Quality Assessment of Biometric Systems: A Comprehensive Perspective Based on Accuracy and Performance Measurement

M. Gamassi, Massimo Lazzaroni, *Member, IEEE*, M. Misino, Vincenzo Piuri, *Fellow, IEEE*, D. Sana, and F. Scotti, *Member, IEEE*

***Abstract*—Despite the efforts of the international biometric community, the measurement of the accuracy of a biometric system is far from being completely investigated and, eventually, standardized. This paper presents a critical analysis of the accuracy and performance measurement of a biometric system. Current approaches to the problem and procedural error have been described and criticized. Finally, a methodology for the measurement of the accuracy of biometric systems with nonsymmetric matching function will be proposed and discussed.**

***Index Terms*—Accuracy, biometric systems, fingerprint, iris recognition, matching algorithms, measurement, uncertainty.**

I. Introduction

BIOMETRIC systems have been defined by the U.S. National Institute of Standards and Technology (NIST) [1] as systems exploiting "automated methods of recognizing a person based on physiological or behavioral characteristics" (*biometric identifiers,* also called *features*). Physiological biometrics is based on data derived from direct measurement of a body part (i.e., fingerprints, face, retina, iris), while behavioral biometrics is based on measurements and data derived from a human action [2] (i.e., gait and signature).

Biometric systems are being used to verify identities and restrict access to buildings, computer networks, and other secure sites [3]. Recent global terrorism is pushing the need for secure, fast, and nonintrusive identification of people as a primary goal for homeland security. As commonly accepted, biometrics seems to be the first candidate to efficiently satisfy these needs. For example, by October 2004, the United States planned to control the accesses to/from country borders by biometric passports [4], [5].

Personal identification has taken the form of token-based or knowledge-based methods, such as secret passwords and personal identification numbers, ID cards, keys, passes etc. The biometric approach completely differs since the identification is based on personal and unique peculiarities of individuals which cannot be easily misplaced, forged, or shared [6].

Given that a biometric system is an identification system, its accuracy can be evaluated by classical techniques [7] but peculiarities are present. Typically, to effectively test biometric systems, a great number of volunteers are required, or a large database of biometric records must be accessed [7]–[9]. Experiments are complex and expensive, and they expose the data maintainer to important problems related to the security and privacy of the biometric records. Furthermore, the protocol of the experiments can directly impact the system accuracy [9], [10], and it is not possible to resume the overall system performance in a single index of accuracy to simply compare two different biometric systems.

This paper aims to present a critical analysis of the accuracy and performance measurement methodology of a biometric system and proposes how to consider in the methodology biometric systems that have a nonsymmetric matching function. Section II presents the more frequently studied biometric systems in the literature and their peculiarities. Section III introduces the terms and the theory of the measurement of accuracy of a biometric system. Section IV describes and criticizes current best practices as well as proposes how to evaluate nonsymmetric matching function systems into the comprehensive framework of accuracy evaluation. Finally, Section V presents statistical considerations concerning the interval of confidence of the accuracy estimation and typical errors in setting up the biometrics experiments.

II. Biometric Systems

From the literature, a biometric system has a general structure. Fig. 1 shows the components of a biometric system according to [10]. First of all, a sensor acquires a *sample* of the user presented to the biometric system (i.e., fingerprint, face, iris images). As defined in [10], a sample is a biometric measure presented by the user and captured by the data collection subsystem as an image or signal. The sample can be transmitted, eventually exploiting by compression/decompression techniques. Some systems store the complete sample data in the storage unit. Storing samples is often deprecated in the literature due to privacy and security issues [11], [12]. It is important to note that the accuracy of the sensor used to pick up the sample of biometric data is only seldom studied and

Manuscript received June 15, 2004; revised April 17, 2005.

M. Gamassi, M. Lazzaroni, V. Piuri, D. Sana, and F. Scotti are with the Dipartimento di Tecnologie dell'Informazione, Università degli Studi di Milano, 26013 Crema (CR), Italy (e-mail: gamassi@dti.unimi.it; lazzaroni@dti.unimi.it; piuri@dti.unimi.it; sana@dti.unimi.it; scotti@dti.unimi.it)

M. Misino was with the Dipartimento di Tecnologie dell'Informazione, Università degli Studi di Milano, 26013 Crema (CR), Italy. He is now with Galileo Avionica, 20157 Milan, Italy (e-mail: misino@dti.unimi.it).

