Biometric Systems Evaluation

Maria De Marsico *

Synonyms

Biometric Performance Evaluation; Biometric Performance Measurement; Biometric Testing; Biometric System assessment.

Definitions

In its wider definition, biometric system evaluation encompasses a set of procedures measuring different aspects of the system performance under objective and quantitative criteria, according to well-defined rules and in well-defined conditions. The latter guarantee reliability as well as generalizability of evaluation results and the possibility to fairly compare different systems. In particular, the evaluation generally relies on a suited benchmark used as ground-truth and should allow predicting the system performance over unseen biometric data under similar, possibly real-world, operating conditions where ground-truth is not available. The guidelines for evaluation activities themselves are continuously revised and the underlying rules and protocols are updated and extended, therefore this entry does not pretend to be exhaustive but rather aims at providing some fundamentals on this crucial topic.

* corresponding author

Background

According to one of the earliest definitions given by Jain, Ross, and Prabhakar Jain et al (2004): *Biometric recognition, or, simply, biometrics, refers to the automatic recognition of individuals based on their physiological and/or behavioral characteristics. By using biometrics, it is possible to confirm or establish an individual's identity based on "who she is", rather than*

by "what she possesses" (e.g., an ID card) or "what she remembers" (e.g., a password). In practice, a biometric system is a pattern recognition system that deals with biometric data, i.e., either physical or behavioural characteristics of individuals. The captured data (*sample*), e.g., a photo of the face or a fingerprint, is processed to extract a feature set (vector) according to the algorithms exploited for matching; for an enrolled individual, such vector is included in the individual's *template* that is inserted in the template set of the system (*gallery*). When an individual must be recognized, the system extracts the feature vector from the sample submitted for recognition (*probe*), and compares this vector with those contained in the gallery, or with suitable classification models built during a training phase. In many biometric systems, the basic operation for recognition relies on matching the feature vectors extracted from two samples of the same biometric modality/characteristic. As in any other case of comparison between natural items, even when samples come from the same individual it is not possible to achieve perfect equality due to several possible reasons, including but not limited to possible raising defects in the sensors, natural changes in the individual's physiological or behavioral characteristics (e.g., face ageing or changing expression), and varying ambient conditions like illumination. Therefore, it is not possible to expect perfect matching. Rather, for each comparison, the response is a matching score (derived, e.g., from either a distance or a similarity measure) between the feature vector extracted from the probe sample (P) and one in the gallery (G_i). As an alternative, in Machine Learning approaches, e.g., based on Support Vector Machines (SVM) or Convolutional Neural Networks (CNN), a training phase either builds a model for each individual, or a multi-class model for the entire system gallery. In this case, a new sample is "matched" against the model/submitted to the trained classifier: for a one-class model the score is, e.g., a level of confidence or the probability that the sample conforms to the model; for a multi-class model, the response is the label of an enrolled individual/class with, e.g., a level of confidence or the probability that the sample conforms to the model for that class. The higher the score, the higher the possibility that the two feature vectors come from the same individual, or that the feature vector from the probe conforms to the trained model. Due to the ease of mapping the definitions related to pairwise matching of feature vectors onto those related to measuring the conformity of a feature vector to a model, from now on, we will rely on the first ones.

Biometric applications

Confirming an identity claim (*verification*) entails a $1:1$ comparison of identities: the individual accessing the system claims an enrolled identity, and the feature vector(s) extracted from the incoming probe(s) is compared with the feature vector(s) stored for that identity. Establishing an identity (*identification*) entails a $1:N$ comparison of identities: the template(s) extracted from the incoming probe(s) (representing an unknown identity) is compared with all the template(s) stored in the gallery to return

the best match, possibly together with an ordered list of candidates (complete or partial). Closed-set identification assumes that all individuals accessing the application are enrolled in the system. Open-set identification is closer to the real-world and releases this assumption so that some probes can also belong to individuals that are not in the gallery (for example in applications relying on a watch-list). In this case, a Reject option must be added to the set of possible system responses. The elasticity of matching that allows recognizing an individual under different variations of the biometric traits is also the source of possible errors, that will be explored in detail in the following. Of course, a final decision cannot just consider the matching scores, but relies on a suitably defined threshold: a score that is higher than or equal to this threshold supports the hypothesis that the matched feature vectors come from samples belonging to the same individual; otherwise, the system will respond that the two feature vectors do not belong to the same individual. One of the aims of biometric evaluation is to estimate a suitable threshold to possibly use during real operation. This results from a compromise between the possible errors that the system can make.

