Evaluation Methodologies for Biometric Presentation Attack Detection

Ivana Chingovska, Amir Mohammadi, André Anjos and Sébastien Marcel

Abstract Presentation attack detection (PAD, also known as anti-spoofing) systems, regardless of the technique, biometric mode or degree of independence of external equipment, are most commonly treated as binary classification systems. The two classes that they differentiate are bona-fide and presentation attack samples. From this perspective, their evaluation is equivalent to the established evaluation standards for the binary classification systems. However, PAD systems are designed to operate in conjunction with recognition systems and as such can affect their performance. From the point of view of a recognition system, the presentation attacks are a separate class that they need to be detected and rejected. As the problem of presentation attack detection grows to this pseudo-ternary status, the evaluation methodologies for the recognition systems need to be revised and updated. Consequentially, the database requirements for presentation attack databases become more specific. The focus of this chapter is the task of biometric verification and its scope is three-fold: firstly, it gives the definition of the presentation attack detection problem from the two perspectives. Secondly, it states the database requirements for a fair and unbiased evaluation. Finally, it gives an overview of the existing evaluation techniques for presentation attacks detection systems and verification systems under presentation attacks.

Ivana Chingovska
Idiap Research Institute, rue Marconi 19, 1920 Martigny, Switzerland e-mail: ivana.cingovska@gmail.com

Amir Mohammadi
Idiap Research Institute, rue Marconi 19, 1920 Martigny, Switzerland e-mail: amir.mohammadi@idiap.ch

André Anjos
Idiap Research Institute, rue Marconi 19, 1920 Martigny, Switzerland e-mail: andre.anjos@idiap.ch

Sébastien Marcel
Idiap Research Institute, rue Marconi 19, 1920 Martigny, Switzerland e-mail: marcel@idiap.ch

1 Introduction

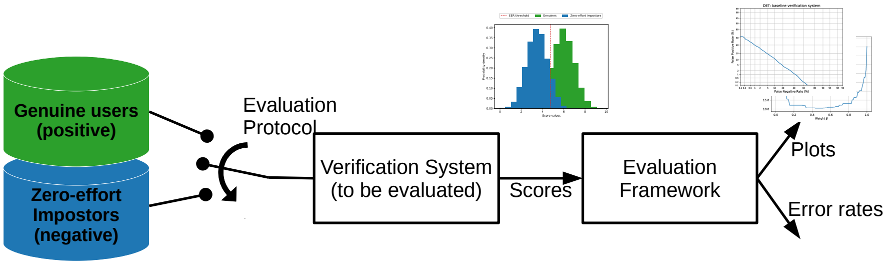


Fig. 1 Evaluation of a (unknown) verification system with regards to its capacity to discriminate genuine samples (positives) from zero-effort impostor samples (negatives). Data from each of the classes are fed into the verification system (treated as a black box) and the scores are collected. Collected scores are fed into an evaluation framework which can compute error rates and draw performance figures.

Biometric person recognition systems are widely adopted nowadays. These systems compare presentations of biometric traits to verify or identify a person. In the typical verification scenario, a biometric system matches a biometric presentation of a claimed identity against a pre-stored reference model of the same identity. The verification problem can be seen as a binary classification problem where presentations that are being matched against the same reference identity are considered positive samples (genuine samples) and the presentations that are being matched against another identity are considered negative samples (zero-effort impostor samples). Evaluation of verification systems as a binary classification problem is done using common metrics (error rates) and plots that are designed for binary classification problems. Figure 1 outlines such an evaluation framework.

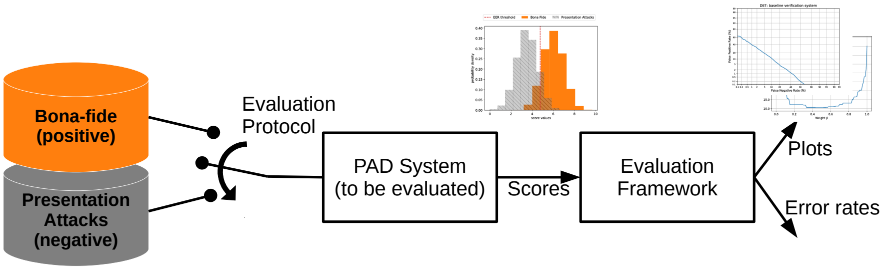


Fig. 2 Evaluation of a (unknown) PAD system with regards to its capacity to discriminate bona-fide samples (positives) from presentation attacks (negatives).

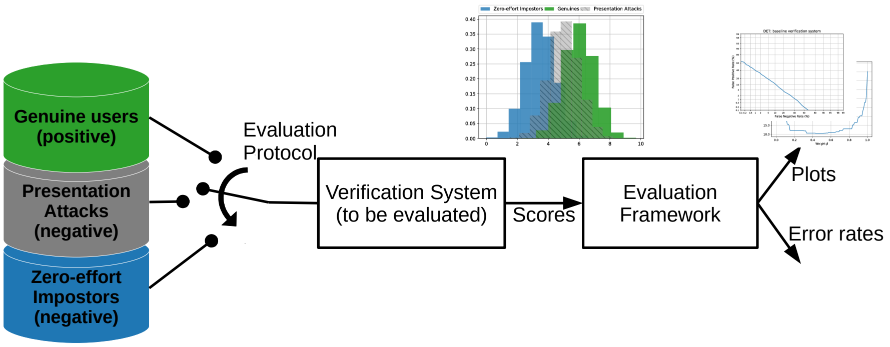


Fig. 3 Evaluation of a (unknown) verification system with regards to its capacity to discriminate genuine accesses from zero-effort impostors *and* presentation attacks.

Moreover, biometric systems are vulnerable to presentation attacks (PA, also known as spoofing). A printed photo of a person presented to a face recognition system with the goal of interfering with the operation of the system is an example of presentation attacks [1]. Presentation attack detection (PAD, also known as anti-spoofing) systems discriminate between bona-fide[1] (positives) and presentation attacks (negatives). The problem of PAs and PAD can be seen from different perspectives. As implied directly by the definition of the task of PAD systems, the problem is most often designed as a binary classification problem as outlined in Figure 2.

On the other hand, presentation attacks are directed towards deceiving recognition systems[2], regardless of whether there is a PAD algorithm to prevent them to do so, or not. From that perspective, the problem of PAs and PAD is not limited only to binary classification systems, as the isolated PAD systems are. It is of equal importance to transfer the problem understanding to the domain of biometric recognition systems (in particular, in this chapter, biometric verification systems).

Biometric verification under presentation attacks can be cast into a pseudo-ternary classification problem. While as binary classifiers, verification systems comply to typical evaluation methods, in this new perspective their concept and evaluation need to be changed accordingly. Figure 3 depicts these new settings. Instead of inputting a single set of negative examples, this new evaluation method requires two sub-classes of negative samples: samples coming from zero-effort impostors and the ones coming from presentation attacks.

This concept shift may influence the biometric verification systems at several levels. First of all, presentation attacks represent another class of input samples for

[1] Bona-fide are also called real or live samples. Both genuine and zero-effort impostor samples are bona-fide samples. While zero-effort impostors are negative samples in a verification system, they are considered positive samples in a standalone PAD system (since they are not PAs).