Digital Object Identifier 10.1109/TIM.2005.851087

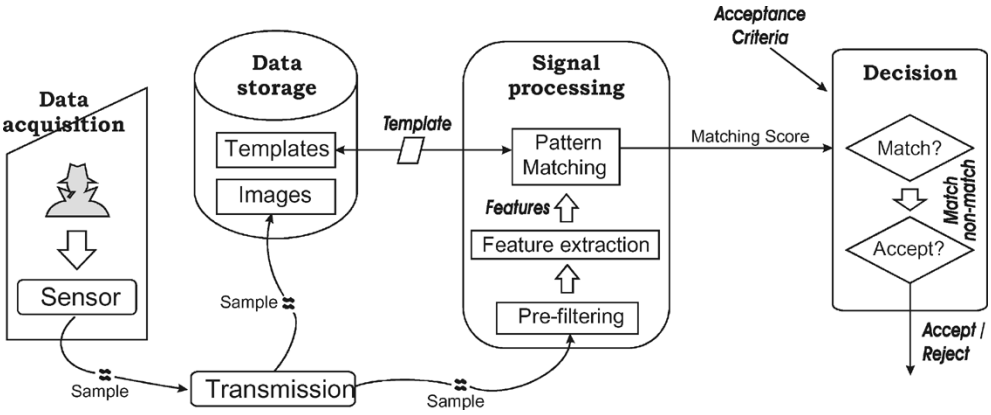


Fig. 1. Structure of a biometric system.

taken into account when the accuracy of a biometric system is evaluated.

Moreover, during the sample pickup phase, the presentation effect would be considered. In fact, there is a broad category of variables impacting the way in which the users' inherent biometric characteristics are displayed to the sensor: for example, in facial recognition, pose angle and illumination; in fingerprinting, finger rotation and skin moisture. In many cases, the distinction between changes in the fundamental biometric characteristic and the presentation effects may not be clear. A typical example of the presentation effect is the facial expression in facial recognition.

More correctly, a biometric system uses and stores only a mathematical representation of the information extracted from the presented sample by the signal processing module that will be used to construct or compare against enrolment templates: the *biometric feature*. Examples of features are minutiae coordinates and iris-codes. If the extracted feature is stored (enrolled) into the biometric system, a *template* for future identification or verification (matching) is added. Each biometric system has a measure of the similarity between features derived from a presented sample and a stored template. The measure produces a typical index called *matching score*. Hence, a match/nonmatch decision may be made according to whether this score exceeds a *decision threshold* or not. The term *transaction* refers to an attempt by a user to validate a claim of identity or nonidentity by consecutively submitting one or more samples, as allowed by the system decision policy [10].

Lastly, a transmission process is implemented to transmit the collected data to the signal processing section. During this transmission, some nondesired effect are presented and would be taken into consideration. In particular, the channel effect is defined as the changes imposed upon the presented signal in the transduction and transmission process due to the sampling, noise, and frequency response characteristics of the sensor and transmission channel.

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

Typically, we can identify the following attributes of a biometric system: 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 cannot assume that all people, for example, have all fingers or have one/two of the two irises not damaged.

The *Permanence* property is related to the requirement that the feature not change over time, or at least that it vary 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.

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.

The most cited biometric features in the literature are fingerprint, signature (handwriting), 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 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 compared *only* the one 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]. An 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 the same single biometric features [14]–[16].

III. Biometric System Evaluation

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 from a selected sensor technology [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 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 available sensors on the market. Of course, performance with this database will depend upon both the environment and the population in which it is collected. Typically, to avoid malicious approaches by the developers, it is possible to distribute to them only some examples of samples. Then, distribute actual evaluation data 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.

Fig. 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 contain complete biometric samples of real individuals. Security and privacy expects are seriously involved [11], [12]. Some synthetic databases/generators are available only for fingerprint biometrics [21].