Types of evaluation

Biometric Systems Evaluation is a multi-faceted activity. The most popular goal is pure performance testing of algorithms. This is generally carried out even in most international challenges but is only one form of biometric testing. Capture devices are often compared according to the quality of samples that they allow to acquire, e.g., the resolution of produced images. It is possible to run technical performance tests that are device-specific. Testing can also assess system security. This can be split into two main branches. The first one is the biometrics-related check for vulnerability to spoofing (or presentation) attacks of the algorithms. The second one is more cybersecurity-related and not specifically bound to biometric techniques, aiming at identifying possible vulnerabilities to intrusion into equipment and communication lines. As such, this entry will not consider it further since related techniques are dealt with by entries in suited chapters. When choosing a system among several competitors for concrete operation, other aspects deserve attention too. For instance, the throughput rate refers to the number of individuals that the system is able to process per unit time; cost/benefit evaluation takes into account computational and storage resources required, and the possibility to use the system in real-time settings; compliance with international standards and regulations assures to avoid legal as well as interoperability problems in wider operational settings; last but not least, especially when a system has to be deployed within an unattended framework, human factors deserve great attention. When no operator guidance is provided, or not a continuous one at least, like in biometric authentication to remote services, ease of use and user acceptance are important parameters. In general, the different measures can be computed for any biometric system.

In order to set up a suitable grounding, it is worth referring to the classification also reported by the International Standard ISO/IEC 19795-1 ISO (2006).

It is possible to classify biometric technical performance testing in three types: technology, scenario, or operational evaluation. The most popular one is technology evaluation, which is an offline evaluation (and possibly comparison) of one or more algorithms for the same biometric modality using either a pre-existing benchmark or a specially collected corpus of samples from the biometric trait at hand. Of course, repeatability is an important characteristic of technology evaluation. In scenario evaluation, the system performance is determined exploiting a prototype application in a real context. In operational evaluation, the performance of a complete biometric system is determined also taking into account a specific real application environment with a specific target population. Both the latter cases entail an end-to-end (E2E) biometric system evaluation, where each activity is monitored from only two observation points: the input and the output, no matter what happens inside the system.

In summary, technical performance testing generally aims at estimating system error rates and throughput rates in standard/abnormal operation conditions, in order to predict/tune the system behaviour in real-world settings.

Evolution

Regarding the evaluation settings (benchmarks and associated protocols) it is possible to identify a precise trend towards repeatability and generalizability of results. The first pioneering works mostly used in-house, small datasets, generally acquired in well-controlled conditions and followed quite simple protocols (see for instance Turk and Pentland (1991)). It is not surprising that such works reported extremely encouraging accuracy. However, as the evaluated systems start tackling more realistic settings (e.g., sample capture in uncontrolled conditions or at-a-distance), the dramatically changed conditions require increasingly challenging datasets. Actually, new ones are collected when moving to novel problems. For instance, the face ageing problem was not among the first ones that were addressed. In addition, precise and accurate protocols allow fair and reliable comparison of system performance, and together with shared benchmarks are the basis for an increasing number of international challenges related to the different biometric modalities and issues. An example of fair organization of biometric evaluation challenges can be found on the related page of NIST (U.S. National Institute of Standards and Technology): https://www.nist.gov/itl/iad/image-group/resources/biometrics-evaluations

Errors of a biometric system

Whatever the kind of recognition application, the main goal of a biometric system design is to achieve a satisfying accuracy, i.e., to maximize the accuracy of biometric decision-making/to minimize the number of erroneous decisions of the classifier underlying the recognition process. The evaluation aims at getting objective quantitative measures of accuracy, that vary according to the type of application. The errors that a system can make are better understood

starting from the verification case, that in the simplest case requires a single matching operation for each input. It is possible to compute two score distributions: the distribution of scores generated from pairs of samples from the same person (often called the *genuine* distribution with special reference to the verification applications) and from different persons (often called the *impostor* distribution, also in this case with special reference to a false identity claim). The errors come from the possible overlap of the two distributions so that the more these distributions are well separated the better. In statistical terms, it is possible to formulate the errors in terms of hypotheses involving the probe P and the Gallery template G_i belonging to a certain enrolled individual. The null and alternate hypotheses and associated decisions are traditionally set as:

- H_0: the input P does not come from the same individual as G_i;
- H_1: the input P comes from the same individual as G_i;
- D_0: the system decides that the samples belong to different individuals (in verification, the identity claim is deemed false and rejected);
- D_1: the system decides that the samples belong to the same individual (in verification, the identity claim is deemed true and accepted).

