at which one ridge divides into two. Many types of minutiae exist, including dots (very small ridges), islands (ridges slightly longer than dots, occupying a middle space between two temporarily divergent ridges), ponds or lakes (empty spaces between two temporarily divergent ridges), spurs (a notch protruding from a ridge), bridges (small ridges joining two longer adjacent ridges), and crossovers (two ridges which cross each other).

Methods which exploit intrinsic properties of a single impression study the skin perspiration phenomenon. The vitality indication can be found by using Wavelet Transform and Fast Fourier Transform [9]. Wavelet analysis is able to capture the non-regular shape typical of the ridges in an image acquired from a live finger. Images taken from artificial

fingers show a more regular shape. Fourier Transform is employed to study the regular periodicity of pores on the ridges in live fingerprints. Such a regularity is not present in signals corresponding to spoof fingerprints. Liveness detection methods which search for morphological characteristics of fingerprint images, are significantly efficient when based on the surface coarseness.

Below, we describe three static *morphology-based* methods which exploit a single fingerprint image for vitality information extraction and which have been used for comparison. Each of them exploits a subset of the features we used in our algorithm.

Moon et *al.* [79] proposed a method based on analyzing the surface coarseness in high resolution (1000dpi) fingertip images. It has been observed that the surface of a fake finger is much coarser than the one of the human skin. The coarseness feature is measured by computing the standard deviation of the residual noise of the fingerprint image. The alternation of the ridges and valleys makes the fingertip surfaces intrinsically coarse because the material used during the fabrication process is composed by molecules which tend to agglomerate. Then, before feature extraction, the effect ridge/valley was minimized by using a wavelet decomposition at different scales. In particular, the image is enhanced through an histogram equalization and converted into a mono-dimensional signal representing the gray level profile of the ridges. The decision is made by using a threshold value of 25. This algorithm is fast and convenient but it works well only in presence of an high resolution sensor (1000dpi, while the common commercial sensors present a resolution of about 500dpi) [9].

An interesting texture-based approach using a single fingerprint image was proposed by

Nikam [50]. They analyzed liveness of a fingerprint image by using the gray level associated to the fingerprint pixels. The gray level distribution in a fingerprint image changes when the physical structure changes. This information is quantified by using several texture features. Real and fake fingerprint images present different textural properties useful for vitality detection. Due to the presence of sweat pores and the perspiration phenomenon, authentic fingerprints exhibit non-uniformity of gray levels along ridges, while due to the characteristics of artificial material surface, such as gelatin or silicon, spoof fingers show high uniformity of gray levels along ridges. The gray level distribution of the single pixels is modeled as first order statistics, while the joint gray level function between pair of pixels is modeled as second order statistics. The authors proposed Gabor filter-based features, since fingerprint exhibit oriented texture-like pattern and Gabor filters can optimally capture local frequency and orientation information. The basic steps of the adopted procedure are listed as follows:

- *Step1*: Fingerprint image is filtered using a bank of 4 Gabor filters oriented in 4 directions 0°, 45°, 90°and 135°.
- *Step2*: A gray level co-occurrence matrix method is applied to filtered images to extract textural details.
- *Step3*: Dimensionality of the features is reduced by Principal Component Analysis (PCA).

Features are used to train three different classifiers: a Neural Network (NN), a Support Vector Machine(SVM) and OneR. A Multilayer Perceptron (MLP) is used as NN and a

Radial basis function (RBF) is used as the SVM kernel, with parameters $C$ and $\gamma$ as 1 and 2.3, respectively. The three classifiers are then fused using the "Max Rule". This approach presents good performance when the *core point* is accurately located, (see Fig.5.10). However, existing core detection algorithms do not work well in presence of poor quality images or with very dry or wet fingerprints, resulting in a *noisy* core.

An approach based on multiresolution texture analysis and the inter-ridge frequency analysis of fingerprint images has been proposed by Abhyankar and Schuckers [2]. They used different texture features to quantify how the gray level distribution in a fingerprint image changes when the physical structure changes. First order statistics model the gray level distribution of the single pixels by using histograms, while second order statistics refer to the joint gray level function between pair of pixels. Two secondary features were used, Cluster Shade and Cluster Prominence, based on the co-occurrence matrix. These features derived from a multi-resolution texture analysis have been combined with features derived from fingerprint local-ridge frequency analysis that was performed as well. The training was performed separately for all the three scanners. Error rates have been computed after processing the statistics and the local ridges frequencies features by using Fuzzy-C-means classifier. This algorithm does not depend on the perspiration phenomenon and it is able to overcome the dependence on more than one fingerprint image. However, it presents limitations in real scenarios, since the computation of the local-ridge frequencies may be affected by cold weather and different skin conditions, including dirty fingers and wet fingers.

### 5.2.3 The proposed approach

Among the approaches proposed in the scientific literature, methods which exploit a single impression are cheaper and faster. None of the developed approaches alone can perfectly separate fake and live fingerprints. The static features previously described are able to capture different aspect of vitality,in particular morphology-based and perspiration-based . Then, it is reasonable that a combination of them is expected to achieve better performance than any of the individual measures. In the current investigation, we combine both perspiration- and morphology-based static features to improve the vitality detection accuracy.

Below we describe the considered morphology-based features.

- *Residual noise of the fingerprint image*: indicates the difference between an original and de-noised image, in which the noise components are due to the coarseness of the fake finger surface [2]. Materials used to make fake fingers such as *Silicon* or *Gelatin* consist of organic molecules which tend to agglomerate, thus the surface of a live finger is generally smoother than an artificial one [79]. In the present work, the coarseness of the image can be measured by computing the standard deviation of the residual noise of an image, where the amount of residual noise was computed by using a wavelet-based approach. According to the approach proposed by Moon [79], we have treated the surface coarseness as a kind of Gaussian white noise added to the image. Firstly, the image was de-noised with a *Symlet* by applying a *soft-threshold* for wavelet shrinkage. The noise residue was achieved by calculating the difference

between the two finger tip images before and after de-noising. The *Noise Residue Standard Deviation* is a good indicator of texture coarseness since the pixel value fluctuation in the noise residue of a coarser surface texture is generally stronger.

- *First order statistics*: measure the likelihood of observing a gray value at a randomly-chosen location in the image. The gray level associated to each pixel is exploited to determine a vitality degree of the fingerprint image. They can be computed from the histogram of pixel intensities in the image. The goal is to quantify the variations of the gray level distribution when the physical structure changes. The distinction between a fake and a live finger is based on the difference of these statistics. If $H(n)$ indicates the normalized histogram and $N$ the number of bin, the set of first order statistical properties used in this work are as follows [2]:
  - Energy:
  $$e = \sum_{n=0}^{N-1} H(n)^2 \tag{5.1}$$
  - Entropy:
  $$s = -\sum_{n=0}^{N-1} H(n) log H(n) \tag{5.2}$$
  - Median:
  $$M = \arg\min_{a} \sum_{n} H(n)|n-a| \tag{5.3}$$
  - Variance:
  $$\sigma^2 = \sum_{n=0}^{N} (n-\mu)^2 H(n) \tag{5.4}$$

– Skewness:

$$\gamma_1 = \frac{1}{\sigma^3} \sum_{n=0}^{N-1} (n-\mu)^3 H(n) \tag{5.5}$$

– Kurtosis:

$$\gamma_2 = \frac{1}{\sigma^4} \sum_{n=0}^{N-1} (n-\mu)^4 H(n) \tag{5.6}$$

– Coefficient of variation:

$$cv = \frac{\sigma}{\mu} \tag{5.7}$$

Below we describe the considered perspiration-based features.

- *Individual pore spacing.* Extensive research has shown that pore patterns are unique to each individual [2]. A photo-micrograph of pores is shown in Figure 5.11. For the purpose of the proposed approach, we focus on analyzing the occurrence of pores that causes a gray value variability in the fingerprint image. This tendency can be studied by using the Fast Fourier Transform (FFT), then the fingerprint image has to be transformed into a *ridge signal*, representing the gray-level value along the ridge. The discrimination between a live finger and a fake one is performed in the space of the total energy of the *ridge signal*. In this method, according to the algorithm proposed in [56], the 2-dimensional fingerprint image was mapped to 1-dimensional signal which represents the gray-level values along the ridges. This technique lets to quantify the perspiration phenomenon in a given image. The gray-level variations in the signal correspond to variations in moisture due to the pores and the presence of perspiration. By transforming the signal in the Fourier domain lets to measure this static variability in gray-level along the ridges. In particular, the focus is on

frequencies corresponding to the spacial frequencies of the pores. Firstly, by using a median filter the image was processed to remove noise and device effects. Such as denoised image was converted into a binary one. Second, a thinning routine was applied on the binary image and the fingerprint ridge paths, composed by only one pixel, were determined. Connections were removed to have only individual curves. Finally, the FFT was computed and the total energy associated to the spacial frequency of the pores was obtained as static feature. The coefficients of interest are from 11 to 33, since these values correspond to the spacial frequencies (0.4 - 1.2 mm) of pores. The formula for this static measure $SM$ is given from the following:

$$SM = \sum_{k=11}^{33} f(k)^2 \tag{5.8}$$

where $f(k)$ is expressed by the following:

$$f(k) = \frac{\sum_{i=1}^{n} |\sum_{p=1}^{256} S_{0i}^{a}(p) e^{-j2\pi(k-1)(p-1)/256}|}{n} \tag{5.9}$$

$$S_{0i}^{a} = S_{0i} - mean(S_{0i}) \tag{5.10}$$

where $n$ is the total number of individual ridges and $S_{0i}$ is the $i^{th}$ ridge.