[2] In this chapter, since we focus on the biometric recognition task, we will only consider PAs aiming to impersonate an identity and not to conceal (hide) an identity.

the verification systems, which may cause changes in their internal algorithms to gain greater spoofing resilience. Two most prominent attempts for such changes are multimodal fusion [41, 24, 4, 3, 5] and fusion of a verification system with a PAD system [49, 30, 29, 12]. Secondly, the problem restatement needs to modify the evaluation methods for verification systems. Finally, it may play a key role in the process of their parameterization.

While the first aspect of Presentation Attack (PA) and PAD problem under the umbrella of a verification system is out of the scope of this chapter, we will thoroughly inspect all the modifications that the evaluation methods need to undergo to accustom to the new setting. The main reason is that once the danger of presentation attacks is acknowledged, the verification performance of the biometric systems is not the only measurement of their quality. Important property to assess is their robustness to presentation attacks. Only in that case one can say that the overall performance of the system is being estimated. In this context, by verification system we could consider any type of system that can produce verification scores given a biometric sample as an input. No assumption on the mechanism the system employs for protection against presentation attacks, if any, is needed. The system may be solely any baseline biometric verification algorithm which disregards the hazard of presentation attacks, or a multi-modal system or a fusion with a PAD algorithm. In any case, the system can be regarded as a black box, and the full evaluation can be done based on the verification scores it outputs for the input samples.

Mutual comparison of verification systems is the second matter of their evaluation with regards to presentation attacks. For example, it is of great importance to observe the performance change of a verification system before and after an integration with a PAD system. Blending in PAD into an existing verification system can increase its robustness to presentation attacks, but at the same time it can affect its verification performance. The evaluation methodology which is going to be deployed should be able to assess the trade-off between these two effects.

Issues regarding the aspect of parameterization and tuning of the verification systems when presentation attacks have a non-negligible prior will be also touched upon in this chapter.

With the previous observations in mind, stating the problem of PAs from the perspective of a PAD system, as well as from the perspective of a verification system is the primary objective of this chapter (Section 2). Thorough review of the evaluation strategies for isolated presentation attack detection systems, as well as for verification systems commonly used in the literature will follow in Section 4. As a prerequisite, the concepts we are going to evaluate entail certain database structure that will be covered in Section 3.

2 Problem statement

When treating PAD as a binary classification problem, designers are interested in determining the capacity of a given system to discriminate between bona-fide (positives) and presentation attacks (negatives)[3]. These systems, which do not have any capacity to perform biometric verification, are only exposed to elements of these two classes. Figure 2 represents these settings in a block diagram. In order to evaluate a given system, one feeds data from each of the two classes involved on the assessment. Scores collected from the evaluated system are fed into an evaluation framework which can compute error rates or draw performance figures. This workflow, typical for evaluation of binary classification systems, is widely deployed by PAD developers as well [34, 7, 53, 43, 16, 52, 49]. The database design and the evaluation of PAD systems comprise to the standards of general binary classification systems and will be revisited in Section 3.1 and Section 4.2, respectively.

A less considered perspective is how biometric verification systems treat presentation attacks. The classical approach puts biometric verification systems into the set of binary classifiers. Normally, such systems are designed to decide between two categories of verification attempts: bona-fide genuine users (positives) and the so-called bona-fide zero-effort impostors (negatives) [28]. Presentation attacks represent a new type of samples that can be presented at the input of this system. Considering that both presentation attacks and zero-effort impostors need to be rejected, it is still possible to regard the problem as a binary classification task where the genuine users are the positives, while the union of presentation attacks and zero-effort impostors are the negatives. Nevertheless, tuning of different properties of the verification system to make it more robust to presentation attacks may require a clearly separated class of presentation attacks. Furthermore, the correct ratio of presentation attacks and impostors in the negative class union is, at most times, unknown at design time. Applications in highly-surveilled environments may consider that the probability of a presentation attack is small, while applications in unsurveilled spaces may consider it very high. Presentation attacks, therefore, should be considered as a third separate category of samples that the verification systems need to handle.

This viewpoint, casts biometric verification into a pseudo-ternary classification problem as depicted in Figure 3. Researchers generally simplify the pseudo-ternary classification problem so that it suits the binary nature of the verification systems. A common approach is to reduce it to two binary classification problems, each of which is responsible for one of the two classes of negatives. According to this, the verification system can be operating in two scenarios or operation modes: (1) when it receives genuine accesses as positives and only zero-effort impostors as negatives, and (2) when it receives only genuine accesses as positives and presentation attacks as negatives. Sometimes the first scenario is called a *normal operation mode* [19, 42, 18]. As it is going to be discussed in Section 4.3, it is beneficial to simplification that the positives (genuine accesses) that are evaluated completely match in both scenarios.

The workflow of the verification system confronted with presentation attacks, from the input to the evaluation stage, is represented in Figure 3. The score his-

[3] In this chapter, we shall treat examples in a (discriminative) binary classification system one wishes to keep as *positive class* or simply as *positives*, and, examples that should be discarded as *negative class* or *negatives*.

togram displays 3 distinctive groups of data: the positive class and the two negative ones. If the mixing factor between the negative classes is known at design time, system evaluation can be carried using known binary classification analysis tools. Since that is usually not the case, the evaluation tools for the verification systems need to be adapted to the new settings.

The new concept for verification systems explained above requires a database design and evaluation methodologies adapted to the enhanced negative class, regardless of the system's robustness to presentation attacks and how it is achieved. An overview of the research efforts in this domain will be given in Section 3.2 and Section 4.3, respectively.

3 Database requirements

The use of databases and associated evaluation protocols allow for objective and comparative performance evaluation of different systems. As discussed on Section 2, the vulnerability (aka *spoofability*) of a system can be evaluated on isolated presentation attack detection systems, but also on fully functional verification systems. The simple evaluation of PAD requires only that database and evaluation protocols consider two data types: bona-fide and presentation attack samples. The evaluation of verification systems, merged with PAD or not, requires the traceability of identities contained in each presented sample, so that tabs are kept for probe-to-model matching and non-matching scenarios. The particular requirements for each of the two cases are given in Sections 3.1 and 3.2. Databases for each of these two settings exist in literature. An exhaustive listing of databases that allow for the evaluation of resilience against presentation attacks in isolated PAD or biometric verification systems is given by the end of this section, in Section 3.3.

3.1 Databases for evaluation of presentation attack detection systems

The primary task of a database for evaluation of presentation attack detection systems is to provide samples of presentation attacks along with samples of bona-fide. The identity information of clients in each sample needs not to be present and can be discarded in case it is. The two sets of samples, which will represent the negative and the positive class for the binary classification problem, are just by themselves sufficient to train and evaluate a PAD system. It is a common practice that a database for binary classification provides a usage protocol which breaks the available data into 3 datasets [22]:

- *Training set* $\mathscr{D}_{train}$, used to train a PAD model;

- *Development set* $\mathscr{D}_{dev}$, also known as the validation set, used to optimize the decisions in terms of model parameters estimation or model selection;
- *Test set* $\mathscr{D}_{test}$, also known as the evaluation set, on which the performance is finally measured.