IV. Accuracy and Performance Indexes

In the case of a *technology evaluation,* the accuracy indexes most commonly accepted in the literature are discussed in this section. This definition of accuracy presents differences with respect to the classical one used in metrology [22] but 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 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 to, for example, pressure, humidity, finger position,

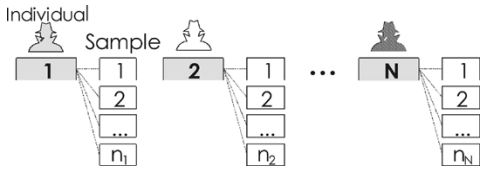


Fig. 2. General samples situation of a biometric dataset.

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 the presence of noise corrupted samples, the preferred method minimizes the effect of noise source on the accuracy.

The following theory is valid for monomodal and multimodal biometric systems. We can assume to have a sample database of identified individuals, as plotted in Fig. 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], [24]. Given two biometric features A and B and naming the matching function M, we have a symmetric matching function if $\mathrm{M(A,B)} = \mathrm{M(B,A)}$. In the following, we describe how to extend the equation for the accuracy evaluation for systems where we have $\mathrm{M(A,B)} \neq \mathrm{M(B,A)}$. 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 to make a fair comparison between different biometric systems by taking into account that issue.

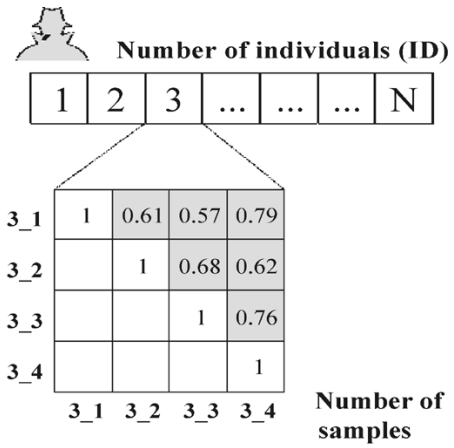
Referring again to Fig. 2, let us define B_{ij} as the jth sample of the ith biometric data (i.e., an image, either filtered or not); T_{ij} as the template computed from B_{ij} (the features extracted); n_i as the number of samples available for the ith biometric data, and N as the number if individuals enrolled. Let us follow the steps to compute the accuracy performance of the systems defining the proper indexes.

A. Step 1: Enrollment

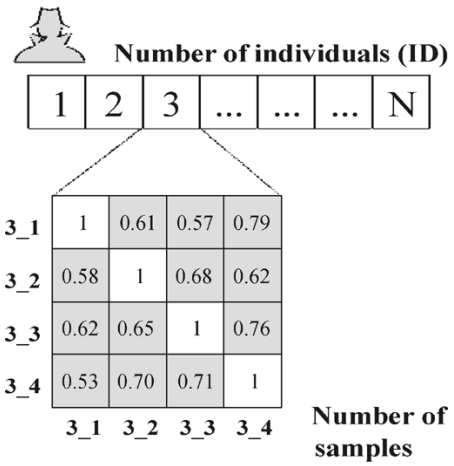
During the *enrollment* step, the presence of errors is monitored by using an index named $\mathbf{REJ_{ENROLL}}$. $\mathrm{REJ_{ENROLL}}$ is the rejection ratio in the enrolment phase, due to *Fail* (the algorithm declares it cannot enroll the biometric data), *Timeout* (the enrollment exceeds the maximum allowed time), and *Crash* (the algorithm crashes during biometric processing) situations. The templates T_{ij}, $i = 1,\ldots,N$, $j = 1,\ldots,\mathrm{n}_i - 1$ are computed from the corresponding B_{ij} and stored on disk; if something wrong happens, index $\mathrm{REJ_{ENROLL}}$, previously defined, has to be increased [10], [17].

B. Step 2: A General Matching Score Computation

For *symmetric matching functions,* the practice 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* $\mathbf{gms_{ijk}}$ [Fig. 3(a)]. The term "genuine"



a) *gms* for symmetric matchings



b) *gms* for asymmetric matchings

Fig. 3. Genuine matching scores.

refers to the fact that the matching is computed between samples of the same identity-certified individual. Since the matrix is symmetric by definition, *only* the upper triangular matrix is computed. Each individual has its squared gms matrix.