The decision is based on the comparison of the matching score with an acceptance threshold, so that a score higher than or equal to the threshold supports H_1, while a score lower than the threshold supports H_0. It is now possible to define the errors in statistical terms:
Type I: false match - FM (D_1 is decided when H_0 is true);
Type II: false nonmatch - FNM (D_0 is decided when H_1 is true).
It is worth considering that the decision score can be based either on a similarity (the higher the better, as above) or on a distance measure. In this latter case, the role of the threshold and how it influences the system decision is of course symmetrical. It is interesting to underline that, once a normalization in the range $[0,1]$ is preliminarily computed, it is always possible to map one measure onto the other. Therefore, in the following, we will refer to similarity scores.

The probability density function representing the distribution of scores obtained from samples of different subjects, and the probability density function representing the distribution of scores obtained from samples of the same subject are typically close to normal ones. The errors happen in the overlapping region of the two distributions (see Figure 1). It is not possible to minimize both types of errors at the same time since they inversely depend on the same threshold. A low threshold takes to increase the number of FMs, while a too high one increases the FNMs to a point that the system is hardly usable. A suitable compromise depends on the needs of the application using the biometric recognition: the higher the requested security, the lower the relevance of FNM.

In evaluation, it is seldom useful to measure the absolute number of errors, but rather the probability that each type of error occurs. This probability is computed as the rate of errors with respect to the corresponding set of matching operations. False Match Rate (FMR) at a given threshold is computed as the ratio between the number of False Matches and the number of pairs where the samples belong to different individuals

(in verification, the number of impostors claiming a false identity and accepted). False Non Match Rate (FNMR) at a given threshold is computed as the ratio between the number of False Non Matches and the number of pairs where the samples belong to the same individual (in verification, the number of genuine users claiming a true identity and rejected). Figure 1 summarizes the main concepts mentioned up to now.

Verification errors are often also referred to as False Accept with its False Acceptance Rate (FAR) and False Reject with its False Reject Rate (FRR) respectively. Even though they may appear as synonyms, there is a difference that comes out when the evaluation does not only take into account the errors made by the matcher/classifier, usually measured "statically" on a given dataset, but also other problems that can occur, e.g., in a real-time system during end-to-end types of evaluation, with online capture of probe samples. In this case, a quality check over the submitted samples can be carried out to increase the reliability of the response. A failure-to-acquire happens when the quality of the captured sample is not deemed sufficient to carry out matching so that the system returns a Reject response for the recognition attempt. If the system allows a single attempt, and the user is a genuine one, then a failure-to-acquire causes a false reject. If the verification system allows several attempts to be matched to an enrolled template, a false rejection can result from any combination of failures-to-acquire and false non-matches over the different attempts, given that the user is a genuine one. On the other hand, a false acceptance always results if a sample is acquired and falsely matched to the enrolled template for the claimed identity on any of the attempts. In other words: *"False match or non-match rates are calculated over the number of comparisons, but false accept or reject rates are calculated over transactions and refer to the acceptance or rejection of the stated hypothesis, whether positive or negative. Furthermore, false accept or reject rates are inclusive of failures-to-acquire "*ISO (2006). In addition: *"The transaction will consist of one or more attempts as allowed or required by the corresponding decision policy. For example, the decision policy may allow three attempts to verify; then the transaction may consist of one attempt, two attempts if the first attempt is rejected, or three attempts if the first two attempts are rejected* ISO (2006). Assuming that the system only allows a single attempt, and given that we are able to estimate a reliable Failure-to-Acquire rate (FTA), it is possible to establish a relation between FRR(FAR) for a given threshold and FNMR(FMR) at the same threshold:

$$\text{FRR} = \text{FTA} + \text{FNMR} \times (1 - FTA)$$

$$\text{FAR} = \text{FMR} \times (1 - FTA)$$

A quality check can also be carried out on enrolment samples so that in end-to-end evaluations it is also possible to report and quantify the Failure-to-Enrol rate (FTE).