In the case of presentation attack databases, it is recommended that the 3 datasets do not contain overlapping client data to avoid bias related to client specific traits and to improve generalization [27]. A database with this setup completely satisfies the requirements of a two-class classification problem, as the isolated presentation attack detection is.

The process of generating presentation attacks requires bona-fide samples that will serve as a basis to create the fake copies of the biometric trait. These may or may not be the same samples as the bona-fide samples of the database. In any case, if they are provided alongside the database, it can be enhanced with new types of presentation attacks in future.

3.2 Databases for vulnerability analysis of verification systems

If a database is to serve for evaluation of a verification system, it needs to possess similar properties of a biometric database. Training and testing through biometric databases require (preferably) disjoint sets of data used for enrollment and verification of different identities. In practice, many databases also present a separation of the data in three sets as described above. Data from the training set can be used to create background models. The development data should contain enrollment (gallery) samples to create the user-specific models, as well as probe samples to match against the models. Similar specifications apply for the test set. The matching of the development probe samples against the user models should be employed to tune algorithms' parameters. Evaluation is carried out by matching probe samples of the test set against models created using the enrollment samples. The identity of the model being tested and the gallery samples are annotated to each of the scores produced so that the problem can be analyzed as a binary classification one: if model identity and probe identity match, the score belongs to the positive class (genuine client), otherwise, the score belongs to the negative class (zero-effort impostors). Usually, all identities in the three datasets are kept disjoint for the same reasons indicated in Section 3.1. Following this reasoning, a first requirement for a presentation attack database aspiring to be equally adapted to the needs of PAD and verification systems, is provision of separate enrollment samples, besides the bona-fide and presentation attack samples.

The pseudo-ternary problem of presentation attacks as explained in Section 2 imposes scenario for matching bona-fide genuine accesses, bona-fide zero-effort impostors, and presentation attacks against the models. In order to conform to this second requirement, the simplification of the pseudo-ternary problem introduced in Section 2 is of great help. In the case of the first scenario, or the normal operation mode, matching entries equivalent to the entries for genuine users and zero-effort impostors for a classical biometric verification database are needed. In the case of the second scenario, the provided entries should match the presentation attack samples to a corresponding model or enrollment sample.

To unify the terminology, we formalize the two scenarios of operation of the verification system as below:

- *Licit* scenario: A scenario consisting of genuine users (positives) and zero-effort impostors (negatives). The positives of this scenario are created by matching the genuine access samples of each client to the model or enrollment samples of the same client. The negatives can be created by matching the genuine access samples of each client to the model or enrollment samples of other clients. This scenario is suitable to evaluate a verification system in a normal operation mode. Evidently, no presentation attacks are present in this scenario;
- *Spoof* scenario: A scenario consisting of genuine users (positives) and presentation attacks (negatives). The positives of this scenario are created by matching genuine access samples of each client to the models or enrollment samples of the same client. The negatives are created by matching the presentation attacks of each client to the model or enrollment samples of the same client. No zero-effort impostors are involved in this scenario.

The licit scenario is necessary for evaluation of the verification performance of the system. The spoof scenario is necessary for evaluation of the system's robustness to PAs. If we follow a convention to match *all* the genuine access samples to the model or enrollment samples of the same client in both scenarios, we will end up having the same set of positives for the two scenarios. This agreement, as will be shown in Section 4.3, plays an important role in some approaches for evaluation of the verification systems.

To better illustrate how to create the scenarios out of the samples present in any presentation attack database, let us assume a simple hypothetical presentation attack database containing one bona-fide and one presentation attack of two clients with identities A and B. Let us assume that the database also contains bona-fide enrollment samples for A and B allowing computation of models for them. The matching of the samples with the models in order to create the positives and the negatives of the two scenarios is given in Table 1. To exemplify an entry in the table, L+ in the first row means that entries that match genuine accesses of client A to the model of client A belong to the subset of positives of the licit scenario. The same applies for L+ in the third row, this time for client B. Similarly, S- in the second row means that entries that match presentation attacks of client A to the model of client A belong to the subset of negatives in the spoof scenario.

Instead of creating a presentation attack database and then creating the licit and spoof scenario from its samples, an alternative way to start with is to use an existing biometric database which already has enrollment samples as well as data for the licit scenario. All that is needed is creating the desirable presentation attacks out of the existing samples. One should note however, that the *complete* system used for the acquisition of samples, including the sensor, should be kept constant through all the recordings as differentiation may introduce biases. For example, consider

Table 1 Creating licit and spoof scenarios out of the samples in a PA database. + stands for positives, - for negatives. L is for licit and S for spoof scenario. Note that the positives are the same for both L and S scenarios. Bona-fide enrollment samples will also be needed for each identity.

Probe	Model for	A	B
A	**bona-fide**	L+, S+	L-
	presentation attack	S-	no match done
B	**bona-fide**	L-	L+, S+
	presentation attack	no match done	S-

a situation in which a speaker verification system is evaluated with data collected with a low-noise microphone, but in which attack samples are collected using noisier equipment. Even if attacks do pass the verification threshold, it is possible that potential PAD may rely on the additional noise produced by the new microphone to identify attacks. If that is the case, then such a study may be producing a less effective PAD system.

3.3 Overview of available databases for presentation attack detection

Table 2 contains an overview of the existing PAD databases that are publicly available. The columns, that refer to properties discussed throughout this section, refer to:

- **Database**: the database name;
- **Trait**: the biometric trait on the database;
- **# Subsets**: the number of subsets in the database referring to existing separate set for training, developing and testing systems;
- **Overlap**: if there is client overlap between the different database subsets (training, development and testing);
- **Vulnerability**: if the database can be used to evaluate the vulnerability of a verification system to presentation attacks (i.e. contains enrollment samples);
- **Existing DB**: if the database is a spin-off of an existing biometric database not originally created for PAD evaluation;
- **Sensor**: If the sensors used to acquire the presentation attack samples are the same as those used for the bona-fide samples.

Table 2 Catalog of evaluation features available on a few presentation attack databases available. For detailed column description, please see section 3.3. This table is not an exhaustive list of presentation attack databases.

Database	Trait	# Subsets	Overlap	Vulnerability	Existing DB	Sensor
ATVS-FFp[4] [18]	Fingerprint	2	No	No	No	Yes
LivDet 2009 [31]	Fingerprint	2	?	No	No	Yes
LivDet 2011 [52]	Fingerprint	2	?	No	No	Yes
LivDet 2013[5] [21]	Fingerprint	2	?	No	No	Yes
CASIA FAS[6] [55]	Face	2	No	No	No	Yes
MSU MFSD[7] [51]	Face	2	No	No	No	Yes
NUAA PI[8] [43]	Face	2	No	No	No	Yes
OULU-NPU[9] [10]	Face	2	No	No	No	Yes
Replay Attack[10] [11]	Face	3	No	Yes	No	Yes
Replay Mobile[11] [14]	Face	3	No	Yes	No	Yes
UVAD[12] [37]	Face	2	No	No	No	Yes
Yale Recaptured[13] [36]	Face	1	Yes	No	Yes	No
VERA Fingervein[14] [48, 46, 45]	Finger-vein	2	No	Yes	No	Yes
VERA Palmvein[15] [44]	Palm-vein	3	No	Yes	Yes	Yes
ASVSpoof 2017[16] [25]	Voice	2	No	Yes	Yes	No
AVSpoof[17] [15]	Voice	3	No	Yes	No	Yes
VoicePA[18] [26]	Voice	3	No	Yes	Yes	Yes

4 Evaluation techniques

Several important concepts about evaluation of binary classification systems have been established and followed by the biometric community. Primarily, they are used to evaluate verification systems, which have a binary nature. They are also applicable in the problem of PAD as a binary classification problem.