We now propose how to include systems that have asymmetric matching-function in the framework proposed in the literature. The next section considers its statistic effect on the estimation of the systems accuracy.

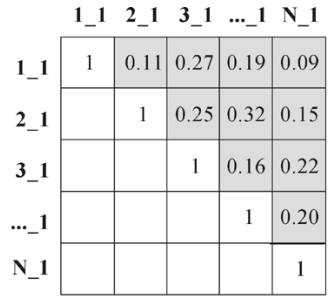
For *asymmetric matchings,* each biometric template T_{ij} successfully created in the previous step is matched against the biometric images B_{ik} $(1 \leq k \leq n_i, k \neq j)$ and the corresponding *genuine matching scores* matrix $\mathbf{gms_{ijk}}$ is stored [Fig. 3(b)]. The matrix is not symmetric but is still square. Then, the number of matches, denoted as *number of genuine recognition attempts* (NGRA), is given by

$$\mathrm{NGRA}_{\mathrm{symMatch}} = \frac{1}{2}\sum_{i=1}^{N} n_i(n_i - 1) \tag{1}$$

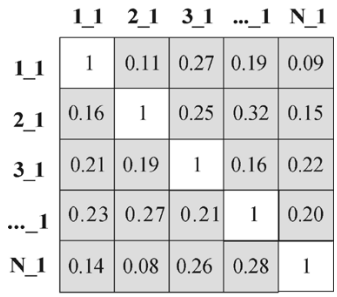
where $\mathrm{REJ}_{\mathrm{ENROLL}} = 0$ for symmetric matching and

$$\mathrm{NGRA}_{\mathrm{asymMatch}} = \sum_{i=1}^{N} n_i(n_i - 1) \tag{2}$$

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



a) *ims* for symmetric matchings.



b) *ims* for asymmetric matchings.

Fig. 4. Impostor matching scores.

Let us now consider the matching values of samples of *different* individuals (*impostors matching*). For symmetric matching, each biometric template T_{i1}, $i = 1, \ldots, N$ is matched against the first biometric image from different data B_{k1} $(1 < k \leq N)$ and then the corresponding *impostor matching scores* $\mathbf{ims_{ik}}$ matrix is stored [Fig. 4(a)]. Impostor matching in the case of asymmetric matching function is computed as follows: each biometric template T_{i1}, $i = 1, \ldots, N$ is matched against the first biometric image from different data B_{k1} $(1 \leq k \leq N, k \neq i)$ and the corresponding impostor matching scores $\mathbf{ims_{ik}}$ matrix is stored [Fig. 4(b)]. The number of matches, denoted as *number of impostor recognition attempts* (NIRA), is given by

$$\mathrm{NIRA}_{\mathrm{symMatch}} = \frac{1}{2}N(N-1) \tag{3}$$

if $\mathrm{REJ}_{\mathrm{ENROLL}} = 0$ for symmetric matching and

$$\mathrm{NIRA}_{\mathrm{asymMatch}} = N(N-1) \tag{4}$$

if $\mathrm{REJ}_{\mathrm{ENROLL}} = 0$ for 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 $\mathrm{REJ}_{\mathrm{NGRA}}$ and $\mathrm{REJ}_{\mathrm{NIRA}}$ counters. It follows that *gms* and *ims* matrixes can have missing values. Commonly, 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 $\mathbf{FMR}(t)$ and false nonmatch rate $\mathbf{FNMR}(t)$. 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 nonmatch 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 $\mathbf{FMR}(t)$ and $\mathbf{FNMR}(t)$ curves are computed from *gms* and *ims* distributions for t typically ranging from zero to one. Given a threshold t, $\mathbf{FMR}(t)$ and $\mathbf{FNMR}(t)$ are defined as follows [16]:

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

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

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). The ROC is a graphical plot of the fraction of true positives versus the fraction of false positives for a binary classifier system as its discrimination threshold is varied. In our case, we have in the plot the (1-FNMR) quantity plotted as a function of FMR for all available values of t.

