The signal processing module represents the core of the system and is generally composed by sub-modules which implement the preprocessing functions (i.e. image filtering and enhancement), the feature extraction and the matching between two features.

Typically a biometric system can be characterized by the following attributes: uniqueness, universality, permanence, measurability, user friendliness, acceptability and circumvention [10]. *Uniqueness* refers to the fact that a feature must be unique: an identical feature should not appear in two different people. *Universality* means that the feature type is present/occurs in as many people as possible. Unfortunately we can not assume, for example, that every individual has all the fingers or has both irises not damaged. The *Permanence* property is related to the need that the feature does not change over time, or at least, it varies very slowly. *Measurability* concerns the possibility to measure the feature with relatively simple technical instruments. *User friendliness* requires that the measure should be easy and comfortable to be done, and *Acceptability* refers to the people's acceptance of the measure in daily lives. *Circumvention* concerns the toughness to deceive the system by fraudulent methods. All these attributes must be taken into account designing a biometric system.

Most cited biometric samples in the literature are: fingerprint, signature (hand-writing), facial geometry, iris, retina, hand geometry, vein structure, ear form, voice, DNA, odor (human scent), keyboard strokes and gait [2]. Each of them has different accuracy, cost and a different fulfillment of the seven attributes previously presented.

A biometric system can work basically in two configurations: identification and verification. *Identification* means that the acquired and processed biometric feature is compared to *all* biometric templates stored in a system. If there is a match, the identification is successful, and the corresponding user name or user ID is put in output. *Verification* means that the user enters her/his identity into the system (i.e. by keyboard or using a card) and a biometric feature is scanned. Then, the system compares the input feature *only* with the previously enrolled reference feature corresponding to the ID. If a match occurs, verification is successful. Systems that use a single biometric feature are defined as *monomodal*. When the identification is computed by comparing the matching values between N biometric features different in type with a specific policy, the system is called *multimodal* [13]. Example of combinations such as face/fingerprint, iris/fingerprint, and face/voice are particularly discussed in the literature [13-15]. Many studies report an improvement in accuracy for multimodal systems with respect to systems working with single biometric features [14-16].

III. BIOMETRIC SYSTEM EVALUTATION

The evaluation of a biometric system can be performed from different perspectives named: *technology*, *scenario* and *operational*. In this paper we deal with the technology evaluation since its goal is to compare competing algorithms when a sensor technology has been selected [7,10].

The *scenario evaluation* aims to determine the overall performance of a complete system in a prototype or simulated application that models a real-world target application. Since each tested system has its own acquisition sensor, it will receive slightly different data even if we acquire samples from the same individuals. Test results will be repeatable only if the simulated scenario can be carefully controlled. The *operational evaluation* tests a complete biometric system in a specific application environment with a specific target population. In general, operational test results will not be repeatable. The *technology evaluation* compares algorithms on a standardized database collected by a "universal" sensor. Of course, performance with this database will depend upon both the environment and the population in which it has been collected. Typically to avoid malicious approaches by the developers, it is possible firstly to provide them only a portion of the sample database, and distribute actual evaluation samples only after the developing of the algorithm's code. Testing is carried out using offline processing of the data. Because the database is fixed, the results of technology tests are repeatable.

Figure 2 shows the most general situation in a biometric database: we have a different number of samples for different individuals. Databases for algorithms comparison are poorly available [1, 17-20] due to the fact that they are very expensive and they contain complete biometric samples of real individuals. Security and privacy expects are seriously involved [11,12]. Some synthetic databases/generators are available only for fingerprints [21].

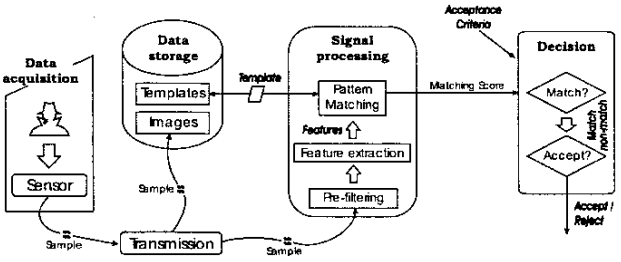


Figure 1: Structure of a biometric system.