In Section 4.1 we revisit the basic notation and statistics for evaluation of any binary classification system. After that recapitulation, we give an overview of how

[4] `http://atvs.ii.uam.es/atvs/ffp_db.html`

[5] `http://livdet.org/`

[6] `http://www.cbsr.ia.ac.cn/english/Databases.asp`

[7] `http://www.cse.msu.edu/rgroups/biometrics/Publications/Databases/MSUMobileFaceSpoofing/`

[8] `http://parnec.nuaa.edu.cn/xtan/data/nuaaimposterdb.html`

[9] `https://sites.google.com/site/oulunpudatabase/`

[10] `http://www.idiap.ch/dataset/replayattack`

[11] `http://www.idiap.ch/dataset/replaymobile`

[12] `http://ieeexplore.ieee.org/abstract/document/7017526/`

[13] `http://ieeexplore.ieee.org/abstract/document/6116484/`

[14] `https://www.idiap.ch/dataset/vera-fingervein`

[15] `https://www.idiap.ch/dataset/vera-palmvein`

[16] `http://dx.doi.org/10.7488/ds/2313`

[17] `https://www.idiap.ch/dataset/avspoof`

[18] `https://www.idiap.ch/dataset/voicepa`

the error rates and methodologies are adapted particularly for PAD systems in Section 4.2 and verification systems under presentation attacks in Section 4.3.

4.1 Evaluation of binary classification systems

The metrics for evaluation of binary classification systems are associated to the types of errors and how to measure them, as well as to the threshold and evaluation criterion [39]. A binary classification system is subject to two types of errors: False Positive (FP) and False Negative (FN). Typically, the error rates that are reported are False Positive Rate (FPR), which corresponds to the ratio between FP and the total number of negative samples and False Negative Rate (FNR), which corresponds to the ratio between FN and the total number of positive samples.

Alternatively, many algorithms for binary classification report different error rates, but still equivalent to FPR and FNR. For example, True Positive Rate (TPR) refers to the ratio of correctly classified positives and can be computed as $1 - \mathrm{FNR}$. True Negative Rate (TNR) gives the ratio of correctly detected negatives, and can be computed as $1 - \mathrm{FPR}$.

To compute the error rates, the system needs to compute a decision threshold τ which will serve as a boundary between the output scores of the genuine accesses and presentation attacks. By changing this threshold one can balance between FPR and FNR: increasing FPR reduces FNR and vice-versa. However, it is often desired that an optimal threshold τ^* is chosen according to some criterion. Two well established criteria are Minimum Weighted Error Rate (WER) and Equal Error Rate (EER) [39]. In the first case, the threshold τ^*_{WER} is chosen so that it minimizes the weighted total error rate as in Eq. 1 where $\beta \in [0,1]$ is a predefined parameter which balances between the importance (cost) of FPR and FNR. Very often, they have the same cost of $\beta = 0.5$, leading to Minimum Half Total Error Rate (HTER) criteria. In the second case, the threshold τ^*_{EER} ensures that the difference between FPR and FNR is as small as possible (Eq. 2). The optimal threshold, also referred to as *operating point* should be determined using the data in the development set, denoted in the equations below as $\mathscr{D}_{dev}$.

$$\tau^*_{\mathrm{WER}} = \arg\min_{\tau} \beta \cdot \mathrm{FPR}(\tau, \mathscr{D}_{dev}) + (1-\beta) \cdot \mathrm{FNR}(\tau, \mathscr{D}_{dev}) \quad (1)$$

$$\tau^*_{\mathrm{EER}} = \arg\min_{\tau} |\mathrm{FPR}(\tau, \mathscr{D}_{dev}) - \mathrm{FNR}(\tau, \mathscr{D}_{dev})| \quad (2)$$

Regarding the evaluation criteria, once the threshold τ^* is determined, the systems usually report the WER (Eq. 3) or its special case for $\beta = 0.5$, HTER (Eq. 4). Since in a real world scenario the final system will be used for data which have not been seen before, the performance measure should be reported on the test set $\mathscr{D}_{test}$.

$$\mathrm{WER}(\tau, \mathscr{D}_{test}) = \beta \cdot \mathrm{FPR}(\tau, \mathscr{D}_{test}) + (1-\beta) \cdot \mathrm{FNR}(\tau, \mathscr{D}_{test}) \quad (3)$$

$$\mathrm{HTER}(\tau, \mathscr{D}_{test}) = \frac{\mathrm{FPR}(\tau, \mathscr{D}_{test}) + \mathrm{FNR}(\tau, \mathscr{D}_{test})}{2} \quad [\%] \tag{4}$$

Graphical analysis

Important tools in evaluation of classification systems are the different graphical representations of the classification results. For example, one can get an intuition about how good the discriminating power of a binary classification system is by plotting its output score distributions for the positive and the negative class, as in Figure 4(a). Better separability between the two classes means better results in terms of error rates.

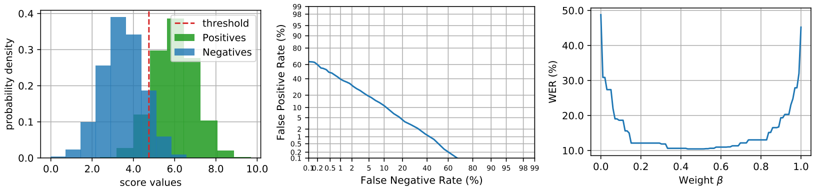
To summarize the performance of a system and to present the trade-off between FPR and FNR depending on the threshold, the performance of the binary classification systems are often visualized using Receiver Operating Characteristic (ROC) and Detection-Error Tradeoff (DET) [32] curves. They plot the FPR versus the FNR (or some of the equivalent error rates) for different values of the threshold. Sometimes, when one number is needed to represent the performance of the system in order to compare several systems, Area Under ROC curve (AUC) values are reported. Usually it is computed for ROC curves plotting FPR and TPR and in this case, the higher the AUC the better the system. Figure 4(b) illustrates the DET curve for a hypothetical binary classification system.

Unfortunately, curves like ROC and DET can only display *a posteriori* performance. When reading values directly from the plotted curves, one implicitly chooses a threshold on a dataset and the error rates are reported on the same dataset. Although ROC and DET give a clear idea about the performance of a single system, as explained in [8], comparing two systems with these curves can lead to biased conclusions. To solve this issue, [8] proposes the so-called Expected Performance Curve (EPC). It fills in for two main disadvantages of the DET and ROC curves: 1. it plots the error rate on an independent test set based on a threshold selected *a-priori* on a development set; and 2. it accounts for the varying relative cost $\beta \in [0;1]$ of FPR and FNR when calculating the threshold.