The second, and most used, is the plot of FNMR versus FMR in a logarithmic chart, called the detection error tradeoff (DET) plot. The DET plot can be used to directly compare biometric systems. Fig. 5 shows patterns of the DET curves computed for six different systems [17]. The best system is the one that has its DET curve below all other systems' curves since it yields lower FNMR and FMR errors with respect to all others biometric systems for *all* the values of the decision threshold t. In order to directly compare biometric systems, the FNMR and FMR errors of the systems plotted in the DET plot must be evaluated on the same dataset. As commonly happens, a system outperforms all the others in *some* intervals of threshold t, not for all values.

In order to evaluate the peculiar behavior 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 Fig. 6), the more ideal the biometric system will be. If no overlap occurs, it means that there exists a threshold value t' that perfectly separates the genuine individuals from the impostors (ideal case).

Others error indexes can complete the accuracy description. The *equal error rate* (EER) is often considered, and it is computed as the point where $\mathbf{FMR}(t) = \mathbf{FNMR}(t)$. Score distributions are typically not continuous and the EER must be often interpolated by the quantized data [17].

Others indexes measures the capability of the biometric system to *acquire* sample or to *process and enroll* templates: performance indexes.

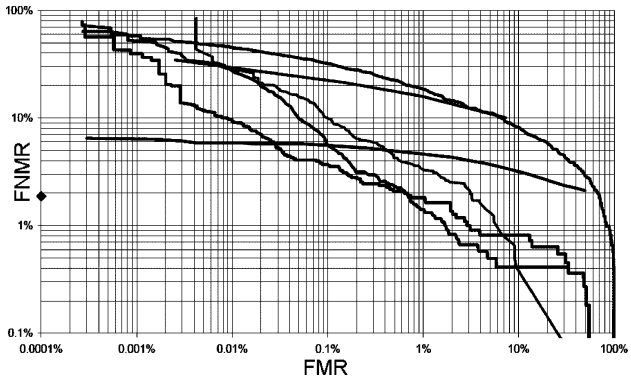


Fig. 5. Examples of DET curve.

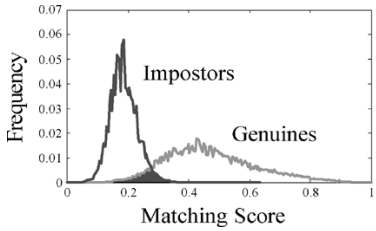


Fig. 6. Examples of genuine and impostor distributions.

The former is the failure to acquire rate (FAR) and 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 enroll rate* (FER) and 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 enrollment, and those that cannot reliably match their template in attempts to confirm the enrollment is usable. For example, it has been estimated that about 2–3.5% of individuals have their fingerprint ridges damaged by friction [20].

In order to shorten 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).

There are many other indexes useful for testing the performances present in the literature which depend on the envisioned system's structure (identification/verification, fixed threshold, number of enrolled users and number of templates per user) [10]. That issue must be taken into account comparing different systems [9]. Most common are FAR and false reject rate (FRR). Considering also the BER and PR, and if the acceptance depends on a single successful match, we can write

$$\mathrm{FAR} = \mathrm{PR} \times \mathrm{FMR} \times (1 - \mathrm{FTA}) \tag{7}$$

$$\mathrm{FRR} = \mathrm{FTA} + (1 - \mathrm{FTA}) \times \mathrm{BER} + (1 - \mathrm{FTA}) \times (1 - \mathrm{BER}) \times \mathrm{FNMR}. \tag{8}$$

It is worth noting that it is nonsense to describe the system performance by only its FAR or FRR. The two indexes must both be furnished since they depend on the fixed threshold t: shifting 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, [9]). In general, an "$\mathrm{N}\%$ *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., $\mathrm{P(x} \in \mathrm{[L, U])} = \mathrm{N}\%$)" [10]. Of course, the smaller the evaluation test size, the wider the confidence interval will be.

The "size" of an evaluation can be thought of in terms of the number of volunteers involved in the testing phase and the number of attempts made.