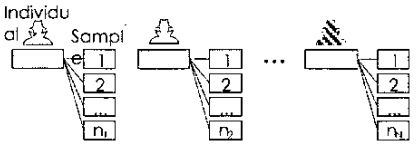


Figure 2: General samples situation of a biometric dataset

IV. ACCURACY AND PERFORMANCE INDEXES

In the case of a *technology evaluation*, the accuracy indexes most commonly accepted in the literature are now discussed.

The following definition of accuracy presents differences with respect to the classical one used in metrology [22] but it is generally accepted in biometric systems. Accuracy of measurements evaluates the agreement between the result of a measurement and the expected value, applying the system on a standardized database, as described in the previous section.

In this paper, accuracy is given by indexes evaluated using the concept of error: this definition is typically used in biometric systems. Readers often confuse this measure of accuracy processed on a standard database with the accuracy of the methodology. However, at least a second source of uncertainty – which affects the overall accuracy – should be considered: the uncertainty introduced by the measurement process due, for example, to pressure, humidity, finger position, electronic noise, quantization, etc. [5]. The authors consider this second source of uncertainty of great interest and it will be the goal of the further research. Moreover, taking into account both methodological and measurement uncertainty is not a trivial task. If the extracted biometric feature comes from an ideal sensor obtained by an ideal collection procedure, the methodological uncertainty should be equal to zero. However, in presence of noise corrupted samples, the preferred method minimizes the effect of noise source on the accuracy.

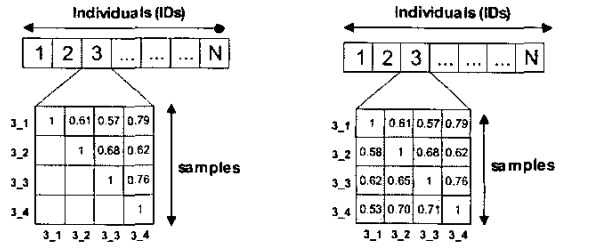
The following theory is valid for both monomodal and multimodal biometric systems. We can assume to have a sample database of identified individuals, as plotted in figure 2. In the literature many methods considered to evaluate the accuracy of a biometric system implicitly assume that the matching function is *symmetric* [15, 23 and 24]. Given two biometric features A and B and naming the matching function M, we have a symmetric matching function if M(A,B) = M(B,A). In the following we describe how to extend the equation for the accuracy evaluation for systems where we have M(A,B) ≠ M(B,A). Such systems are present in the literature, for example as described in [25] and [26]. In this paper, we do not comment if the symmetry is preferable to asymmetry in the matching function, but we will describe how is possible to make a fair comparison between different biometric systems by taking into account that issue.

Referring again to figure 2, let's define B_{ij} as the j^{th} sample of the i^{th} individual (i.e. a fingerprint or iris image, either filtered or not); T_{ij} as the template computed from B_{ij} (the features extracted); n_i as the number of samples available for the i^{th} individual and, finally, N as the number of individuals enrolled. Let's follow the steps to compute the accuracy performance of the systems defining the proper indexes.

A. Step 1 – Enrolment:

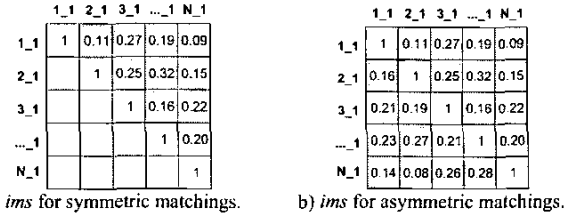
The templates T_{ij}, where i=1..N, j=1..n_i, are computed from the corresponding sample B_{ij} and stored on disk; if something wrong happens, an index (**REJ$_{ENROLL}$**) has to be increased.

REJ$_{ENROLL}$ is the rejection ratio in the enrolment phase, due to *Fail* (the algorithm declares it cannot enrol the biometric data), *Timeout* (the enrolment exceeds the maximum allowed time) and *Crash* (the algorithm crashes during biometric processing) situations [10,17].



a) *gms* for symmetric matchings b) *gms* for asymmetric matchings

Figure 3: Genuine Matching Scores.



a) *ims* for symmetric matchings. b) *ims* for asymmetric matchings.