Hence, in the EPC framework, an optimal threshold τ^* is computed using Eq. 1 for different values of β, which is the variable parameter plotted on the abscissa. Performance for the calculated values of τ^* is then computed on the test set. WER, HTER or any other measure of importance can be plotted on the ordinate axis. The EPC curve is illustrated in Figure 4(c) for a hypothetical classification system.

4.2 Evaluation of presentation attack detection systems

In the domain of PAD, bona-fide samples are the positive samples and presentation attacks are negative. Moreover, False Positive Rate (FPR) was renamed by the ISO



(a) Score distributions (b) DET curve (c) EPC curve

Fig. 4 Evaluation plots for a hypothetical binary classification system

standards [1] to Attack Presentation Classification Error Rate (APCER), and False Negative Rate (FNR) was renamed to Bona Fide Presentation Classification Error Rate (BPCER). Before the ISO standardization, since the positives and the negatives are associated with the action of *acceptance* and *rejection* by the PAD system, False Accept Rate (FAR) and False Reject Rate (FRR) were used commonly in place of APCER and BPCER, respectively. Some publications utilize other synonyms which are listed in Table 3.

Table 3 Typically used error rates in PAD and their synonyms.

Error rate	Acronym	Synonyms
False Positive Rate	FPR	Attack Presentation Classification Error Rate (APCER), False Accept Rate (FAR), False Spoof Accept Rate [29], False Living Rate (FLR) [16]
False Negative Rate	FNR	Bona-fide Presentation Classification Error Rate (BPCER), False Reject Rate (FRR), False Alarm Rate [34], False Live Rejection Rate [29], False Fake Rate (FFR) [16]
True Positive Rate	TPR	True Accept Rate
True Negative Rate	TNR	True Reject Rate, detection rate [34, 7, 50] , detection accuracy [54]
Half Total Error Rate	HTER	Average Classification Error (ACE) [16]

For a more general framework, where the system is specialized to detect any kind of suspicious or subversive presentation of samples, be it a presentation attack, altered sample or artifact, [38] has assembled a different set of notations for error measurements. Such a system reports False Suspicious Presentation Detection (FSPD) in the place of FNR and False Non-Suspicious Presentation Detection (FNSPD) in the place of FPR. To summarize the error rates into one value, some authors use accuracy [36, 20, 53], which is the ratio of the overall errors that the system made and the total number of samples. Finally, to graphically represent the performance of the PAD systems, score distribution plots [47], ROC, DET and EPC curves are often used.

4.3 Evaluation of verification systems under presentation attacks

The classical approach regards a biometric verification system as a binary classification system. In the scope of biometric verification systems, False Match Rate (FMR) and False Non-Match Rate (FNMR) are the most commonly used terms for the error rates FPR and FNR. FMR stands for the ratio of incorrectly accepted zero-effort impostors and FNMR for the ratio of incorrectly rejected genuine users. These and the equivalent error rates are often substituted with other synonyms which are different by different authors. The most common of them are listed in Table 4. Although not always equivalent [28], sometimes FMR and FNMR are substituted with FAR and FRR, respectively [23].

The simplification of ternary classification into two binary classification problems, as explained in Section 2, is the key step that set the standards for the evaluation of verification systems. Systems are usually evaluated separately in the two modes of operation associated with the two scenarios stated in Section 3.2. This section focuses on the error rates and plots typical for this evaluation.

While verification performance metrics are well established and widely used, metrics for PA evaluation is not as well defined and adopted. Some authors do not make a clear distinction between a presentation attack and a zero-effort impostor and refer to both types of samples as impostors. The nature of the sample can be concluded by the scenario in which it is being used: licit or spoof.

The importance of a clear distinction between the terminology for error rate reporting on misclassified zero-effort impostors and presentation attacks was outlined in [2]. Besides Liveness False Acceptance Rate (LFAR) as a ratio of presentation attacks that are incorrectly accepted by the system, [2] defines error rates connected to the total number of accepted negatives, regardless of whether they come from zero-effort impostors or presentation attacks. For example, the union of FAR in licit scenario and LFAR in spoof scenario is called System False Acceptance Rate (SFAR). However, since the introduction of [1], LFAR (also sometimes called Spoof False Accept Rate (SFAR)) was renamed to Impostor Attack Presentation Match Rate (IAPMR). A detailed overview of all the metrics utilized by various authors is given in Table 4. The table contains two metrics of error rates for negatives: for the licit and spoof scenario. It also reports the overall error rates that occur when both scenarios are considered as a union.

The adopted terminology in the remainder of this text is as follows:

- FNMR - ratio of incorrectly rejected genuine users (both licit and spoof scenario)
- FMR - ratio of incorrectly accepted zero-effort impostors (in the licit scenario)
- IAPMR - ratio of incorrectly accepted presentation attacks [24] (in the spoof scenario)

Table 4 Typically used error rates in biometric verification and their synonyms

Scenario	Error rate	Synonyms
Licit	False Positive Rate	False Match Rate (FMR), False Accept Rate (FAR) [17, 29], Pfa [49]
Spoof	False Positive Rate	Impostor Attack Presentation Match Rate (IAPMR), False Accept Rate (FAR) [19], Spoof False Acceptance Rate [24, 5], Liveness False Acceptance Rate [2], Success Rate [18, 42], Attack Success Rate [17]
Both	False Negative Rate	False Non-Match Rate (FNMR), False Reject Rate (FRR) [17, 29], Pmiss [49]
Both	True Positive Rate	True Positive Rate, True Accept Rate, Genuine Acceptance Rate [38, 40]
Union	False Positive Rate	Global False Acceptance Rate (GFAR) [29], System False Acceptance Rate (SFAR) [2]
	False Negative Rate	Global False Rejection Rate (GFRR)

- GFAR - ratio of incorrectly accepted zero-effort impostors and presentation attacks.

Researchers generally follow three main methodologies for determining the effect of presentation attacks over the verification systems and obtaining the error rates. The differences between the three evaluation methodologies are in the way of computation of the decision threshold.

Evaluation methodology 1

Two decision threshold calculations are performed separately for the two scenarios, resulting in two separate values of the error rate (HTER or EER) [33, 35, 19, 24, 6]. FNMR, FMR and IAPMR are reported depending on the decision threshold obtained for the scenario they are derived from. One weak point of this type of evaluation is that it neglects that there is only one verification system at disposal and it should have only one operating point corresponding to one decision threshold. Furthermore, the decision threshold and the reported error rates of the spoof scenario are irrelevant in a real-world scenario. The problem arises because the spoof scenario assumes that all the possible misuses of the system come from spoofing attacks. It is not likely that any system needs to be tuned to operate in such a scenario. Therefore, the error rates depending on the threshold obtained under the spoof scenario are not a relevant estimate of the system's performance under presentation attacks. Furthermore, the error rates for the licit and spoof scenarios can not be compared, because they rely on different thresholds.