The criterion to choose volunteers/samples will influence how accurately error rates can be measured. In the literature, the term *nonself* 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 also be 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 *nonself* 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 nonindependent 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 variation in the biometric trait of 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 underrepresented in the volunteer set, this can give rise to a "bias" in the results [10]. Any efforts would be taken in account to limit the bias. The remaining bias would be estimated and taken into account or declared with the proposed results. The uncertainty associated with the measurement/estimation of the aforementioned remaining bias would be also evaluated, when possible.

It is interesting to note that some biometric producers state parts-per-million (p.p.m.) errors in their systems, but errors in the data collection are typically considered much higher due to "human errors" previously described or factors such as iris/fingertip illness/injures [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 can be expressed as follows: "To be 90% confident that the true error rate is within $\pm 30\%$ of the observed error rate, there must be at least 30 errors." So, for example, if we have 30 false nonmatch errors in 3000 independent genuine trials, we can say with 90% confidence that the true error rate is between 0.7% and 1.3%. This rule has 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 overoptimistic [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*. Under the assumption of normality, $100 \bullet (1-\alpha)\%$ confidence bounds on the observed error rates are given by the following formula:

$$\hat{p} \pm \mathrm{z}\left(1 - \frac{\alpha}{2}\right)\sqrt{\hat{V}\left(\hat{p}\right)} \tag{9}$$

where

$\hat{p}$ observed error rate;
$\hat{V}(\hat{p})$ estimated variance of observed error rate [9];
$z(.)$ inverse of the standard normal cumulative distribution.

For 95% confidence limits, the value $z(0.975)$ is 1.96.

Often this formula gives rise to negative values for the error rate—but negative error rates are impossible. This is due to nonnormality of the distribution of observed error rates. When a case like that occurs, nonparametric methods, such as *bootstrap*[35], can be used to obtain confidence intervals.

Finally, it must be noted that a biometric system is not more accurate just because it uses a more complicated feature than other systems. Statements such as "iris biometrics is more accurate than fingerprint because its biometric feature is much more complicated" are not correct. Under quite general assumptions, in [36] it has been demonstrated that the accuracy does not depend only on the number of degrees of freedom of the biometric features utilized. Of course, the accuracy depends on *how* information on the biometric features is collected/measured and, finally, used much more than the features' "complexity."

VI. Conclusion

In this paper, we summarize and critically discuss the main issues to be taken into account for the evaluation of the accuracy and performance of a biometric system. The case of technology evaluation has been considered according to current best practices. The discussed methodology has a general application to different sample database formats, and we propose how to support asymmetric matching algorithms. Our analysis shows that more efforts should be made to analyze the accuracy of the biometric systems from a stricter metrological point of view.

The estimation of uncertainty in biological and clinical measurements is a true critical point of such measurements and will be considered with an in-depth metrological approach.

REFERENCES

[1] A. K. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*. Norwell, MA: Kluwer, 1999.

[2] R. Bolle, S. Pankanti, and A. K. Jain, "Guest editorial," *IEEE Computer (Special Issue on Biometrics)*, vol. 33, no. 2, pp. 46–49, Feb. 2000.

[3] J. D. M. Ashbourn, *Biometrics: Advanced Identify Verification—The Complete Guide*. Berlin, Germany: Springer-Verlag, 2000.

[4] S. Waterman, "Biometric borders coming," *Wash. Times*, 2003.

[5] "Technology assessment: Using biometrics for border security," U.S. General Accounting Office, GAO-03–174, 2002.

[6] A. K. Jain *et al.*, "Biometrics: Promising frontiers for emerging identification market," *Computer*, vol. 33, no. 2, Feb. 2000.

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

[8] P. J. Phillips, A. Martin, and W. M. Przybocki, "An introduction to evaluating biometric systems," *IEEE Comput.*, vol. 33, no. 2, pp. 56–63, Feb. 2000.

[9] V. S. Valencia, "Biometric testing: It's not as easy as you think," in *Biometric Consortium Conf.*, Arlington, VA, Sep. 2003.

[10] A. J. Mansfield and J. L. Wayman, "Best practices in testing and reporting performance of biometric devices," National Physical Lab., Center for Mathematics and Scientific Computing, NPL Rep. CMSC 14/02, 2.01 ed., 2002.

[11] R. Clarke, "Biometrics and privacy," in *Proc. Computers, Freedom Privacy* San Francisco, CA, 2002.