Even in offline evaluation using a static dataset, the biometric system may also include other modules besides the matcher, e.g., a segmentation module to extract the biometric trait from the sample. In this case, the accuracy of, e.g., the face detection module of a face recognition system can be included in the overall system evaluation by measuring the frequency of wrong detection, deleting the involved samples,

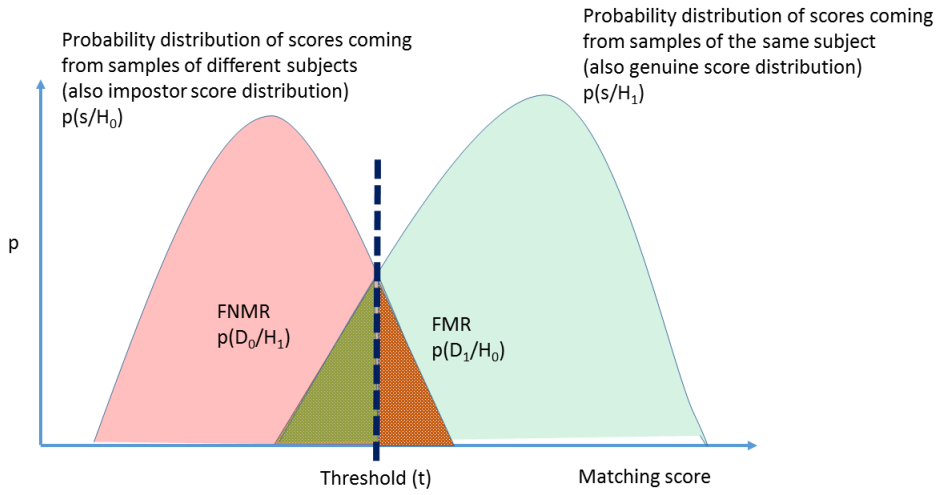

Fig. 1 Probability distributions of similarity scores corresponding to matching samples either of different individuals or of the same individual, with the possible errors.

and expressing that frequency as FTA. After having underlined their difference with FMR and FNMR, in the following, we will use FAR and FRR.

In open-set identification, the incoming sample is generally compared with all those in the gallery, possibly a watchlist, so that an ordered list of scores is returned (possibly limited to a certain rank or number of positions, or even to the first one only). It is still possible to consider FNMR(t) and FMR(t), with the threshold t determining the possible Reject response. In this case, a correct identification requires that the matching score of the probe with a template of the correct identity both exceeds the acceptance threshold (detection) and is the first one in the returned list (identification at rank 1). Detection and Identification Rate at threshold t and rank k, denoted as DIR(t,k), is an estimate of the probability that a subject in the gallery/watchlist is correctly detected and returned within the k-th position in the returned ordered list. False Rejection Rate at threshold t denoted as FRR(t) is defined as 1-DIR(t,1), while False Alarm Rate at threshold t, denoted as FAR(t), is an estimate of the probability of an incorrect detection on an individual who is not in the gallery/watchlist. By explicitly taking into account Failure-to-Acquire (FTA) it is possible to adopt the notation also used in ISO (2006), that is especially suited to end-to-end evaluation, and to use the following acronyms: FPIR is the false-positive identification-error rate; FNIR is the false-negative identification-error rate; FTA is the failure-to-acquire rate; FMR is the false match rate; FNMR is the false non-match rate; N is the number of templates in the database. Error definitions follow:

$$\text{FNIR} = \text{FTA} + (1 - \text{FTA}) \times FNMR$$

FPIR = (1 - FTA)$\times(1-(1-FMR)^N)$

In closed-set identification, each probe always belongs to a subject in the gallery. It is compared with all those in the gallery, and also in this case an ordered list of scores is returned. The only error that it is possible to make in closed-set identification is to return the correct identity in a position (rank) higher than the first one.