- *Intensity-based.* From the intensity distribution perspective, among the 256 different possible intensities, the spoof and cadaver fingerprints images are distributed in the dark (<150) [71]. The current study uses image histograms showing the number of pixels at each different intensity values found in the image and it focuses on the gray level values along the ridge, represented by the *ridge signal*. We have computed two particular features: *i) Gray Level 1 ratio*, corresponding to the ratio between the

number of pixels having a gray level belonging to the range (150, 253) and the number of pixels having a gray level belonging to the range (1, 149); *ii) Gray Level 2 ratio*, corresponding to the ratio between the number of pixels having a gray level belonging to the range (246, 256) and the number of pixels having a gray level belonging to the range (1, 245). Moreover, we have analyzed the uniformity of gray levels along ridge lines and the contrast between valleys and ridges. As Figure 5.12 shows, real fingerprints exhibit non-uniformity of gray levels and high ridge/valley contrast values. Then, the general variation in gray-level values of in a spoof fingerprint is less than a live one. To capture this information we have computed as additional feature the *Gradient* of the gray-level matrix of the image.

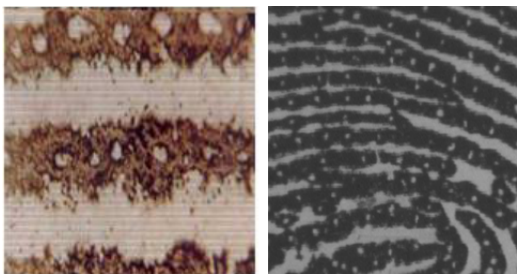

Figure 5.11: The image on the left shows a photo-graphical example of pores. The image on the right is output from a high resolution sensor (1000dpi) that captures the location of pores in detail. Both are taken from [20].

The time to perform the recognition process is a fundamental parameter which affects the performance of the proposed system. A feature selection phase reduces the number of features to be extracted and subsequently the time needed for feature extraction. We have selected the subset of features with highest discriminant power on the training set by using a *Sequential Forward Selection* technique. The feature selection was performed for each

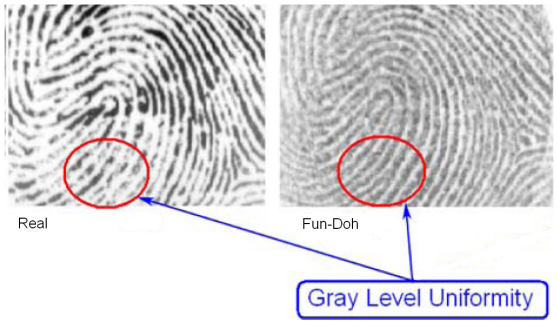


Figure 5.12: Gray level uniformity analysis in fingerprint images: high level value for a real fingerprint and low for a spoof. The image was taken from [50]

sensor.

Different classifiers have been trained, such as a Support Vector Machine, a Decision Tree, a Multilayer Perceptron and a Bayesian classifier. For each sensor, we have chosen the classifier with the highest accuracy on the training set.

### 5.2.4 Results and Discussion

#### Datasets

Our experimental phase was carried out by using three databases composed by live and spoof fingerprint images. Each database refers to a different sensor (*Biometrika*, *CrossMatch* e *Identix*), see Table 3. They have been taken from the Liveness Detection Competition 2009 and each one of them is composed by two subsets, one for training and the other one for testing the algorithm [42]. *Biometrika* training dataset is made up by 520 silicone images and 520 live images (13 subjects x 20 acquisitions x 2 frames), with 2 time-series (0 sec and 5 sec). The corresponding test set is made up by 1440 silicone images and 1440 live images (37 subjects x 20 acquisitions x 2 frames), with 2 time-series (0 sec and 5 sec). *CrossMatch* training dataset is made up by 500 live images and 500 fake images produced by using

Table 5.1: Datasets for training

| Database | Subjects | Live Images | Fake Images | Frames |
|---|---|---|---|---|
| *Biometrika* | 13 | 520 | 520 | 0 and 5 sec |
| *Identix* | 35 | 375 | 375 | 0 and 2 sec |
| *CrossMatch* | 63 | 500 | 500 | 0 and 2 sec |

Table 5.2: Datasets for testing

| Database | Subjects | Live Images | Fake Images | Frames |
|---|---|---|---|---|
| *Biometrika* | 37 | 1440 | 1440 | 0 and 5 sec |
| *Identix* | 125 | 1125 | 1125 | 0 and 2 sec |
| *CrossMatch* | 191 | 1500 | 1500 | 0 and 2 sec |

silicone, gelatin and *Play-Doh*, with 2 time-series (0 sec and 2 sec). The corresponding test set is made up by 1500 live images and 1500 fake images produced by using *Silicon*, *Gelatin* and *Play-Doh*, with 2 time-series (0 sec and 2 sec). *Identix* training dataset is made up by 375 live images and 375 spoof images produced by using *Silicon*, *Gelatin* and *Play-Doh*, with 2 time-series (0 sec and 2 sec). The corresponding test set is made up by 1125 live images and 1125 spoof images produced by using *Silicon*, *Gelatin* and *Play-Doh*, with 2 time-series (0 sec and 2 sec). The details about the data collection are shown in the tables 1 and 2. In the three cases, the subjects using for training are different than those considered for testing. Table 3 reports details about the sensors used for LivDet 2009 Competition.

Table 5.3: Fingerprint sensors used for LivDet 2009.

| Sensors | Model No. | Resolution (dpi) | Image size |
|---|---|---|---|
| *Biometrika* | FX2000 | 569 | (312x372) |
| *Identix* | DFR2100 | 686 | (720x720) |
| *CrossMatch* | Verifier 300 LC | 500 | (480x640) |

**Performance of the proposed method**

Firstly, we analyzed the fingerprint images in the space of the selected features. The Figures 5.13, 5.14, 5.15 and 5.16 correspond to the entropy, the mean, the variance and the coefficient of variation of the fingerprint image. These three first statistics present a good separability between the classes live and fake.

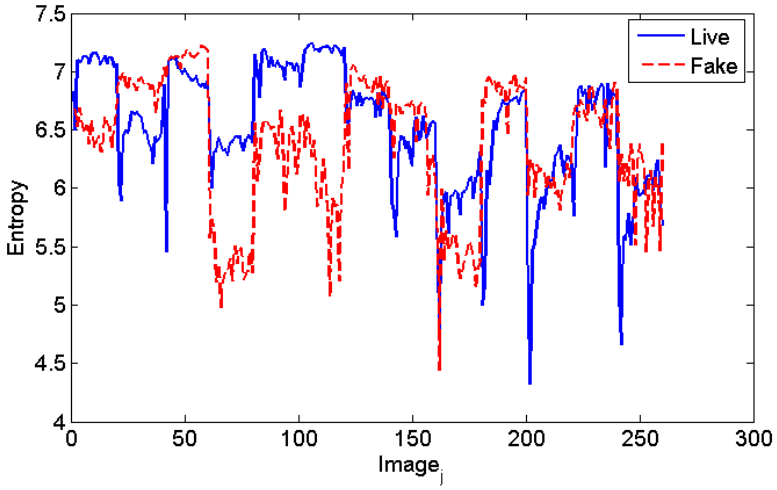


Figure 5.13: Entropy for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

The standard deviation of the residual noise also presents a good separability, as the Figure 5.17 shows.

Finally, Figure 5.18 and Figure 5.19 report the two intensity-based features, the Gray Level 2 and the gradient of the fingerprint image.

The classification performance evaluation was performed by adopting the same parameters used during the Liveness Detection Competition 2009, defined as follows:

- *Ferrlive*: rate of misclassified live fingerprints.

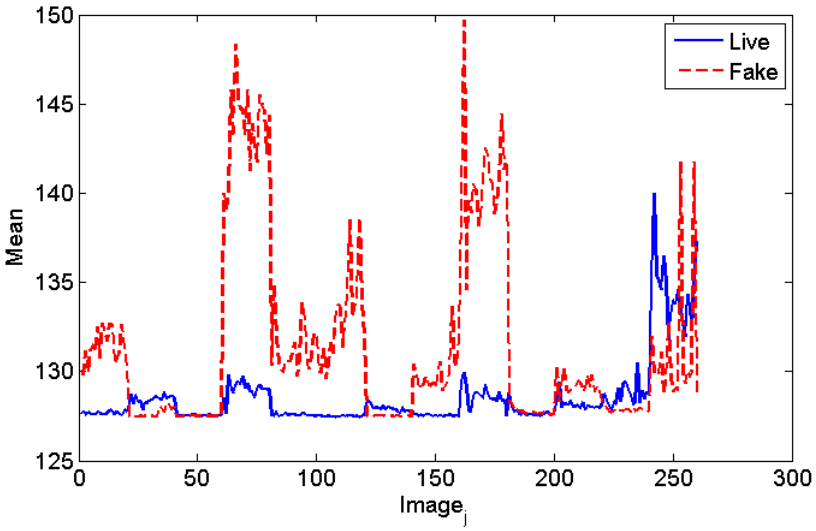


Figure 5.14: Mean for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

- *Ferrfake*: rate of misclassified fake fingerprints.