[12] M. Crompton, "Biometrics and privacy: The End Of The World as We Know It Or The White Knight Of Privacy?," in *Proc. 1st Biometrics Inst. Conf.*, Mar. 20, 2003, [Online] http://www.privacy.gov.au/news/speeches/sp80notes.htm.

[13] R. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *IEEE Comput.*, vol. 33, no. 2, pp. 64–68, Feb. 2000.

[14] A. Ross and A. K. Jain, "Information fusion in biometrics," *Pattern Recognit. Lett. (Special Issue on Multimodal Biometrics)*, vol. 24, no. 13, pp. 2115–2125, Sep. 2003.

[15] A. Kumar *et al.*, "Personal verification using palmprint and hand geometry biometric," in *Proc. 4th Int. Conf. AVBPA*, Guildford, U.K., Jun. 2003, pp. 668–678.

[16] D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, *Handbook of Fingerprint Recognition*. Berlin, Germany: Springer-Verlag, 2003.

[17] D. Maio, D. Maltoni, R. Cappelli, J. L. Wayman, and A. K. Jain, "FVC2002: Second fingerprint verification competition," in *Proc. 16th Int. Conf. Pattern Recognition (ICPR)*, vol. 3, Québec City, Canada, Aug. 2002, pp. 811–814.

[18] FPVC2000 database [Online]. Available: http://bias.csr.unibo.it/fvc2000/databases.asp

[19] CASIA database [Online]. Available: http://www.sinobiometrics.com/casiairis.htm

[20] (2005) "Summary of NIST standards for biometric accuracy, tamper resistance, and interoperability"Internal Rep.. National Institute of Standards and Technology. [Online]. Available: http://www.itl.nist.gov/iad/894.03

[21] R. Cappelli, D. Maio, and D. Maltoni, "Synthetic fingerprint-database generation," in *Proc. ICPR*, vol. 3, Québec City, QC, Canada, 2002, pp. 744–747.

[22] International Bureau of Weights and Measures, ISO, IEC, OIML, IFCC, IUPAC, and IUPAP *et al.*, *International Vocabulary of Basic and General Terms in Metrology*, 2003.

[23] X. Jiang and W. Y. Yau, "Fingerprint minutiae matching based on the local and global structures," in *Proc. ICPR*, vol. 2, 2000, pp. 1038–1041.

[24] A. K. Jain, L. Hong, S. Pankanti, and R. Bolle, "An identity authentication system using fingerprints," *Proc. IEEE*, vol. 85, no. 9, pp. 1365–1388, Sep. 1997.

[25] Y. P. Huang *et al.*, "An efficient iris recognition system," in *Proc. Int. Conf. Machine Learning Cybernetics*, Beijing, China, Nov. 4–5, 2002, pp. 450–454.

[26] L. Yu, K. Q. Wang, C. F. Wang, and D. Zhang, "Iris verification based on fractional Fourier transform," in *Proc. 1st Int. Conf. Machine Learning Cybernetics*, Beijing, China, Nov. 4–5, 2002, pp. 1470–1473.

[27] H. H. Newman *et al.*, *Twins*. Chicago, IL: Chicago Univ. Press, 1937.

[28] J. Daugman and C. Downing, "Epigenetic randomness, complexity, and singularity of human iris patterns," in *Proc. Royal Soc. Biol. Sci.*, vol. 268, 2001, pp. 1737–1740.

[29] A. K. Jain, S. Prabhakar, and S. Pankanti, "On the similarity of identical twin fingerprints," *Pattern Recognit.*, vol. 35, no. 11, pp. 2653–2663, 2002.

[30] T. A. Louis, "Confidence intervals for a binomial parameter after observing no successes," *Amer. Statist.*, vol. 35, no. 3, p. 154, 1981.

[31] B. D. Jovanovic and P. S. Levy, "A look at the rule of three," *Amer. Statist.*, vol. 51, no. 2, pp. 137–139, 1997.

[32] J. L. Wayman *et al.*, "Technical testing and evaluation of biometric identification devices," in *Biometrics: Personal Identification in Networked Society*, A. K. Jain *et al.*, Eds. Norwell, MA: Kluwer, 2000, pp. 345–368.