More recently, another type of error for a biometric system has attracted the research interest: it is the failure to detect either a spoofing attack or a fake sample. The main difference is that in the first case the "direct" aim of the attacker is to impersonate an enrolled subject by using a forged sample, e.g. a photo of the face or a silicon fingerprint. This attack is especially targeted at verification systems, therefore it is possible to measure the matching errors affecting the biometric system with the addition of the failures to detect spoofing. The problem of fake samples is recently attracting attention thanks to artificial intelligence-based deepfake techniques, whose aim is, e.g., to deceitfully attribute actions or words to a certain individual. In both cases, it is possible to reduce the problem of forgery detection to a binary classification genuine/fake. Its solution depends on the biometric trait at hand (e.g., face or fingerprint) and the type of attack (e.g., photo or video). The biometric community is organizing specific challenges based on suited datasets and establishing an ad-hoc terminology for possible errors. When either there is no specific presentation attack detection (PAD) module or some strategy is embedded directly in the biometric system, and the probe set includes attacks, it is possible to measure the Impostor Attack Presentation Match Rate (IAPMR). This is the rate of presentation attacks which are accepted as genuine samples, and can also be referred to as Spoofing False Accept Rate (SFAR). It is possible to measure the distribution of verification scores that come from spoofing attacks. They usually fall in a central region of the complete range of scores, therefore, overlapping both the distribution of scores coming from samples of the same individual and that of scores coming from samples of different individuals (see 1). When a separate PAD module is present, it is possible to measure its classification errors: Attack Presentation Classification Error Rate (APCER) is the rate of PAs incorrectly classified as normal presentations, while Normal Presentation Classification Error Rate (NPCER) is the rate of normal presentations incorrectly classified as PAs. It is possible to combine the results of a PAD module with those of a biometric system in different ways. For each probe submitted to a system, spoofing classification can run in parallel or precede the actual individual recognition. In the first case, the compound system can either return a single Accept/Reject response, or a pair of responses to be suitably combined.

Theory and Application

Technology evaluation

This kind of evaluation will be discussed in more detail because it is by far the most popular and most often found in literature, and also because several Fig-

ures of Merit (FOMs) are also used for the other two types of evaluation.

It is possible to identify different FOMs for different biometric applications. This entry discusses the ones that are most frequently used in literature.

In order to evaluate the performance of a biometric verification system, the error rates (FAR and FRR) achieved over a given dataset and according to a certain protocol are plotted at all the operating points (thresholds). The two curves present of course opposite behaviours (the higher the required acceptance score, the lower the FAR and the higher the FRR). Expressing system performance with a single value is better suited for comparison, and this value is traditionally the Equal Error Rate (EER) that is achieved at the threshold where FAR=FRR. The lower the EER, the better is the system performance. It is possible to report a cumulative error measure for each threshold t by taking the average (Half Total Error Rate) of FAR(t) and FRR(t): this is denoted as HTER(t). Other operating points and associated error values can be used for evaluation. A frequently considered one when dealing with a high-security application is the FRR corresponding to FAR at either 0 (ZeroFAR) or at very small values, e.g. FAR=10^{-3}: the lower the FRR, the better (the lower are the troubles for genuine users). The corresponding estimated thresholds can be used as parameters during real operation. Figure 3 shows FAR and FRR curves as functions of the threshold, and the mentioned relevant values. It is also possible to use a similar criterion considering FAR where FRR is very low but, of course, this is far less used.

An overall idea of the system behaviour can be obtained by depicting it in the form of a Receiver Operating Characteristic (ROC) curve. A ROC curve is a plot of FAR against (1-FRR) or Genuine Acceptance Rate (GAR) at the same threshold values (Figure **??**). Also in this case, a single value can be produced that is the Area Under Curve (AUC), expressing in a single value the "quality" of the ROC (the closer to 1 the better).

A further alternative is the Detection Error Tredeoff (DET) curve, which rather plots the FRR vs. FAR and that with suitable value scaling appears more linear than the corresponding ROC curve.

Similarly to verification, in open-set identification, it is possible to identify the threshold producing the Equal Error Rate (EER) where FRR=FAR. Plotting DIR versus FAR produces open-set ROC or DET.

Regarding closed-set identification, the most important FOM is the Recognition Rate, expressing the probability that the system returns the correct identity as the first element in the ordered score list. This probability is estimated as the rate of correctly identified probes over the total number of submitted ones since all of them belong to individuals in the gallery. It is also possible to get a deeper insight into the system behaviour by considering the positions after the first one. The Cumulative Match Score (CMS) at rank k, denoted as CMS(k), estimates the probability of returning the correct identity within the first k positions. The Cumulative Match Curve (CMC) plots the sequence of CMSs at increasing ranks: the steeper the curve, the higher the possibility to find the correct identity close to the head of the list. This can be useful in forensic applications, where a human expert