In particular, the indicator of performance is given from the value $e$ averaged on the three databases *Biometrika*, *CrossMatch* and *Identix*. The value $e$ is computed as follows:

$$e = \frac{Ferrlive + Ferrfake}{2} \tag{5.11}$$

Table 4 reports the average time needed for extracting each of the 12 features which has been exploited in our approach. Table 5 reports the features selected for each sensor by using a *Sequential Forward Selection* technique. We also observed that, when each feature was individually used, its discriminant power changed by varying the resolution of the images and the size of the dataset, while the joint usage of both perspiration- and morphology-based features showed a high discriminant power on all the considered databases. Moreover,

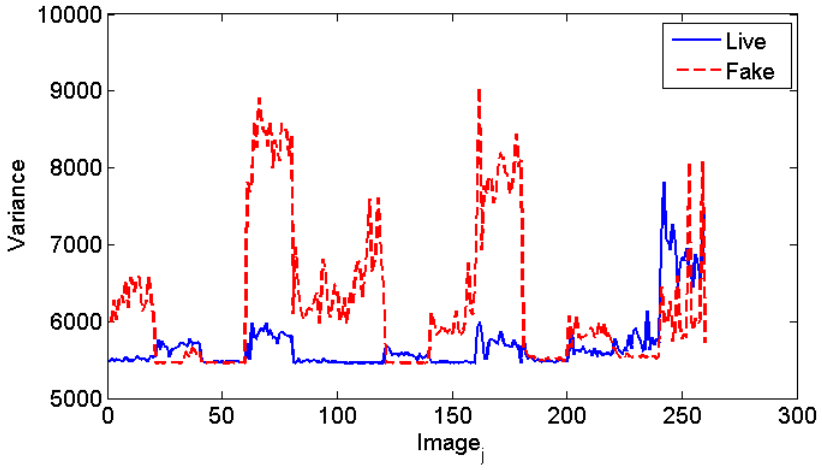


Figure 5.15: Variance for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

a subset composed by mean, gradient, standard deviation of the residual noise and the coefficient of variations has been selected. The average error rate achieved by the proposed method is of 12.47%, as reported in Table 6. On *Biometrika* and *Identix* datasets, the higher percentage accuracy has been achieved by using a Multilayer Perceptron, while on *CrossMatch* dataset, the Decision Tree classifier achieved the best performance. The performance achieved by the best algorithm submitted to the LivDet09 Competition was of 14.67%, as reported in Table 7. Our approach reduced this average error with a low variance on the three LivDet09 databases.

**Performance of the existing methods**

Table 8 reports the best performance of the method of Moon by varying the de-noising filter. The fingertip images have been first enhanced through a histogram equalization. Then the de-noising was performed by adopting different filters. Median filter produces standard

Table 5.4: Time required for extracting the proposed set of features on a *Core Duo T8100 2,1 Ghz Intel* Processor.

| Feature | Average Extraction Time |
|---|---|
| $Energy$ | 0.15 sec |
| $Entropy$ | 0.02 sec |
| $Mean$ | 0.02 sec |
| $Variance$ | 0.02 sec |
| $Skewness$ | 0.06 sec |
| $Kurtosis$ | 0.06 sec |
| $Coefficientofvariation$ | 0.02 sec |
| $NoiseResidueStd$ | 0.59 sec |
| $IndivPoreSpacing$ | 1 sec |
| $GrayLevel1$ | 0.02 sec |
| $GrayLevel2$ | 0.02 sec |
| $Gradient$ | 0.06 sec |

Table 5.5: Selected features for each database.

| Feature | | Biometrika | CrossMatch | Identix |
|---|---|---|---|---|
| $Morphology-based$ | $Energy$ | | x | x |
| $Morphology-based$ | $Entropy$ | x | | x |
| $Morphology-based$ | $Mean$ | x | x | x |
| $Morphology-based$ | $Variance$ | | x | x |
| $Morphology-based$ | $Skewness$ | | x | x |
| $Morphology-based$ | $Kurtosis$ | | x | x |
| $Morphology-based$ | $CoefficientOfVariation$ | x | x | x |
| $Morphology-based$ | $NoiseResidueStd$ | x | x | x |
| $Perspiration-based$ | $PoreSpacing$ | x | x | |
| $Perspiration-based$ | $GrayLevel1$ | | x | |
| $Perspiration-based$ | $GrayLevel2$ | x | | x |
| $Perspiration-based$ | $Gradient$ | x | x | x |

Table 5.6: Performance of the proposed algorithm.

| | $Ferrlive$ | $Errfake$ | $e$ |
|---|---|---|---|
| $Biometrika$ | 12.20% | 13.00% | 12.60% |
| $CrossMatch$ | 17.40% | 12.90% | 15.20% |
| $Identix$ | 8.30% | 11.0% | 9.70% |
| $Average$ | 12.60% | 12.30% | 12.47% |

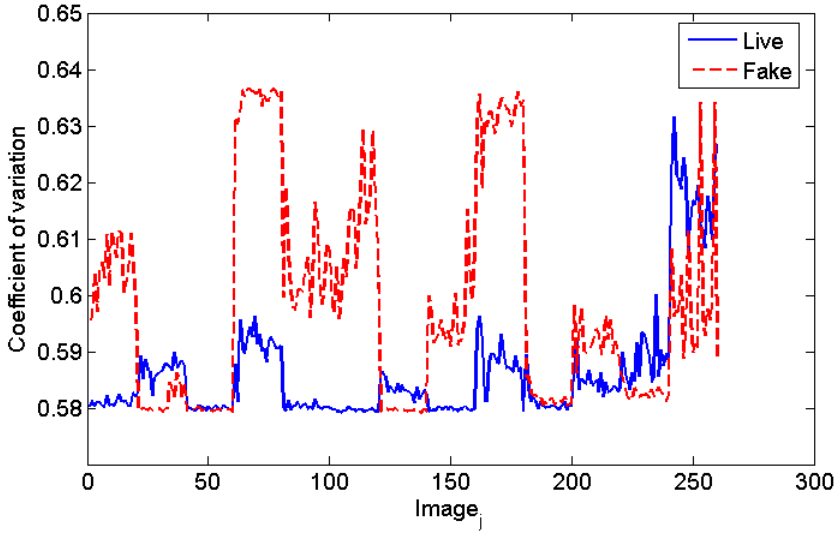


Figure 5.16: Coefficient of variation for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

deviation values similar to the value 25 employed in the approach proposed by Moon *et al.*, while the Wavelet-based procedure presents lower values of the considered feature. According to the procedure proposed in [79], the wavelet shrinkage was performed by applying a soft-threshold. The threshold assumes the lower value on the database *CrossMatch* having the lower resolution (500dpi) and composed by image with poor quality. In our experiments, we also used wavelet packets that work using high frequencies at each filtering of the image. They seem to be good for fingerprint images that present the most of the components at high frequencies. The time frequency analysis is performed by repeating the filtering of the signal. At each filter step, the frequency domain is cut in the middle and the high-frequency components are kept [75]. Wavelet packets are able to improve the classification accuracy only when the resolution is high enough, on *Identix* database it increases from 61.80% to 64.10% using a Symlet wavelet and from 62.00% to 66.80%, in both cases the threshold value

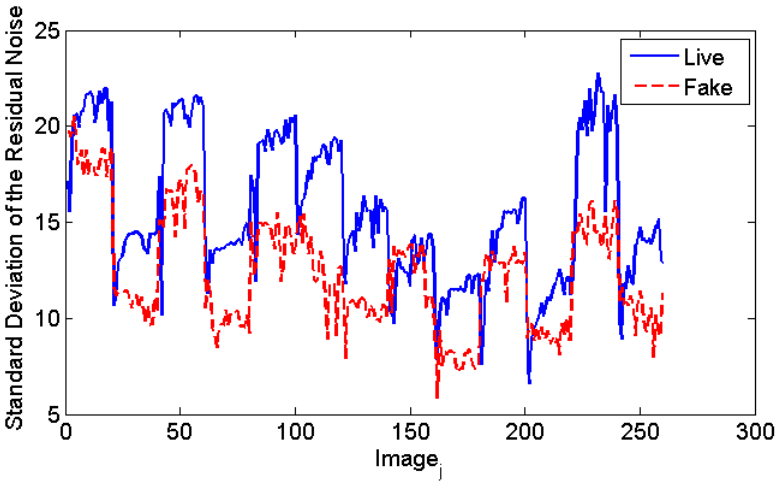


Figure 5.17: Standard deviation of the residual noise for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

Table 5.7: Performance of the best algorithm submitted to the Liveness Detection Competition 2009.

| | $Ferrlive$ | $Errfake$ | $e$ |
|---|---|---|---|
| $Biometrika$ | 15.60% | 20.70% | 18.20% |
| $CrossMatch$ | 14.40% | 15.90% | 15.20% |
| $Identix$ | 9.80% | 11.30% | 10.60% |
| $Average$ | 13.20% | 16.10% | 14.67% |

is close to 1. On fingerprint images taken *Identix*, Meyer wavelet packet worked better than the standard wavelet, while on the other two databases the usage of wavelet packets made a performances decrease.

Table 9 and 10 show the performance of the methods proposed by Nikam and Abhyankar-Schuckers respectively, on the three LivDet09 databases. The first approach achieved the lowest error rate, equal to 18.70%, on the *CrossMatch* having the lowest resolution, while the second approach achieved the highest error rate, equal to 47.20%, on the *Identix* database having the highest resolution [41].