Evaluation methodology 2

This methodology is adopted for more realistic performance evaluation. It takes advantage of the assumption that the licit and spoof scenarios share the same positive samples: a requirement mentioned to be beneficial in Section 3.2. In this case, the system will obtain the same FNMR for the both scenarios regardless of the threshold. Once the threshold of the system is chosen, FMR and IAPMR can be reported and compared. The threshold can be chosen using various criteria, but almost always using the licit scenario. Most of the publications report error rates for the two scenarios using a threshold chosen to achieve a particular desired value of FRR [17, 18, 49, 9, 42, 41, 40, 4, 3, 5, 30].

The issue that the evaluation methodology 2 oversees is that a system whose decision threshold is optimized for one type of negatives (for example, the zero-effort impostors), can not be evaluated in a fair manner for another type of negatives (the presentation attacks). If the system is expected to be exposed to two types of negatives in the test or deployment stage, it is fair that the two types of negatives play a role in the decision of the threshold in the development stage.

Evaluation methodology 3

This methodology, introduced as Expected Performance and Spoofability (EPS) framework in [13], aims at filling in the gaps of the evaluation methodology 2 and establishes a criteria for determining a decision threshold which considers the two types of negatives and also the cost of rejecting positives. Two parameters are defined: $\omega \in [0,1]$, which denotes the relative cost of presentation attacks with respect to zero-effort impostors; and $\beta \in [0,1]$, which denotes the relative cost of the negative classes (zero-effort impostors and presentation attacks) with respect to the positive class. FAR_ω is introduced which is a weighted error rate for the two negative classes (zero-effort impostors and presentation attacks). It is calculated as in Eq. 5.

$$\mathrm{FAR}_\omega = \omega \cdot \mathrm{IAPMR} + (1-\omega) \cdot \mathrm{FMR} \tag{5}$$

The optimal classification threshold $\tau^*_{\omega,\beta}$ depends on both parameters. It is chosen to minimize the weighted difference between FAR_ω and FNMR on the development set, as in Eq. 6.

$$\tau^*_{\omega,\beta} = \arg\min_\tau |\beta \cdot \mathrm{FAR}_\omega(\tau, \mathscr{D}_{dev}) - (1-\beta) \cdot \mathrm{FNMR}(\tau, \mathscr{D}_{dev})| \tag{6}$$

Once an optimal threshold $\tau^*_{\omega,\beta}$ is calculated for certain values of ω and β, different error rates can be computed on the test set. Probably the most important is $\mathrm{WER}_{\omega,\beta}$, which can be accounted as a measurement summarizing both the verification performance and the vulnerability of the system to presentation attacks and which is calculated as in Eq. 7.

$$\begin{aligned}\mathrm{WER}_{\omega,\beta}(\tau^*_{\omega,\beta},\mathscr{D}_{test}) = \beta\cdot\mathrm{FAR}_{\omega}(\tau^*_{\omega,\beta},\mathscr{D}_{test})\\ +(1-\beta)\cdot\mathrm{FNMR}(\tau^*_{\omega,\beta},\mathscr{D}_{test})\end{aligned} \tag{7}$$

A special case of $\mathrm{WER}_{\omega,\beta}$, obtained by assigning equal cost $\beta = 0.5$ to FAR_w and FNMR can be defined as HTER_{ω} and computed as in Eq. 8. In such a case, the criteria for optimal decision threshold is analogous to the EER criteria given in Section 4.2.

$$\mathrm{HTER}_{\omega}(\tau^*_{\omega},\mathscr{D}_{test}) = \frac{\mathrm{FAR}_{\omega}(\tau^*_{\omega},\mathscr{D}_{test})+\mathrm{FNMR}(\tau^*_{\omega},\mathscr{D}_{test})}{2} \tag{8}$$

The parameter ω could be interpreted as relative cost of the error rate related to presentation attacks. Alternatively, it could be connected to the expected relative number of presentation attacks among all the negative samples presented to the system. In other words, it could be understood as the prior probability of the system being exposed to presentation attacks when it is deployed. If it is expected that there is no danger of presentation attacks for some particular setup, it can be set to 0. In this case, $\mathrm{WER}_{\omega,\beta}$ corresponds to WER in the traditional evaluation scheme for biometric verification systems. When it is expected that some portion of the illegitimate accesses to the system will be presentation attacks, ω will reflect their prior and ensure they are not neglected in the process of determining the decision threshold.

As in the computation of WER in Section 4.2, the parameter β could be interpreted as the relative cost of the error rate related to the negative class consisting of both zero-effort impostors and presentation attacks. This parameter can be controlled according to the needs or to the deployment scenario of the system. For example, if we want to reduce the wrong acceptance of samples to the minimum, while allowing increased number of rejected genuine users, we need to penalize FAR_{ω} by setting β as close as possible to 1.

Graphical Analysis

Following the typical convention for binary classification system, biometric verification systems use score distributions, ROC or DET curves to graphically present their performance. The plots for a traditional biometric verification system regard the genuine users as a positive and the zero-effort impostors as a negative class. The details about these types of plots are given in Section 4.2.

When using graphical representation of the results, the researchers usually follow the evaluation methodology 2. This means that all the tuning of the algorithms, in particular in computation of the decision thresholds, is performed using the licit scenario, while the plots may represent the results for one of the scenarios or for the both of them.

When only the licit scenario is of interest, the score distribution plot contains the distributions only for the genuine users and the zero-effort impostors. If evaluation with regards to the vulnerability to presentation attacks is desired, the score distribution plot gets an additional distribution corresponding to the scores that the system outputs for the presentation attack samples in the spoof scenario. As a result, the score distribution plot presents three score distributions, which, illustratively for a hypothetical verification system, are given in Figure 5(a).

An information about the dependence of IAPMR on the chosen threshold can be obtained directly from the score distribution plot. An example is shown in Figure 5(b), where the full red line represents how IAPMR varies with shifting the threshold, while the vertical dashed red line represents the threshold at a chosen operating point.