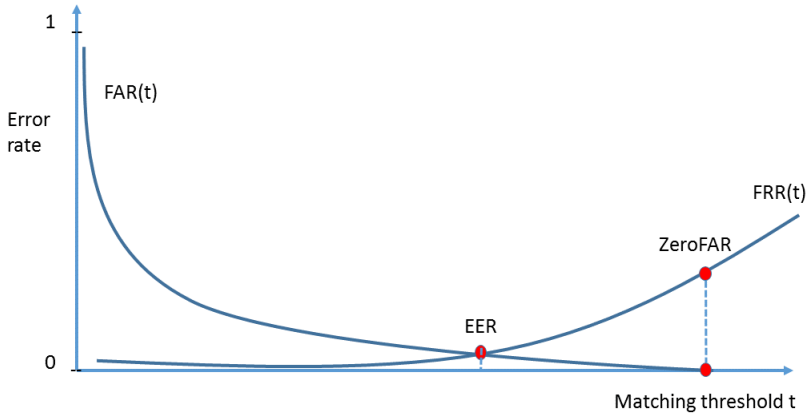

Fig. 2 FAR and FRR curves, with EER and ZeroFAR points.

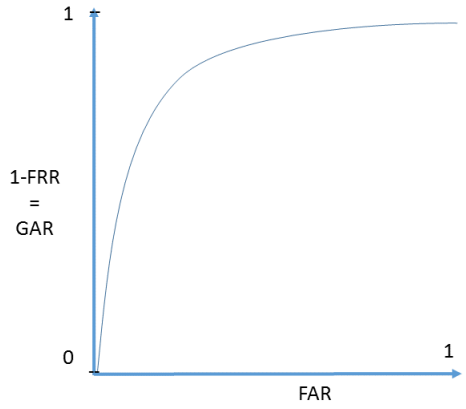

Fig. 3 The ROC curve from a ”good” matcher.

evaluates a set of candidates returned by the system for a final decision. Further performance parameters that are often used in literature are CMS(5) and CMS(10) that, together with CMS(1) or RR are an alternative to the curve to express the speed of convergence towards the correct identity.

Regarding the evaluation of spoofing attack detection, it is possible to obtain FOMs similar to those used for the biometric systems. In order to take into account false acceptance due to undetected spoofing attacks, it is possible to use a weighted combination of the error rates for the two negative classes (FAR from impostors and SFAR from presentation attacks) and find the operating point where the difference with FRR is minimized (similar to classical EER). It is also possible to compute a corresponding modified HTER and to plot it together with SFAR, to have more insight into the behaviour of the compound system.

The gallery composition, i.e., whether it contains either one or multiple templates per enrolled subject, can affect the system performance. Several works in the literature show a significant increase in performance when the gallery contains multiple templates. This happens especially when it is possible to implement a multiple-session enrollment, with sessions sufficiently separated in time to increase the representativeness of the trait variations. Of course, all the above FOMs can be computed given that suitable policies specify how to compare a probe with a set of templates per user. In general, the best matching score per individual is used in measuring the possible recognition outcome.

It is natural to wonder whether a particular partition of the dataset used as a benchmark can influence the results of the evaluation. Of course, this can happen. For instance, for an open-set identification system, if only the best samples appear as probes, e.g., face images with good resolution, semi-neutral expression, homogeneous illumination, etc., the system will eventually achieve much better results than using samples with varying distortions. Similarly, when using a machine learning approach and exploiting classifier training, overfitting can be caused by a poor choice of the training set with respect to validation and testing sets. The training set is used to train the classifier, the validation set is used to tune parameters by analyzing how they affect performance, and finally, the testing set is used to evaluate system performance. Whatever is the partition of the benchmark, there must be no overlap among the different subsets. In order to evaluate the generalization ability of the system to unknown subjects, some of them may only be included in the testing set. Of course, wherever a gallery and a probe set appear in the process, they must have no overlap in terms of samples; in open-set identification testing, not all subjects have samples in the gallery. In all these cases, cross-validation is one of the most used techniques to avoid possible bias due to a particular choice of subsets. In any case, when comparing different systems, it is necessary to assure that all of them work on the same partitions of the benchmark. This consideration underlies the set-up of well-organized challenges, where the performance evaluation protocol prescribes both FOMs and the benchmark partitioning(s). Moreover, it has become common to

maintain a sequestered subset of the benchmark dataset to avoid a too strict dependence of the algorithms on the known samples. Actually, the possible subsets structuring entailed by the evaluation protocols can be extremely articulated, and the same dataset with different partitioning can be used to assess performance in different operation scenarios. Figure 4 summarizes some of the main choices; TR is the training subset (for simplicity, this also includes the validation subset) that may not contain all the subjects included in the testing subset TS; the probe set used during testing possibly includes subjects not registered in the gallery (P_G vs. P_N). Regarding the possible bias caused by benchmarks, it is interesting to refer to the analysis in Torralba and Efros (2011).