Table 5.8: Performance of the method of Moon on the three databases LivDet09 using a Median filter for de-noising.

| | Threshold | Ferrlive | Ferrfake | e |
|---|---|---|---|---|
| *Biometrika* | 16.50 | 54.30% | 24.80% | 39.55% |
| *CrossMatch* | 25.00 | 6.00% | 70.00% | 38.00% |
| *Identix* | 16.50 | 31.20% | 44.50% | 37.85% |
| *Avg* | 19.33 | 30.50% | 46.43% | 38.47% |

Table 5.9: Performance of the method of Moon on the three databases LivDet09 using Symlet wavelet for de-noising.

| | $Threshold$ | $Ferrlive$ | $Ferrfake$ | $e$ |
|---|---|---|---|---|
| *Biometrika* | 20.60 | 20.80% | 25.00% | 23.00% |
| *CrossMatch* | $3.1^{-11}$ | 27.40% | 19.60% | 23.50% |
| *Identix* | 10.50 | 74.70% | 1.60% | 38.20% |
| *Avg* | | 40.97% | 15.40% | 28.23% |

Table 5.10: Performance of the method of Moon on the three databases LivDet09 using Symlet wavelet packet for de-noising.

| | Threshold | Ferrlive | Ferrfake | e |
|---|---|---|---|---|
| *Biometrika* | 1.15 | 46.50% | 8.80% | 27.70% |
| *CrossMatch* | 0.8 | 14.00% | 52.60% | 33.30% |
| *Identix* | 1.1 | 30.10% | 41.30% | 35.90% |

Table 5.11: Performance of the method of Moon on the three databases LivDet09 using Meyer wavelet for de-noising.

| | Threshold | Ferrlive | Ferrfake | e |
|---|---|---|---|---|
| *Biometrika* | 20.9 | 20.80% | 26.20% | 23.50% |
| *CrossMatch* | 0 | 45.00% | 28.00% | 36.50% |
| *Identix* | 10.8 | 74.40% | 1.30% | 38.00% |

Table 5.12: Performance of the method of Moon on the three databases LivDet09 using Meyer wavelet packet for de-noising.

| | Threshold | Ferrlive | Ferrfake | e |
|---|---|---|---|---|
| *Biometrika* | 1.2 | 38.10% | 15.00% | 26.50% |
| *CrossMatch* | 0.82 | 20.40% | 38.80% | 29.60% |
| *Identix* | 1.1 | 43.20% | 23.20% | 33.20% |

Table 5.13: Accuracy of the method of Nikam on the three databases LivDet09.

| | MLP | SVM | OneR | MaxRule |
|---|---|---|---|---|
| *Biometrika* | 76.1% | 73.6% | 70.7% | 76.46% |
| *CrossMatch* | 70% | 71.7% | 67.5% | 70.3% |
| *Identix* | 76.4% | 73.2% | 64.8% | 77.9% |

Table 5.14: Performance of the method of Nikam (Max Rule) on the three databases LivDet09.

| | $Ferrlive$ | $Ferrfake$ | $e$ |
|---|---|---|---|
| *Biometrika* | 14.30% | 42.30% | 28.30% |
| *CrossMatch* | 19.00% | 18.40% | 18.70% |
| *Identix* | 23.70% | 37.00% | 30.35% |
| *Avg* | 19.00% | 32.57% | 25.78% |

Table 5.15: Performance of the method of Abhyankar and Schuckers on the three databases LivDet09.

| | $Ferrlive$ | $Ferrfake$ | $e$ |
|---|---|---|---|
| *Biometrika* | 24.20% | 39.20% | 31.70% |
| *CrossMatch* | 39.75% | 23.30% | 31.53% |
| *Identix* | 48.40% | 46.00% | 47.20% |
| *Avg* | 37.45% | 36.17% | 36.81% |

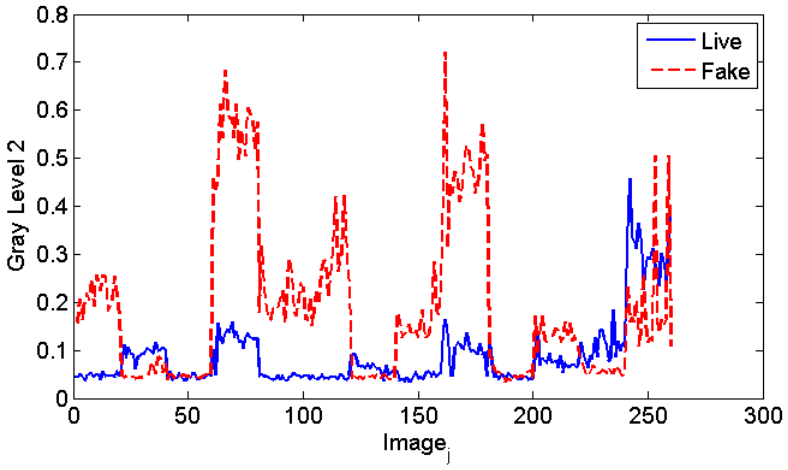


Figure 5.18: Gray Level 2 for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

## 5.3 Robustness of Liveness Detection Algorithms against New Materials used for Spoofing

In our previous experiments, the classifier was trained by using features extracted from fake samples made with all the materials available in each database. In particular, *Gelatin*, *Silicon* and *Play-Doh* are the materials employed in both *Identix* and *CrossMatch* databases. However, a good liveness detection algorithm is expected to be robust when the material used to learn the fake class changes. This aspect is a challenging problem in fingerprint liveness detection, since nowadays materials used for fraudulent spoof attacks are going to become very sophisticated. In this section, we analyze the performance of the existing liveness algorithms in scenarios reproducing the real conditions, where the material used to attack the system is not *a priori* known.

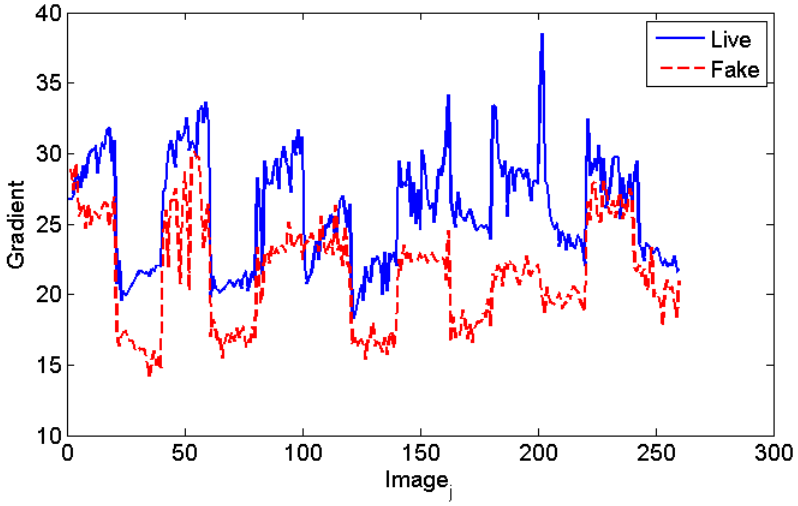


Figure 5.19: Gradient for live(blue line) and fake(red line) fingerprint images taken from Biometrika database.

### 5.3.1 Existing methods employed for comparison

We compare our method to the most efficient approaches existing in the literature, which are those proposed by Moon [79], Nikam [50], Shuckers and Abhyankar [2]. We also considered a *perspiration-based* method using the joint contribution of dynamic and static features which was experimented by Tan and Schuckers [71]. They studied the perspiration phenomenon from the intensity distribution perspective, by observing that live fingers present a distinctive contrast between white (>250, ASCII gray level range 0:255) and dark (<20) gray level, while spoof images have very small contrast difference. The decision rules to perform liveness classification is generated after considering static and dynamic features. The static features used in this work are based on the following parameters:

$$S1 = \frac{sum(151:254)}{sum(0:150)} \tag{5.12}$$

and

$$S2 = sum(151 : 254) \tag{5.13}$$

The dynamic features are based on the difference in the histogram distribution between zero and fifth second that is larger in live finger compared to spoof subjects. In the live fingers, perspiration makes dry (white) regions between the pores moister (darker) in time. This approach may present some limitations in cases of fingers too dry or too moist and other perspiration disorders.