Figure 4: Impostor Matching Scores

B. Step 2 – A general matching score computation:

For *symmetric matching functions* the consuetude is as follows [17]: each biometric template T_{ij} successfully created in the previous step is matched against the biometric sample B_{ik} ($j<k\leq n_i$). The matching values are stored in a matrix called *Genuine Matching Scores* **gms$_{ijk}$** (figure 3.a). The term "genuine" refers to the fact that the matching is computed between samples of the same certified individual. Since the matrix is symmetric by definition, *only* the upper triangular matrix is computed. Each individual has its own squared gms matrix.

We now propose how to include systems that have asymmetric matching-function into the framework proposed in the literature. Next section considers statistic effects on the estimation of the systems accuracy.

For *asymmetric matchings* each biometric template T_{ij} successfully created in the previous step is matched against the biometric sample B_{ik} ($1\leq k\leq n_i$, $k \neq j$) and the corresponding *Genuine Matching Scores* matrix **gms$_{ijk}$** is stored (figure 3.b). The matrix is not symmetric but it is still square. Then, the number of matches, denoted as **NGRA** (*Number of Genuine Recognition Attempts*) is given by

$$NGRA_{symMatch} = \frac{1}{2}\sum_{i=1}^{N} n_i(n_i - 1) \quad (1)$$

where REJ$_{ENROLL}$ = 0 (symmetric matching)

$$NGRA_{asymMatch} = \sum_{i=1}^{N} n_i(n_i - 1) \quad (2)$$

where REJ$_{ENROLL}$ = 0 (asymmetric matching).

Let's now consider the matching values of samples of *different* individuals (*impostors matching*). For symmetric

matching, each biometric template T_{i1}, $i=1..N$ is matched against the first biometric sample from different individual B_{k1} ($i<k\leq N$) and then the corresponding *Impostor Matching Scores* $\mathbf{ims_{ik}}$ matrix is stored (Figure 4.a). Impostor matching in the case of asymmetric matching function is computed as follows: each biometric template T_{i1}, $i=1..N$ is matched against the first biometric sample from different individual B_{k1} ($1\leq k\leq N$, $k \neq i$) and the corresponding *Impostor Matching Scores* $\mathbf{ims_{ik}}$ matrix is stored (Figure 4.b). The number of matches, denoted as **NIRA** (*Number of Impostor Recognition Attempts*) is given by

$$NIRA_{symMatch} = \frac{1}{2}N(N-1) \tag{3}$$

if $REJ_{ENROLL} = 0$ (symmetric matching), and

$$NIRA_{asymMatch} = N(N-1) \tag{4}$$

if $REJ_{ENROLL} = 0$ (asymmetric matching). Higher scores of matching values are associated with more closely matching images.

Finally, in the determination of *gms* and *ims* matrixes it is possible to have Fail, Timeout or Crash rejections. These events are respectively accumulated into REJ_{NGRA} and REJ_{NIRA} counters. It leads that *gms* and *ims* matrixes can have missing values. Commonly, in this case, special values are stored, i.e. "NULL" or negative matching values.

C. Step 3 – Accuracy Indexes

In this section we describe how to evaluate the confidence of the accuracy indexes, as defined in the literature, for a biometric system. Considering systems allowing multiple attempts or having multiple templates, a general definition defines errors of the matching algorithms considering *single* comparisons of a submitted sample against a *single* enrolled template. The rates are: False Match Rate **FMR(*t*)** and False Non-Match rate **FNMR(*t*)**. They are functions of the threshold value t used to compare the matching value to make the decision.

The False Match Rate is the expected probability that a sample will be falsely declared to match a single randomly-selected template (*false positive*). The False Non-Match Rate is the expected probability that a sample will be falsely declared not to match a template of the same measure from the same user supplying the sample (*false negative*) [9].

The **FMR(*t*)** and **FNMR(*t*)** curves are computed from *gms* and *ims* distributions for t typically ranging from 0 to 1. Given a threshold t, **FMR(*t*)** and **FNMR(*t*)** are defined as follows [16]:

$$FMR(t) = \frac{card\{ims_{ik} | ims_{ik} \geq t\}}{NIRA} \tag{5}$$

$$FNMR(t) = \frac{card\{gms_{ijk} | gms_{ijk} < t\} + REJ_{NGRA}}{NGRA} \tag{6}$$

where *card* represents the cardinality.