Scenario evaluation

In a scenario evaluation, both a real environment and a real-world application are possibly simulated. The mentioned FOMs are used according to the kind of application. However, in this case, testing does not only involve the matcher/classifier, but it is carried out on a complete system that will be equipped with its own acquisition sensors. Therefore, contrarily to the use of a pre-collected dataset, each system may receive data with different quality and different characteristics. Test results could be repeatable, given that the reference scenario is completely and carefully controlled. In scenario evaluation, it can be possible to predict end-to-end throughput rates. These show the number of users that can be processed per unit time. Throughput rates can be critical in real-world applications, especially when large scale operations are entailed. They are not only based on computational speed. They can also be influenced by the quality of the equipment, e.g., a poor sensor may cause a sequence of failures to acquire that possibly call for new captures. Therefore, scenario evaluation can include hardware and software tests of the biometric acquisition device(s) in order to assess the quality and fidelity of the acquired samples in operational conditions that are as similar as possible to real ones. In order to get a lower bound, the test should be also performed in the most adverse yet realistic conditions that can be anticipated for the actual system deployment. A further important factor is represented by the quality (ease, effectiveness) of the provided human-machine interaction/interface, possibly guiding the user in the process of autonomously acquiring good quality samples. The time required by the latter can be very difficult to estimate since users' actions may follow different flows than the optimal ones, especially in unattended conditions. A scenario evaluation should therefore also include usability and deployment ergonomics tests to assess the ease of use of the system in particular operating conditions. Ease of use and learnability of the biometric system should be evaluated both for the typical population of target users, and, if entailed by the system use cases, for users that can occasionally access the system though not belonging to such group. Notwithstanding the increasing functionalities offered by biometric systems, it is still true that *"user behavior plays a part in many security failures, and it has become*