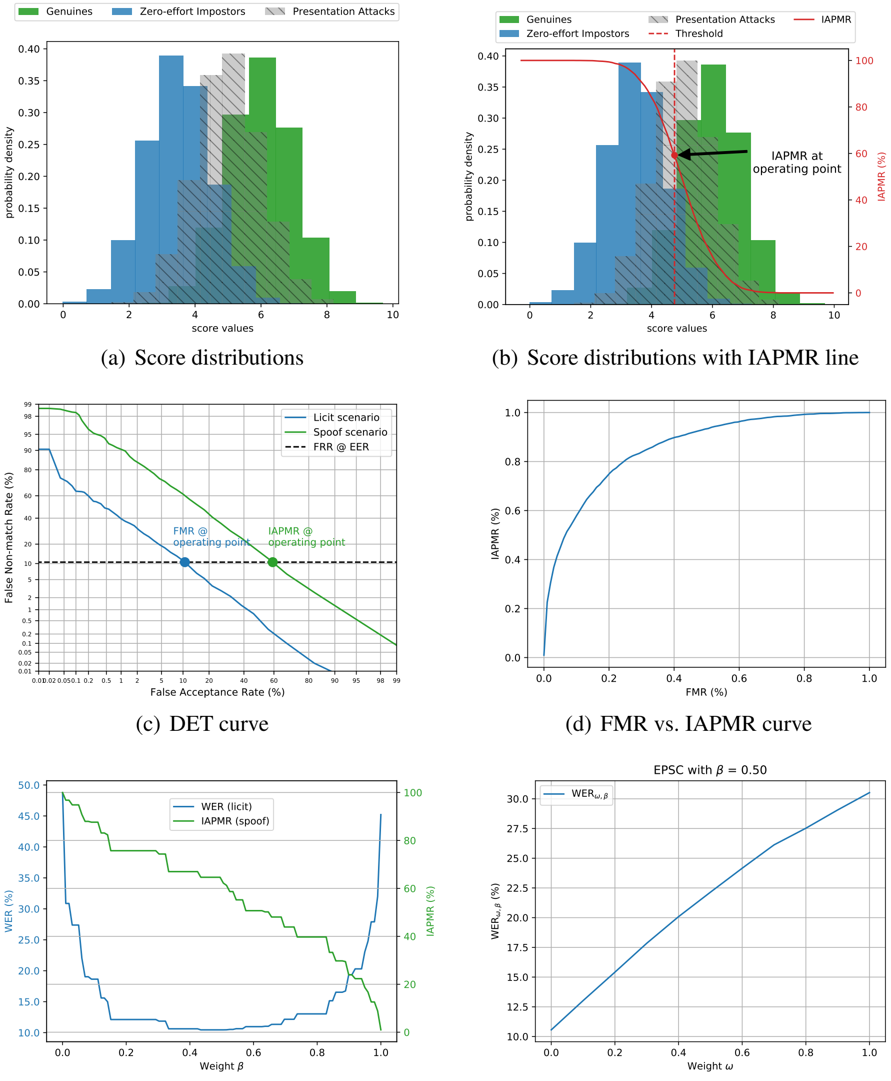
Typically, ROC and DET curves visualize the trade-off between FMR and FNMR for a biometric system with no danger of presentation attacks anticipated. The closest analogy to the ROC and DET curves when evaluating a system exposed to presentation attacks, can be found using the evaluation methodology 2. Firstly, the curve using the licit scenario is plotted. Then, it can be overlaid with a curve for the spoof scenario. For the licit scenario the horizontal axis represents FMR, while for the spoof scenario it represents IAPMR. However, meaningful comparison of the two curves is possible only if the number of genuine access samples in both licit and spoof scenario is the same. In such a case, a certain selected threshold will result in the same value of FNMR for the both scenarios. By drawing a horizontal line at the point of the obtained FNMR, one can examine the points where it cuts the curves for the licit and spoof scenario, and can compare FMR and IAPMR for the given system. Illustration of this analysis is given in Figure 5(c).

The drawback of the DET curve coming from its a-posteriori evaluation feature explained in [8] and obstructing fair comparison of two systems, is not a concern here. The plot does not compare different systems, but the same system with a single operating point under different set of negative samples.

As an alternative figure delivering similar information as DET, [41] and [40] suggest to plot FMR vs. IAPMR. Thresholds are fixed in order to obtain all the possible values of FMR for the licit scenario and IAPMR is evaluated on the spoof scenario and plotted on the ordinate axis. By plotting the curves for different verification systems, the plot enables to compare which of them is less prone to spoofing given a particular verification performance. However, this comparison suffers from the same drawback as the DET: a-posteriori evaluation. As such, its fairness is limited. This plot is illustrated in Figure 5(d).

The logic for plotting the EPC curve is similar if one wants to follow the evaluation methodology 2. One has to vary the cost parameter β which balances between FMR and FNMR of the licit scenario and choose the threshold accordingly. Using the selected threshold, one can plot WER on the licit scenario. Afterwards, to see the method's vulnerability to presentation attacks depending on β, the WER curve can be overlaid with the IAPMR curve using the spoof scenario, as shown in Figure 5(e) for a hypothetical system.

A graphical evaluation for the evaluation methodology 3 can not be easily derived from the existing ROC or DET curves. The EPS framework computes error rates for a range of decision thresholds obtained by varying the parameters ω and β. The visualization of the error rates parameterized over two parameters will result in a 3D



(a) Score distributions

(b) Score distributions with IAPMR line

(c) DET curve

(d) FMR vs. IAPMR curve

(e) EPC curve

(f) EPSC curve

Fig. 5 Performance and spoofing vulnerability evaluation plots for hypothetical verification system

surface, which may not be convenient for evaluation and analysis, especially when one needs to compare two or more systems. Instead, plotting the Expected Performance and Spoofability Curve (EPSC) is suggested, showing $\mathrm{WER}_{\omega,\beta}$ with respect to one of the parameters, while the other parameter is fixed to a predefined value. For example, we can fix the parameter $\beta = \beta_0$ and draw a 2D curve which plots $\mathrm{WER}_{\omega,\beta}$ on the ordinate with respect to the varying parameter ω on the abscissa.

Having in mind that the relative cost given to FAR_ω and FNMR depends mostly on the security preferences for the system, it is not difficult to imagine that particular values for β can be selected by an expert. Similarly, if the cost of IAPMR and FMR or the prior of presentation attacks with regards to the zero-effort impostors can be precisely estimated for a particular application, one can set $\omega = \omega_0$ and draw a 2D curve plotting $\text{WER}_{\omega,\beta}$ on the ordinate, with respect to the varying parameter β on the abscissa. Unlike for EPC, in the decision threshold calculation for EPSC both the licit and spoof scenario take place, because both FMR and IAPMR contribute with a certain weight.

The convenience of EPSC for evaluation of verification systems under presentation attacks is covered by several properties. Firstly, since it follows the evaluation methodology 3, it provides that both types of negatives participate in threshold decision process. Secondly, it presents a-priori results: the thresholds are calculated on the development set, while the error rates are reported on the test set. This ensures unbiased comparison between algorithms. Furthermore, this comparison is enabled for a range of values for the cost parameters ω and β.

Besides $\text{WEE}_{\omega,\beta}$, other error rates of interest may be plotted on the EPSC plot, like IAPMR or FAR_ω.

5 Conclusions

Presentation attack detection systems in biometrics can rarely be imagined working as stand-alone. Their task is to perform an additional check on the decision of a biometric verification systems in order to detect a fraudulent user who possesses a copy of a biometric trait of a genuine user. Unless they have perfect detection rate, they inevitably affect the performance of the verification system they protect.

Traditionally, the presentation attack detection systems have been evaluated as binary classification systems, and in reason: by nature they need to distinguish between two classes - bona-fide and presentation attack samples. However, the above observation throws a light on the critical issue of establishing a methodology for evaluation of verification systems with regards to presentation attacks. This equally applies for verification systems with or without any mechanism for handling presentation attacks.

This task requires reformulation of the problem of biometric verification. They, as well, are, by definition, binary classification systems distinguishing between genuine accesses and zero-effort impostors. With the presentation attacks in play, the problem scales to pseudo-ternary classification problem, with two types of negatives: zero-effort impostors and presentation attacks.