### 5.3.2 Experimental Results

In order to study the robustness of the existing liveness detection approaches with respect to *unknown* materials used for producing fake fingers, we have carried out a further evaluation. In our experiments, each system was trained by using spoof fingerprints realized with all but one of the available materials, while the excluded material was used for testing. Table 4 reports the performance of the method proposed by our approach. In presence of high resolution images, taken from the *Identix* database, the testing performed using *Gelatin* and *Silicon*, when the training is performed by employing fake fingers made in Play-Doh, gives rise to a good spoofing recognition rate. Table 5 shows that the method proposed by Moon et *al.* wrongly classifies the majority of the fake fingerprints taken from *CrossMatch* database, while for a higher resolution factor, such a method presents a better behavior in presence of *unknown* materials using for spoofing. Table 6 and 7 show that the variation in fake materials does not significantly affect the performance of both Nikam-Agarwal and Abhyankar-Schuckers approaches, when the training set is only composed by samples made

Table 5.16: Performance of the method proposed by Marasco and Sansone on CrossMatch and Identix databases.

| | CrossMatch | | | Identix | | |
|---|---|---|---|---|---|---|
| | *Gelatin* | *Play − Doh* | *Silicon* | *Gelatin* | *Play − Doh* | *Silicon* |
| $Ferrlive$ | 6.5% | 5.7% | 12.6% | 3.8% | 19.2% | 9.7% |
| $Ferrfake$ | 25.9% | 16.7% | 10.0% | 42.3% | 5.5% | 30.6% |
| $e$ | 16.2% | 11.2% | 11.3% | 23.05% | 12.35% | 20.15% |

Table 5.17: Performance of the method proposed by Moon et *al.* on CrossMatch and Identix databases.

| | CrossMatch | | | Identix | | |
|---|---|---|---|---|---|---|
| | *Gelatin* | *Play − Doh* | *Silicon* | *Gelatin* | *Play − Doh* | *Silicon* |
| $Ferrlive$ | 12.30% | 15.00% | 35.70% | 45.20% | 79.60% | 40.80% |
| $Ferrfake$ | 63.10% | 61.80% | 47.30% | 31.80% | 4.20% | 36.80% |
| $e$ | 37.70% | 38.40% | 41.50% | 38.50% | 41.90% | 38.80% |

with *Gelatin*. On the contrary, as reported in Table 8, the performance of the Tan-Schuckers method seems quite dependent on the material as well as on the considered dataset.

As resumed in Table 9, when the material used to attack the system is not known during the training, most of the algorithms decrease in performance. This confirms our claim that the performance of liveness detection algorithms reported by the authors typically represents an overestimate of that obtainable in real scenarios. Among the considered methods, the one based on a single feature [79] is the most dependent on the use of *unknown* materials for testing. Also the dynamic method proposed in [71] had a significant decrement in performance when classifying fake fingerprints realized with materials different from those present in the training set. The other methods are instead more robust, and the one proposed by our approach, which is based on a combination of multiple features, exhibited the best average error $e$ when the material used for testing is *unknown* at training time.

Table 5.18: Performance of the method proposed by Nikam and Agarwal on CrossMatch and Identix databases.

| | CrossMatch | | | Identix | | |
|---|---|---|---|---|---|---|
| | *Gelatin* | *Play − Doh* | *Silicon* | *Gelatin* | *Play − Doh* | *Silicon* |
| $Ferrlive$ | 27.20% | 43.70% | 24.20% | 23.50% | 29.30% | 20.00% |
| $Ferrfake$ | 22.00% | 32.90% | 31.60% | 16.00% | 28.80% | 31.50% |
| $e$ | 24.60% | 38.30% | 27.90% | 19.75% | 29.05% | 25.75% |

Table 5.19: Performance of the method proposed by Abhyankar and Schuckers on CrossMatch and Identix databases.

| | CrossMatch | | | Identix | | |
|---|---|---|---|---|---|---|
| | *Gelatin* | *Play − Doh* | *Silicon* | *Gelatin* | *Play − Doh* | *Silicon* |
| $Ferrlive$ | 45.80% | 29.80% | 58.60% | 65.50% | 61.60% | 37.90% |
| $Ferrfake$ | 12.20% | 24.40% | 17.00% | 2.40% | 46.40% | 27.70% |
| $e$ | 29.00% | 27.10% | 37.80% | 33.45% | 54.00% | 32.80% |

Table 5.20: Performance of the method proposed by Tan and Schuckers on CrossMatch and Identix databases.

| | CrossMatch | | | Identix | | |
|---|---|---|---|---|---|---|
| | *Gelatin* | *Play − Doh* | *Silicon* | *Gelatin* | *Play − Doh* | *Silicon* |
| $Ferrlive$ | 38.60% | 24.40% | 54.80% | 64.10% | 36.00% | 38.80% |
| $Ferrfake$ | 32.20% | 39.20% | 43.00% | 28.70% | 42.40% | 13.20% |
| $e$ | 35.40% | 31.80% | 48.90% | 46.40% | 39.20% | 26.00% |

Table 5.21: Performance of the analyzed approaches in terms of the average error $e$ on Identix and CrossMatch databases.

| | Marasco-Sansone | Moon et *al.* | Nikam-Agar. | Abh.-Sch. | Tan-Sch. |
|---|---|---|---|---|---|
| *Gelatin* | 19.63% | 38.10% | 22.18% | 31.23% | 40.90% |
| *Play-Doh* | 11.78% | 40.15% | 33.68% | 40.55% | 35.50% |
| *Silicon* | 15.73% | 40.15% | 26.83% | 35.30% | 37.45% |
| *Avg* | **15.71%** | **39.47%** | **27.53%** | **35.79%** | **37.45%** |
| *All materials* | **12.45%** | **30.85%** | **24.53%** | **39.37%** | **29.20%** |

## 5.4 Evaluation of Fingerprint Liveness Detection Algorithms in a Fusion Scheme

### 5.4.1 Verification Scenario

In this section, we investigated whether incorporating a fingerprint liveness detection in a fusion scheme, under spoof attacks, may lead up to performance improvement. Our analysis involves the simple score sum and the statistical Likelihood Ratio test.

**Sum of scores**

The experiments were carried out on the Nist, Biosecure and WVU databases, described in the previous chapters 3 and 4. As said in Section 5.1, when considering a multimodal biometric system working in presence of a spoof attack, the worst case is obtained by the exact coincidence between the *fake-live* match score and the *live-live* match score. Then, it is important to evaluate the system performance under the assumption that, live-spoof match scores are similarly distributed as live-live match scores. Thus, we simulated each unimodal spoof attack by substituting a genuine match score in place of an impostor match score. We evaluated the performance of a multimodal system composed by face and fingerprint traits under normal operation (i.e., without spoofing), when only the fingerprint trait is spoofed and when only the face trait is spoofed. Further, we simulated the integration of our liveness detection algorithm in the fusion scheme, based on $Ferrlive$ and $Ferrfake$ percentages. The spoofed modalities, as assessed by the incorporated algorithm, do not have to give any contribution to the final decision. $Ferrfake$ indicates the percentage of spoofed scores (impostor substituted by genuine) that have to be reset, while (100% -

$Ferrlive$) indicates the percentage of genuine scores that have to be reset in the score sum rule. When one modality is spoofed, the presence of the liveness detection algorithm helps a $SFAR$ reduction, since $SFAR$ value of $10^{-3}$. This improvement is more significant when the algorithm is applied to both fingerprint modalities. We performed 10 iterations by randomly varying the set of fake samples detected by the algorithm. The average performance is reported in Fig.5.20. This plot shows that the $SFAR$ reduction can be achieved since very low $SFAR$ values $10^{-3}$. Finally, the procedure was experimented also for the sum among four modalities (see Fig.**??** and Fig.5.21). The EER point corresponds to 2.96% fixed on the curve without spoofing, for this FRR value, when only one fingerprint modality is spoofed, SFAR becomes equal to 87.75%. Incorporating our fingerprint liveness detection algorithm in the fusion scheme, aids to significantly decrease SFAR to a value of 8.20%. When two fingerprint modalities are spoofed, SFAR becomes equal to 96.91%. Here, incorporating our fingerprint liveness detection algorithm for both spoofed modalities in the fusion scheme, aids to achieve a SFAR value of 5.03%.

The same experiments were carried out on Biosecure database, where three fingerprint modalities and one face were available (see Fig.5.23 and Fig.5.25); The EER point corresponds to 0.19% fixed on the curve without spoofing, for this FRR value, when only one fingerprint modality is spoofed, SFAR becomes equal to 71.69%. Incorporating our fingerprint liveness detection algorithm in the fusion scheme, aids to significantly decrease SFAR to a value of 0.01%. When two fingerprint modalities are spoofed, SFAR becomes equal to 96.03%. Here, incorporating our fingerprint liveness detection algorithm for both spoofed modalities in the fusion scheme, aids to achieve a SFAR value of 0.20%. and on WVU

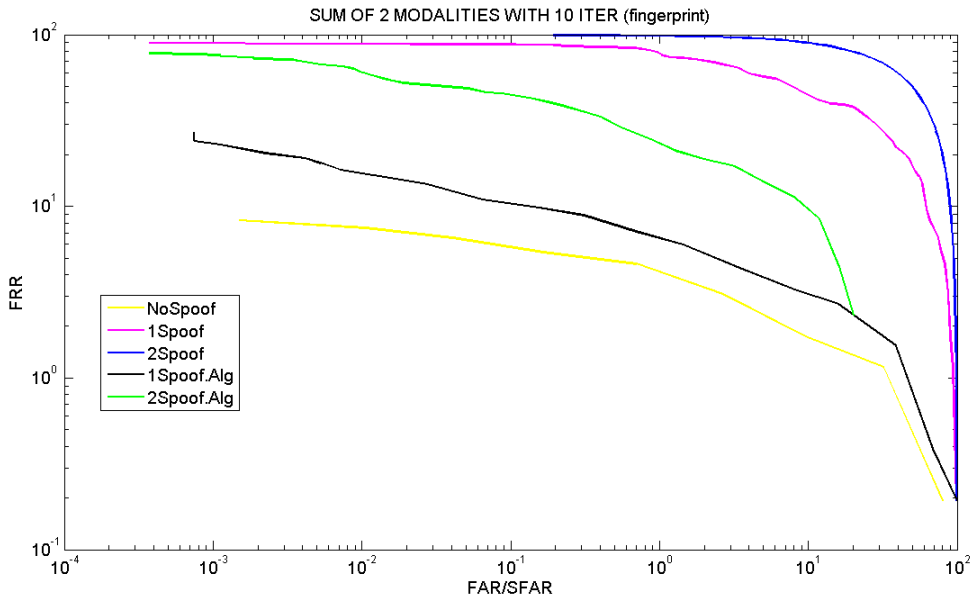


Figure 5.20: Average performance of the score sum between two fingerprint modalities taken from the Nist database over 10 iterations, where the fingerprint modalities have been spoofed.

database, where four fingerprint modalities and one face were available (see Fig.5.26 and Fig.5.27). The EER point corresponds to 0.19% fixed on the curve without spoofing, for this FRR value, when only one fingerprint modality is spoofed, SFAR becomes equal to 28.18%. Incorporating our fingerprint liveness detection algorithm in the fusion scheme, aids to significantly decrease SFAR to a value of 0.62%. When two fingerprint modalities are spoofed, SFAR becomes equal to 79.64%. Here, incorporating our fingerprint liveness detection algorithm for both spoofed modalities in the fusion scheme, aids to achieve a SFAR value of 0.004%.