The evaluation of the overall accuracy level of a biometric system is often evaluated by considering two error plots. The first is the Receiving Operating Curve (**ROC**), where (1– **FNMR**) is plotted as a function of **FMR** for all available values of t. The second, and most used, is the plot of FNMR *vs.* FMR in a logarithmic chart, called the Detection Error Trade-off (**DET**) plot. Figure 5 shows patterns of the DET curves computed for 6 different systems [17]. The best system is the one that has its DET curve below all the others. It would mean that, for all the values of its threshold t, the system yields the lowest FMR and FNMR with respect to the others. Typically a system outperforms all the others in *some* intervals of threshold t, not for all the values. DET plots are suitable to compare biometric systems.

In order to evaluate the peculiar behaviour of a selected system in separating the genuine from the impostor attempts, the *distributions* of the matching function values of the genuine population ($\mathbf{gms_{ijk}}$) and of the impostor population ($\mathbf{ims_{ik}}$) can be plotted. The smaller the overlap (the darker area in Figure 6), the better the biometric system will be. If no overlap occurs, it means that it exists a threshold value t' which perfectly separates the genuine individuals from the impostors (ideal case).

Other error-indexes can complete the accuracy description. The **EER** (Equal Error Rate) is often considered, and it is computed as the point where **FMR(*t*)** = **FNMR(*t*)**. Score distributions are typically not continuous and the **EER** must be often interpolated by the quantized data [17].

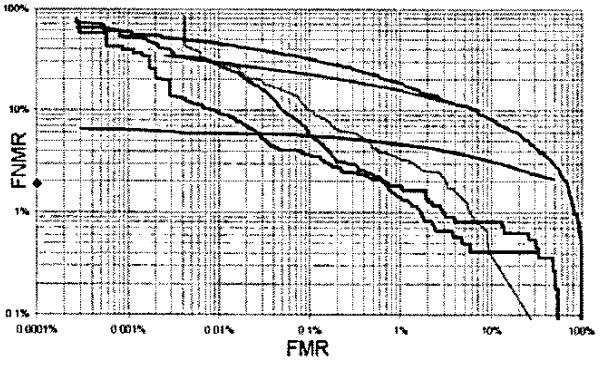


Figure 5: Examples of DET curve.

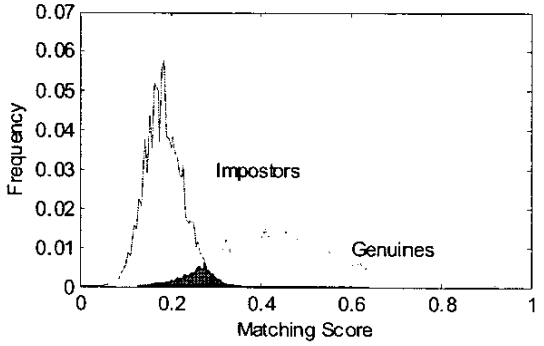


Figure 6: Examples of genuine and impostor distributions.

Other indexes measure the capability of the biometric system to *acquire* sample or to *process and enrol* templates: performance indexes. The former is the Failure to Acquire Rate (**FTA**) and it is "*the expected proportion of transactions for which the system is unable to capture or locate an image or signal of sufficient quality*" [10]. The latter is named Failure to Enrol Rate (**FER**) and it represents the "*expected proportion of the population for whom the system is unable to generate repeatable templates*" [10]. Examples are: individuals that are unable to present the required biometric feature, samples that have insufficient quality at enrolment, and those who cannot reliably match their template. For example, it has been estimated that about 2%–3.5% of individuals have their fingerprint ridges damaged by friction during a two-year period [20].

In order to shorter the matching time, some systems can sort/organize templates into bins. The Penetration Rate (**PR**) is defined as "*the expected proportion of the templates to be searched over all input samples under the rule that the search proceeds through the entire partition regardless of whether a match is found*" [10]. Of course, if the system fails to recognize the proper partition of a new sample we have a binning error. This proportion of misplaced samples represents the Binning Error Rate (**BER**).