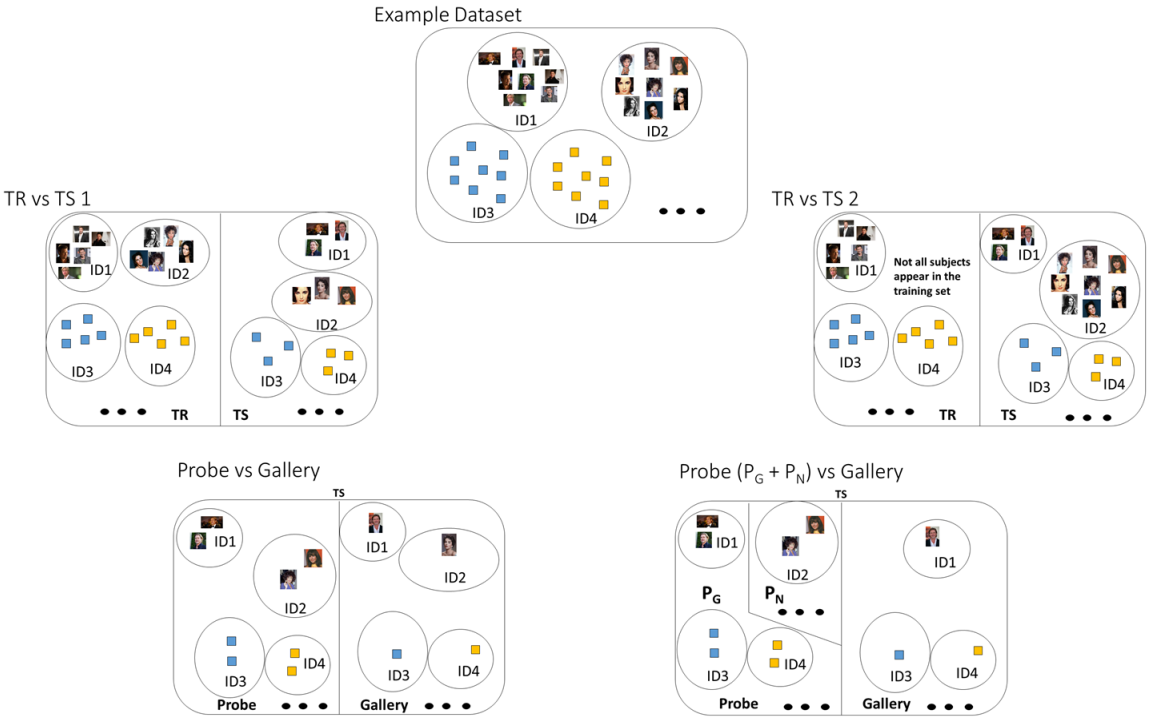

Fig. 4 Given a dataset, when training is required the subset TR may or may not contain all the subjects of the testing subset TS; the probe set used during testing does not necessarily contain all the subjects in the system gallery (P_G vs. P_N).

*common to refer to users as the 'weakest link in the security chain'. We argue that simply blaming users will not lead to more effective security systems. "*Sasse et al (2001) Sensor device efficiency and usability especially affect two error measures: Failure-to-Enroll (FTE) and Failure-To-Acquire (FTA), both including the percentage of failed attempts from the target population to successfully acquire a biometric sample of sufficient quality, together with a possible difficulty by the system to extract a good template. Of course, the criteria to assess the quality of both samples and templates depend on the biometric modality.

Operational evaluation

An end-to-end operational evaluation is similar to a scenario evaluation, also regarding the kind of FOMs that can be measured. The difference is that any kind of test is usually carried out at the actual site using either actual end-users, possibly divided into different categories, or a set of subjects that is sufficiently representative of actual end-users. For instance, the goal of an operational evaluation can be to determine how a biometric system can affect the efficiency of a traditional workflow. It will be seldom possible to repeat operational test results. Furthermore, in general, it is not possible to rely on any kind of ground-truth, especially in unsupervised/unattended conditions where there is no operator or observer controlling the operations to either assist the users or possibly limit/correct/ascertain system errors. End-to-end throughput rates can be estimated during real operations.

Open problems and Future directions

Protocols and guidelines for the evaluation of different performance aspects of biometric systems are rapidly evolving together with the increasing number of traits used as biometrics. Moreover, the use of biometric recognition is expanding also in contexts different from typical authentication scenarios, e.g., in ambient intelligence and smart cities, or using mobile devices. It is interesting to underline the increasing interest in evaluating the influence of several different environmental factors on biometric performance. In particular, a suitably set of scenario evaluations can be carried out to obtain an environmental evaluation Fernández Saavedra et al (2010). In each scenario evaluation, the influence of one specific environmental parameter is analysed. Different biometric traits and different capture technologies entail specific testing and different environmental factors can influence biometric performance ISO (2007); however, it is still an open problem to determine the necessary aspects allowing to perform repeatable and reproducible evaluations.

Strangely enough, a problem that has not been completely solved by the biometric community is the collection of datasets able to really simulate a complete cross-demographic population. The different performance achieved by biometric systems with different gender and ethnicity is an open problem, especially for biometric matchers exploiting a training phase. When

the dataset used as a benchmark, and therefore its training subset, does not contain a sufficient number of samples for all possible groups of subjects that can access the system, the recognition performance over poorly represented groups can dramatically fall during real operation. A large scale and demographically/geographically wide collection of well-balanced datasets is feasible nowadays, but demanding in terms of time and resources. Therefore, this is still a crucial open-problem to tackle.

References

(2006) ISO/IEC 19795-1, Information Technology - Biometric Performance Testing and Reporting - Part 1: Principles and framework

(2007) ISO/IEC TR 19795-3, Information Technology - Biometric Performance Testing and Reporting - Part 3: Modality specific testing.

Fernández Saavedra B, Sánchez Reíllo R, Alonso Moreno R, Miguel Hurtado Ó (2010) Environmental testing methodology in biometrics. In: 1st International Biometric Performance Conference (IBPC 2010), National Institute of Standards and Technology (NIST)

Jain AK, Ross A, Prabhakar S (2004) An introduction to biometric recognition. IEEE Transactions on circuits and systems for video technology 14(1):4–20

Sasse MA, Brostoff S, Weirich D (2001) Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. BT technology journal 19(3):122–131

Torralba A, Efros AA (2011) Unbiased look at dataset bias. In: CVPR 2011, IEEE, pp 1521–1528

Turk M, Pentland A (1991) Face recognition using eigenfaces. In: Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition, pp 586–587