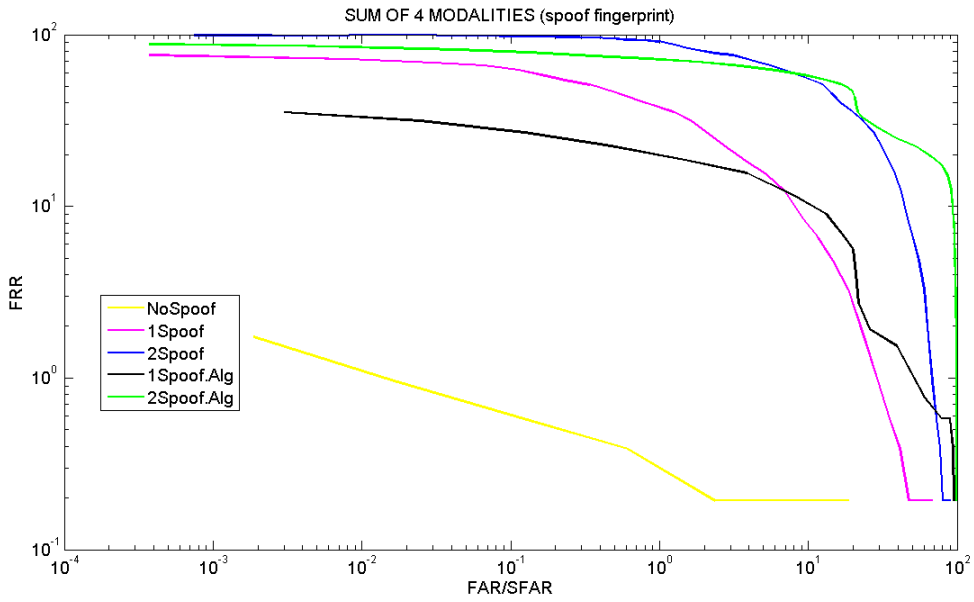


Figure 5.21: Average performance of the score sum between two fingerprint modalities and two face modalities taken from the Nist database over 10 iterations, where the fingerprint modalities have been spoofed.

### Summary

In this chapter, we have investigated a multimodal system composed of face and fingerprint modalities under different spoof attack scenarios. The experiments showed that, the multimodal systems present a high probability to be deceived by spoofing only one or a subset of its modalities. We have also proposed a novel fingerprint liveness detection algorithm which combines morphology- and perspiration- based features. The proposed algorithm has been tested on three different types of sensor technologies.

Our experiments demonstrated that, in presence of low resolution fingerprint images, it overcomes the limitations of the existing approaches. The overall system will also be faster, since the required information can be extracted from only one image without asking

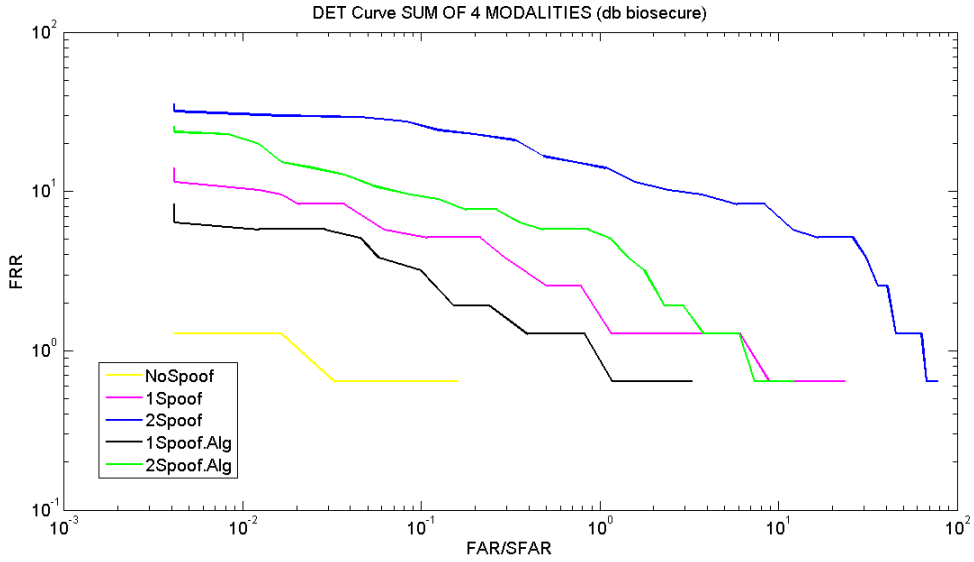


Figure 5.22: Performance of the score sum between three fingerprint modalities and one face modality taken from Biosecure database.

the user to scan twice his finger. Moreover, since our method does not require additional hardware, the cost of the fingerprint sensor does not increase.

Our experiments demonstrated also that, the performance of liveness detection approaches in which only one feature is exploited, decreases in presence of new materials employed for spoofing. This weakness is reduced when multiple vitality features are extracted. In particular, the combination of morphology- and perspiration-based features showed a high robustness in such a real scenario.

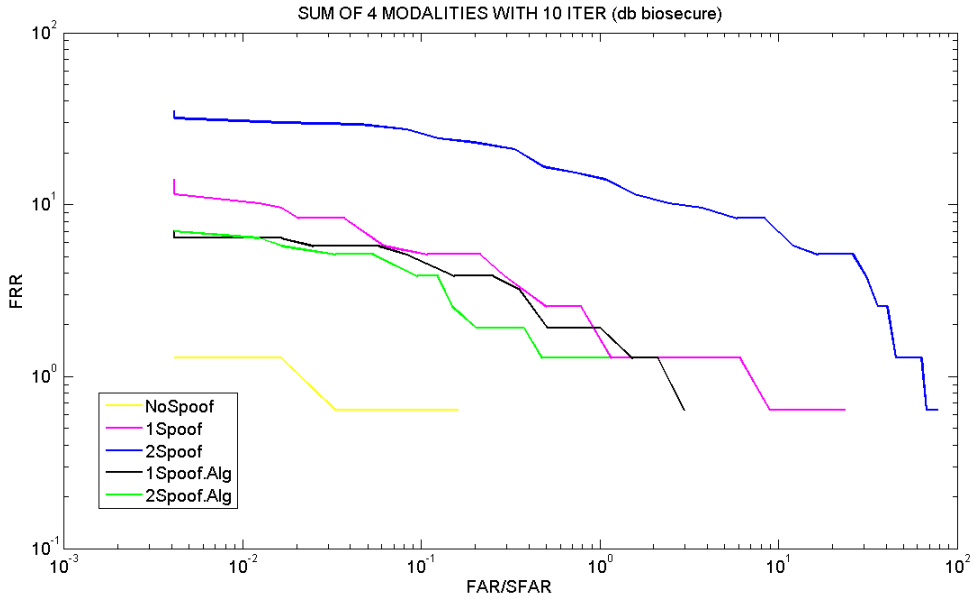


Figure 5.23: Performance of the score sum between three fingerprint modalities and one face modality taken from Biosecure database, over 10 iterations.

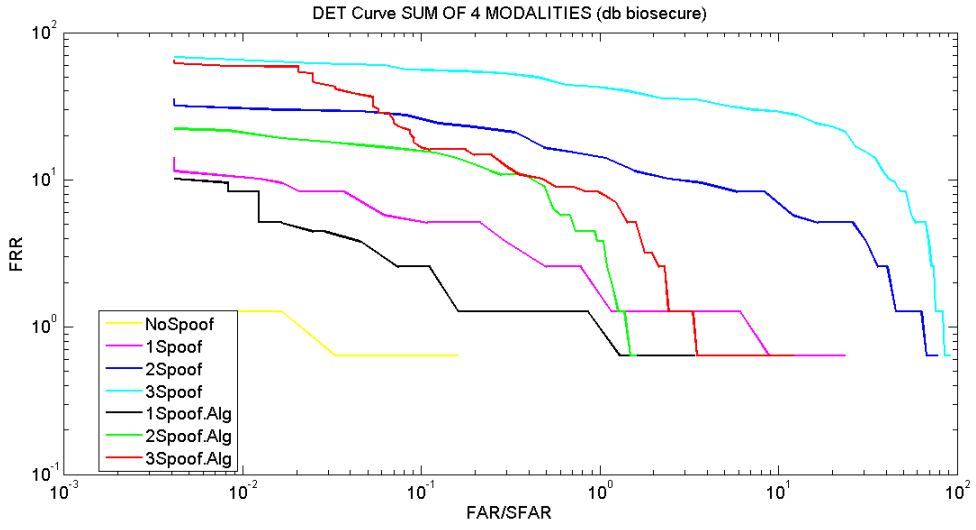


Figure 5.24: Performance of the score sum between three fingerprint modalities and one face modality taken from Biosecure database, where the three fingerprint modalities have been spoofed.