[33] G. R. Doddington *et al.*, "The NIST speaker recognition evaluation: Overview methodology, systems, results, perspective," *Speech Commun.*, vol. 31, no. 2–3, pp. 225–254, 2000.

[34] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, 6th ed. Ames, IA: Iowa State Univ. Press, 1967.

[35] K. V. Diegert, "Estimating performance characteristics of biometric identifiers," in *Proc. Biometrics Consortium Conf.*, San Jose, CA, Jun. 1996.

[36] J. L. Wayman. (1999) Degrees of freedom as related to biometric device performance. *AVANTI* [Online] http://www.engr.sjsu.edu/biometrics/publications_degrees.html

M. Gamassi received the Ing. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

Since 2003, he has been a Research Associate with the Department of Information Technologies, University of Milan, Crema (CR), Italy. His research interests include biometrics identification systems, signal and image processing, and soft-computing technologies for high-level system design.

Massimo Lazzaroni (M'05) received the M.Sc. degree in electronic engineering and the Ph.D. degree in electrical engineering from the Politecnico di Milano, Milano, Italy, in 1993 and 1998, respectively.

From 1994 to 1995, he was with the Design Department, Tecnint HTE. From 2001 to 2002, he was an Assistant Professor of electrical and electronic measurements at the Dipartimento di Elettrotecnica, Politecnico di Milano. Since 2002, he has been Associate Professor of electrical and electronic measurements at the Department of Information Technologies, University of Milan, Crema (CR), Italy. The scientific activity he deals with includes digital signal-processing techniques applied to electrical measurements. In particular, his current research interests are concerned with the application of digital methods to electrical measurements and measurements on electric power systems under distorted conditions. Moreover, he is also involved in research activities for the development of industrial sensors and partial discharge measurement methods and techniques.

M. Misino received the Ing. degree in electronic engineering from the Politecnico di Milano, Milan, Italy, in 2003.

From 2003 to 2004, he was a Research Associate with the Department of Information Technologies, University of Milan, Crema (CR), Italy. His research interests included image processing and biometric technologies. He is now with Galileo Avionica, Milan, Italy.

Vincenzo Piuri (F'01) received the Ph.D. degree in computer engineering from Politecnico di Milano, Milano, Italy, in 1989.

From 1992 to 2000, he was Associate Professor in operating systems at Politecnico di Milano. Since October 2000, he has been a full Professor in computer engineering at the University of Milan, Crema (CR), Italy. He was a Visiting Professor at the University of Texas at Austin during the summers of 1993 to 1999. His research interests include distributed and parallel computing systems, computer arithmetic, application-specific processing architectures, digital signal-processing architectures, fault tolerance, neural network architectures, theory and industrial applications of neural techniques for identification, prediction, control, and signal and image processing. His original results have been published in more than 200 book chapters, international journals, and proceedings of international conferences.

Prof. Piuri is a member of ACM, INNS, and AEI. He was Associate Editor of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT and the IEEE TRANSACTIONS ON NEURAL NETWORKS. He was Vice President for Publications of the IEEE Instrumentation and Measurement Society, Vice President for Member Activities of the IEEE Neural Networks Society, and a member of the Administrative Committee of the IEEE Instrumentation and Measurement Society and the IEEE Computational Intelligence Society. He is President-Elect of the IEEE Computational Intelligence Society. In 2003, he received the IEEE Instrumentation and Measurement Society Technical Award for his contributions to the advancement of computational intelligence theory and practice in measurement systems and industrial applications.

D. Sana received the Ing. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

Since 2003, he has been a Research Associate with the Department of Information Technologies, University of Milan, Crema (CR), Italy. His research interests include multimodal biometrics identification systems, signal and image processing, and multiagent systems.

F. Scotti (M'03) received the Ing. degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 1998 and 2003, respectively.

Since 2003, he has been an Assistant Professor with the Department of Information Technologies, University of Milan, Crema (CR), Italy. His research interests include signal and image processing, soft-computing technologies for industrial applications, and high-level system design. His current research focuses on design methodologies and algorithms for multimodal biometric systems.