In the literature many other indexes are present for testing biometric system's performances, but unfortunately they depend on the envisioned system's structure (identification/verification, fixed threshold, number of enrolled users and number of templates per user) [10]. This issue must be carefully taken into account comparing different systems [9]. The most common are **False Accept Rate** (FAR) and **False Reject Rate** (FRR). Considering also the Binning Error Rate (BER) and penetration rate (PR), and if the acceptance depends on a single successful match, we can write

$$FAR = PR \times FMR \times (1 - FTA) \quad (7)$$

$$FRR = FTA + (1-FTA) \times BER + (1-FTA) \times (1-BER) \times FNMR \quad (8)$$

It is worth noting that it is a non-sense to describe the system performance by only its FAR or FRR. The two indexes must be both provided since they depend on the fixed threshold t: changing t it is possible to arbitrarily reduce one of the two.

V. CONFIDENCE OF ACCURACY ESTIMATION

The evaluation of confidence of the accuracy computed in previous sections and its relationship to the dataset size are now discussed. The proposed approach and definitions are generally used when describing a biometric system (see for example ref. 9). In general, "a N% *confidence interval* for parameter x consists of a lower estimate L, and an upper estimate U, such that the probability of the true value being within the interval estimated is the stated value (e.g.: $P(x \in [L, U]) = N\%$)" [10]. Of course, the smaller the evaluation test size, the wider the confidence interval will be.

The "size" of an evaluation test can be thought in terms of the number of volunteers involved in the testing phase and the number of attempts made. The criterion used to choose volunteers/samples will influence how accurately error rates can be measured. In the literature, the term "*Non-self*" is used in the sense of "genetically different". It has been noted [27-29] that comparison of genetically identical biometric characteristics (for instance, between a person's left and right eyes or across identical twins) yields, on average, more similar score distributions than comparison of genetically different characteristics. Consequently, such genetically similar comparisons could not be considered in computing the false match rate.

It must be also noticed that the assumption about independency of all trials is not always satisfied (i.e. asymmetric/symmetric matching values in the ***igm*** matrix, problem related to "*Non-Self*" samples). The alternative is to compromise the independence of the samples by reusing a subset of all the volunteers and to expect a loss of statistical significance [10]. The actual consequence of not-independent samples in the test-database for a biometric system is not well understood yet [9].

Furthermore, performance estimates will be affected by both systematic errors and random errors. In biometric systems, by definition, random errors are due to the natural variation in people employed in the test, samples *etc.* Instead, systematic errors are due to bias in the test procedures, *etc.* For example, if certain types of individuals are under-represented in the volunteer set, this can give rise to a "bias" in the results [10]. It is fundamental to reduce the bias as much as possible and to report it into the results of the analysis. This allow for further fair comparisons between experiments.

It is interesting to note that some biometric producers state part-per-million (p.p.m.) errors in their systems, but errors in the data-collection procedures are typically considered much higher (due to "human errors" or factors such as iris/fingertips illness/injures previously described) [9, 20].

Dimensioning the test size, two main rules can be followed. They are known in the literature as the *rule of 3*, and the *rule of 30*. The *Rule of 3* [30-32] addresses the question "What is the lowest error rate that can be statistically established with a given number N of independent comparisons?". This value is the error rate p for which the probability of zero errors in N trials is, for example, 5%. This gives $p \approx 3/N$, for a 95% confidence level. For example, a test of 300 independent samples returning no errors can be said with 95% confidence to have an error rate $\leq 1\%$ [10]. The *Rule of 30* [33] is utilized to determine the evaluation test size and it can be expressed as follows: "To be 90% confident that the true error rate is within ±30% of the observed error rate, there must be at least 30 errors". So, for example, if we have 30 false non-match errors in 3,000 independent genuine trials, we can say with 90% confidence that the true error rate is between 0.7% and 1.3%. These rules have been derived from the *binomial distribution* assuming independent trials, and may be applied by considering the performance expectations for the evaluation. The two rules should be considered as over-optimistic [9].

Using a number of samples sufficiently large, the *central limit theorem* [34] implies that the observed error rates should

follow an approximately *Gaussian (or normal) distribution.*

[7] The Biometric Evaluation Methodology Working Group, 'Common Methodology for Information Technology Security Evaluation', 2002