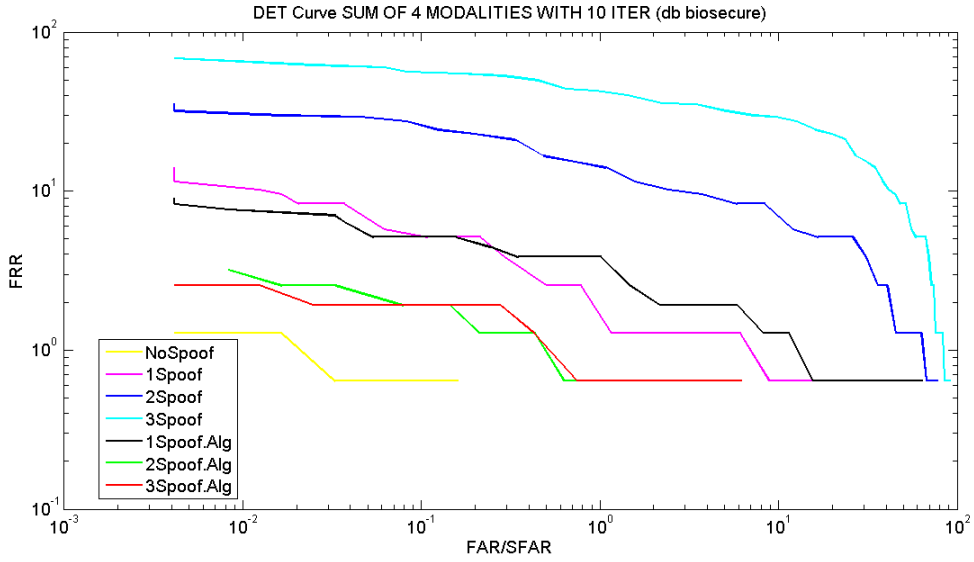


Figure 5.25: Performance of the score sum between three fingerprint modalities and one face modality taken from Biosecure database, where the three fingerprint modalities have been spoofed, over 10 iterations.

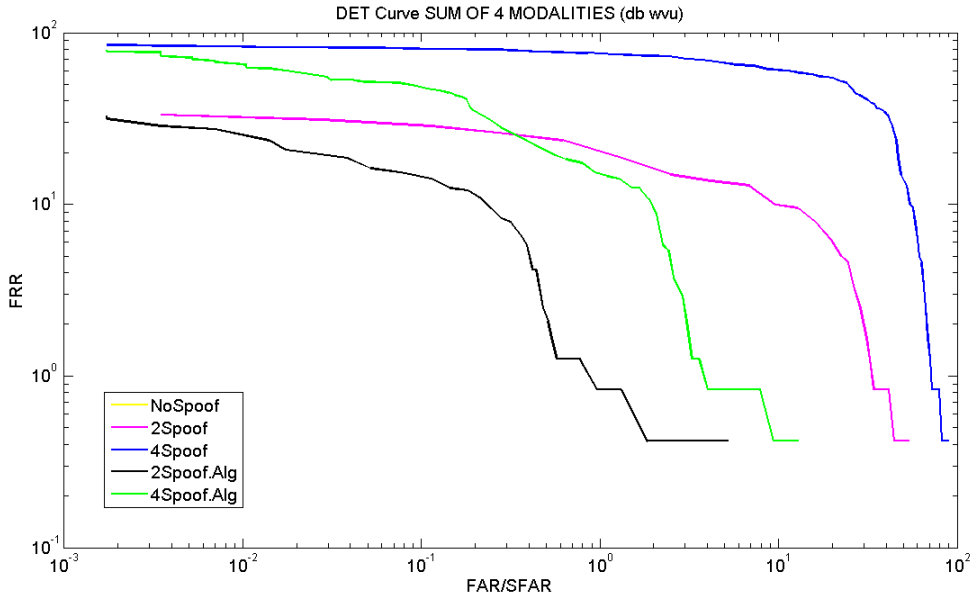


Figure 5.26: DET curve of the score sum involving one face and four fingerprint modalities taken from WVU database.

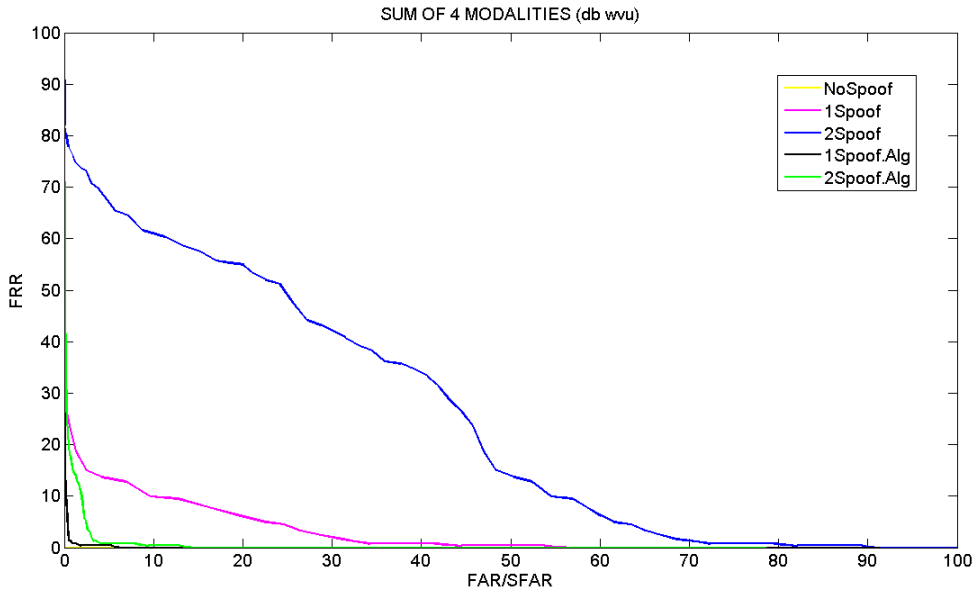


Figure 5.27: ROC curve of the score sum involving one face and four fingerprint modalities taken from WVU database.

# Bibliography

[1] A. Abaza and A. Ross. Quality-based rank level fusion in biometrics. *Third IEEE International Conference on Biometrics: Theory, Applications and Systems*, September 2009.

[2] A. Abhyankar and S. Schuckers. Fingerprint liveness detection using local ridge frequencies and multiresolution texture analysis techniques. *IEEE International Conference on Image Processing*, pages 321–324, October 2006.

[3] A. Abhyankar and S. Schuckers. Integrating a wavelet based perspiration liveness check with fingerprint recognition. *Pattern Recognition*, 42:452–464, 2009.

[4] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. The relation between the roc curve and the cmc. *Fourth IEEE Workshop on Automatic Identification Advanced Technologies*, pages 15–20, October 2005.

[6] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 17(10), October 1995.

[7] H. Kim C. Jin and S. Elliott. Liveness detection of fingerprint based on band-selective fourier spectrum. *Information Security and Cryptology*, 4817:168–179, 2007.

[8] T. Chow. On optimum recognition error and rejection trade-off. *IEEE Transaction on Information Theory*, 16:41–46, 1970.

[9] P. Coli, G. Marcialis, and F. Roli. Vitality detection from fingerprint images: a critical survey. *Lecture Notes in Computer Science*, 4642:722–731, 2007.

[10] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Neural network classification reliability: Problems and applications. *Image Processing and Pattern Recognition*, 5:161–200, 1998.

[11] S. Crihalmeanu, A. Ross, S. Schuckers, and L. Hornak. A protocol for multibiometric data acquisition, storage and dissemination. *Technical Report, West Virginia University*, 2007.

[12] A. Jain D. Maltoni, D. Maio and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer, 2003.

[13] S. Dass, K. Nandakumar, and A. Jain. A principled approach to score level fusion in multimodal biometric systems. *Fifth AVBPA*, pages 1049–1058, July 2005.

[14] R. Duda, P. Hart, and D. Stork. *Pattern Classification 2nd Edition*. Wiley-Interscience, 2001.

[15] M. Figuereido and A. Jain. Unsupervised learning of finite mixture models. *IEEE Transaction on Patterns Analysis and Machine Intelligence*, 24(3):381–396, March 2002.

[16] N. Ratha G. Aggarwal and R. Bolle. Biometric verification: Looking beyond raw similarity scores. *IEEE Transaction on Circuits and Systems for Video*, 14(1):4–20, January 2004.

[17] F. Gargiulo, E. Marasco, C. Mazzariello, and C. Sansone. Pattern recognition in adversarial environments. *5° Convegno Gruppo Italiano Ricercatori in Pattern Recognition (GIRPR)*, pages 1–12, 2010.

[18] S. Graves. On the neyman-pearson theory of testing. *The British Journal for the Philosophy of Science*, 29(1):1–23, March 1978.

[19] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Transaction On Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[20] K. Choi H. Choi, R. Kang and J. Kim. Aliveness detection of fingerprint using multiple static features. *World Academy of Science, Engineering and Technology*, 28:157–162, 2007.

[21] A. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, December 2005.

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

[23] A. Jain and A. Ross. Multibiometric systems. *Communications of the ACM*, 47(1):34–40, January 2004.

[24] A. Jain and A. Ross. Multibiometric systems. *Comm. ACM*, 47(1):34–40, January 2004.

[25] A. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transaction on Circuits and Systems for Video*, 14(1):4–20, January 2004.

[26] P. A. Johnson, B. Tan, and S. Schuckers. Multimodal fusion vulnerability to non-zero effort (spoof) imposters. *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2010.

[27] A. Kandel and H. Bunke. *Applied Graph Theory in Computer Vision and Pattern Recognition*. Springer, 2007.

[28] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[29] J. Kittler, Y. P. Li, J. Matas, and M. U. R. Sanchez. Combining evidence in multimodal personal identity recognition systems. *International Conference on Audio- and Video-based Biometric Person Authentication*, 1997.

[30] J. Kittler and N. Poh. Multibiometrics for identity authentication: Issues, benefits and challenges. *IEEE Transaction On Pattern Analysis and Machine Intelligence*, 2009.

[31] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo. Quality dependent fusion of intramodal and multimodal biometric experts. *SPIE Biometric Technology for Human Identification IV*, 6539, 2007.

[32] K. Kryszczuk. *École Polytechnique Fédéral de Lausanne*. Classification with class-independent quality information for biometric verification, 2007.

[33] K. Kryszczuk, J. Richiardi, and A. Drygajlo. Reliability estimation for multimodal error prediction and fusion. In *7th International Workshop on Pattern Recognition in Information Systems (PRIS)*, 2007.

[34] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Erros handling in multimodal biometric systems using reliability measures. In *Proc. 13th European Signal Processing Conference (EUSIPCO)*, 2005.

[35] L.I. Kuncheva. *Combining Pattern Classifiers Method and Algorithms*. Wiley, 2004.

[36] E. Lehmann and J. Romano. *Testing of Statistical Hypotheses*. Springer, 2005.

[37] Y. Ma, B. Cukic, and H. Singh. A classification approach to multi-biometric score fusion. In T. Kanade, A.K. Jain, and N.K. Ratha, editors, *AVBPA*, volume 3546 of *Lecture Notes in Computer Science*, pages 484–493. Springer, 2005.

[38] V. M. Mane and D. V. Jadhav. Review of multimodal biometrics: Applications, challenges and research areas. *International Journal of Biometrics and Bioinformatics (IJBB)*, 3.

[39] E. Marasco, A. Ross, and C. Sansone. Predicting errors in a multibiometric system based on ranks and scores. *IEEE Third International Conference on Biometrics (BTAS)*, 2010.

[40] E. Marasco and C. Sansone. Improving the accuracy of a score fusion approach based on likelihood ratio in multimodal biometric systems. *Lecture Notes in Computer Science*, 5716:509–518, 2009.

[41] E. Marasco and C. Sansone. An anti-spoofing technique using multiple textural features in fingerprint scanners. *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BioMs)*, pages 8–14, 2010.

[42] G. Marcialis, A. Lewicke, B. Tan, P. Coli, D. Grimberg, A. Congiu, A. Tidu, F. Roli, and S. Schuckers. First international fingerprint liveness detection competition - livdet 2009. *Lecture Notes in Computer Science*, 5716:12–23, August 2009.

[43] G.L. Marcialis and F.Roli. Serial fusion of fingerprint and face matchers. *Lecture Note in Computer Science*, 4472:151–160, June 2007.

[44] O. Melnik, Y. Vardi, and C. Zhang. Mixed group ranks: Preference and confidence in classifier combination. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 26(8):973–981, August 2004.

[45] A. Mishra. Multimodal biometrics it is: Need for future systems. *International Journal of Computer Applications*, 3(4):28–33, 2010.

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

[47] K. Nandakumar, Y. Chen, A. Jain, and S. Dass. Quality-based score level fusion in multimodal biometric systems. *Pattern Recognition*, 4:473–476, September 2006.

[48] K. Nandakumar, A. Jain, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[49] K. Nandakumar, A. Jain, and A. Ross. Fusion in multibiometric systems: What about the missing data? *International Conference on Biometrics*, June 2009.

[50] S. B. Nikam and S. Agarwal. Curvelet-based fingerprint anti-spoofing. *Signal, Image and Video Processing*, 4(1):75–87, January 2009.

[51] K. Nixon, V. Aimale, and R. Rowe. Spoof detection schemes. 2007.

[52] G. Marcialis P. Coli and F. Roli. Fingerprint silicon replicas: static and dynamic features for vitality detection using an optical capture device. *International Journal of Image and Graphics (IJIG)*, 8(4):495–512, 2008.

[53] N. Poh. *École Polytechnique Fédéral de Lausanne.* Multi-system biometric authentication : optimal fusion and user-specific information, 2006.

[54] N. Poh, T. Bourlai, and J. Kittler. A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms. *Pattern Recognition*, 43:1094–1105, 2010.

[55] N. Poh and J. Kittler. Multimodal information fusion. *Multimodal Signal Processing: Theory and applications for human-computer interaction*, 2009.

[56] L. Hornak R. Derakhshani, S. Schuckers and L. O'Gorman. Determination of vitality from non-invasive biomedical measurement for use in fingerprint scanners. *Pattern Recognition*, 36:383–396, 2003.

[57] R. N. Rodrigues, N. Kamat, and V. Govindaraju. Evaluation of biometric spoofing in a multimodal system. *IEEE International Conference on Biometrics (BTAS)*, 2010.

[58] F. Roli, J. Kittler, G. Fumera, and D. Muntoni. An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. *Proc. Multiple Classifier Systems, Sringer-Verlag*, 2364:325–336, 2002.

[59] A. Ross. *Integration of Multiple Cues in Biometric Systems*. Michigan State University, 2005.

[60] A. Ross. An introduction to multibiometrics. *Proc. of the 15th European Signal Processing Conference (EUSIPCO)*, 2007.

[61] A. Ross and A. Jain. Information fusion in biometrics. *Pattern Recognition Letters 24*, pages 2115–2125, 2003.

[62] A. Ross and A. Jain. Fusion techniques in multibiometric systems. *Face Recognition for Personal Identification*, pages 185–212, 2007.

[63] A. Ross, K. Nandakumar, and A. Jain. *Handbook of MultiBiometrics*. Springer, 2006.

[64] U. Sanchez and J. Kittler. Fusion of talking face biometric modalities for personal identity verification. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 5:V–V, 2006.

[65] F. Sanderson and K. Paliwal. Information fusion and person verification using speech and face information. *IDIAP RR*, pages 02–33, 2002.

[66] A. Saranli and M. Demirekler. A statistical unified framework for rank-based multiple classifier decision combination. *Pattern Recognition*, 34:865–884, 2001.

[67] W. Scheirer and T. Boult. A fusion-based approach to enhancing multi-modal biometric recognition system failure prediction and overall performance. *IEEE Int. Conf. on Biometrics Theory Application and Systems*, 2008.

[68] W. Scheirer and T. Boult. Predicting biometric facial recognition failure with similarity surfaces and support vector machines. *IEEE Computer Society Workshop on Biometrics*, 2008.

[69] S. Schuckers, R. Derakhshani, S. Parthasaradhi, and L. Hornak. Liveness detection in biometric devices. 2006.

[70] K. Yamada T. Matsumoto, H. Matsumoto and S. Hoshino. Impact of artificial gummy fingers on fingerprint systems. *Optical Security and Counterfait Deterrence Techniques IV*, 4677:275–289, January 2002.

[71] B. Tan and S. Schuckers. Liveness detection using an intensity based approach in fingerprint scanner. In *Proceedings of Biometrics Symposium (BSYM)*, September 2005.

[72] R. Tronci, G. Giacinto, and F. Roli. Combination of experts by classifiers in similarity score spaces. In N. da Vitoria Lobo, T. Kasparis, F. Roli, J.T.-Y. Kwok, M. Georgiopoulos, G.C. Anagnostopoulos, and M. Loog, editors, *SSPR/SPR*, volume 5342 of *Lecture Notes in Computer Science*, pages 821–830. Springer, 2008.

[73] S. Tulyakov and V. Govindaraju. Use of identification trial statistics for the combination of biometric matchers. *IEEE Transactions on Information Forensics and Security*, 3(4):1556–6013, December 2008.

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

[75] B. Walczak, B. Bogaert, and D. Massart. Application of wavelet packet transform in pattern recognition of near-ir data. *Analytical Chemistry*, 68:1742–1747, 1996.

[76] A. Wald. Sequential tests of statistical hypotheses. *The annals of Mathematical Statistics*, 16(2):117–186, June 1945.

[77] Y. Wang, T. Tan, and A. Jain. Combining face and iris biometrics for identity verification. *Proc. of the 4th International Conference AVBPA*, pages 805–813, 2003.

[78] A. Jain Y. Chen and S. Dass. Fingerprint deformation for spoof detection. *Biometric Symposium*, 2005.

[79] K. C. Chan K. So. Y. S. Moon, J. S. Chen and K. So. Woo. Wavelet based fingerprint liveness detection. *Electronic Letters*, 41(20):1112–1113, 2005.