As a result of the above observations, this chapter covers the problem of presentation attacks evaluation from two perspectives: evaluation of presentation attack detection systems alone and evaluation of verification systems with respect to presentation attacks. The evaluation in the first case means straight-forward application of well established evaluation methodologies for binary classification systems, in error rates (FAR, FRR, HTER etc.), decisions on operating point (Minimum WER, EER etc.) and graphical representation of results (ROC, DET and EPC curves). The second perspective requires a simplification of the pseudo-ternary problem, in, for example, two binary classification problems. This, on the other hand, imposes certain database requirements, and presentation attacks databases which do not satisfy them can not be used for evaluation of biometric verification systems under presentation attacks. Depending on the steps undertaken to simplify the pseudo-ternary problem, the evaluation paradigm for the system differs. In particular, in this chapter, we discussed three evaluation methodologies, together with the error rates and the plots associated with them[19].

As the interest for presentation attack detection in almost all biometric modes is growing both in research, but even more in industrial environment, a common fair criteria for evaluation of presentation attack detection systems and of verification systems under presentation attacks is becoming of essential importance. For the time being, there is a lot of inconsistency in the error rates conventions, as well as the evaluation strategies used in different publications.

Acknowledgements The authors would like to thank the projects BEAT (`http://www.beat-eu.org`) and TABULA RASA (`http://www.tabularasa-euproject.org`) both funded under the 7th Framework Programme of the European Union (EU) (grant agreement number 284989 and 257289) respectively. The revision of this chapter was supported under the project on Secure Access Control over Wide Area Networks (SWAN) funded by the Research Council of Norway (grant no. IKTPLUSS 248030/O70). and by the Swiss Center for Biometrics Research and Testing.

[19] The software to reproduce the plots of this chapter is available in `https://gitlab.idiap.ch/bob/bob.hobpad2.chapter24`

Index

A

Area Under ROC 12
Attack Presentation Classification Error Rate 13
Attack Success Rate 15

B

Bona-fide Presentation Classification Error Rate 13

D

Detection-Error Tradeoff 12

E

Equal Error Rate 11
evaluation 4
Expected Performance and Spoofability Curve 19
Expected Performance Curve 12

F

False Accept Rate 13
False Alarm Rate 14
False Fake Rate 14
False Live Rejection Rate 14
False Living Rate 14
False Match Rate 14
False Negative 11
False Negative Rate 11
False Non-Match Rate 14
False Non-Suspicious Presentation Detection 14
False Positive 11
False Positive Rate 11
False Reject Rate 13
False Spoof Accept Rate 14
False Suspicious Presentation Detection 14

G

Global False Accept Rate 15
Global False Reject Rate 15

H

Half Total Error Rate 11

I

Impostor Attack Presentation Match Rate 15

L

licit scenario 8
Liveness False Accept Rate 15

P

PAD 4

R

Receiver Operating Characteristic 12

S

spoof scenario 8
spoofing 4
Success Rate 15

T

True Positive Rate 11

W

Weighted Error Rate 11

Glossary

APCER: Attack Presentation Classification Error Rate, 13
AUC: Area Under ROC, 12

BPCER: Bona-fide Presentation Classification Error Rate, 13

DET: Detection-Error Tradeoff, 12

EER: Equal Error Rate, 11
EPC: Expected Performance Curve, 12
EPSC: Expected Performance and Spoofability Curve, 19

FAR: False Accept Rate, 13
FFR: False Fake Rate, 14
FLR: False Living Rate, 14
FMR: False Match Rate, 14
FN: False Negative, 11
FNMR: False Non-Match Rate, 14
FNR: False Negative Rate, 11
FNSPD: False Non-Suspicious Presentation Detection, 14
FP: False Positive, 11
FPR: False Positive Rate, 11
FRR: False Reject Rate, 13
FSPD: False Suspicious Presentation Detection, 14

GFAR: Global False Accept Rate, 15
GFRR: Global False Reject Rate, 15

HTER: Half Total Error Rate, 11

IAPMR: Impostor Attack Presentation Match Rate, 15

LFAR: Liveness False Accept Rate, 15

PA: Presentation attack, 3
PAD: Presentation attack detection, 3

ROC: Receiver Operating Characteristic, 12

TPR: True Positive Rate, 11

WER: Weighted Error Rate, 11

References

1. Information technology – Biometric presentation attack detection – Part 3: Testing and reporting. Standard, International Organization for Standardization, Geneva, CH (2017). URL `https://www.iso.org/standard/67381.html`
2. Adler, A., Schuckers, S.: Encyclopedia of Biometrics, chap. Security and Liveness, Overview, pp. 1146–1152. Springer-Verlag (2009)
3. Akhtar, Z., Fumera, G., Marcialis, G.L., Roli, F.: Robustness analysis of likelihood ration score fusion rule for multi-modal biometric systems under spoof attacks. In: 45th IEEE International Carnahan Conference on Security Technology, pp. 237–244
4. Akhtar, Z., Fumera, G., Marcialis, G.L., Roli, F.: Robustness evaluation of biometric systems under spoof attacks. In: 16th International Conference on Image Analysis and Processing, pp. 159–168
5. Akhtar, Z., Fumera, G., Marcialis, G.L., Roli, F.: Evaluation of serial and parallel multibiometric systems under spoofing attacks. In: 5th IEEE International Conference on Biometrics: Theory, Applications and Systems (2012)
6. Alegre, F., Vipperla, R., Evans, N., Fauve, B.: On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In: Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, pp. 36–40 (2012)
7. Bao, W., Li, H., Li, N., Jiang, W.: A liveness detection method for face recognition based on optical flow field. In: Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on, pp. 233–236 (2009)
8. Bengio, S., Keller, M., Mariéthoz, J.: The expected performance curve. Tech. Rep. Idiap-RR-85-2003, IDIAP (2003)
9. Bonastre, J.F., Matrouf, D., Fredouille, C.: Artificial impostor voice transformation effects on false acceptance rates. In: INTERSPEECH, pp. 2053–2056 (2007)
10. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pp. 612–618. IEEE (2017)
11. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proc. IEEE Int. Conf. of the Biometrics Special Interest Group (BIOSIG), pp. 1–7 (2012)
12. Chingovska, I., Anjos, A., Marcel, S.: Anti-spoofing in action: joint operation with a verification system. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Biometrics (2013)
13. Chingovska, I., Rabello dos Anjos, A., Marcel, S.: Biometrics evaluation under spoofing attacks. Information Forensics and Security, IEEE Transactions on **9**(12), 2264–2276 (2014). URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6879440`
14. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The REPLAY-MOBILE Face Presentation-Attack Database. In: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, pp. 1–7. IEEE (2016). URL `http://ieeexplore.ieee.org/abstract/document/7736936/`
15. Ergünay, S.K., Khoury, E., Lazaridis, A., Marcel, S.: On the vulnerability of speaker verification to realistic voice spoofing. In: IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS) (2015). URL `https://publidiap.idiap.ch/downloads//papers/2015/KucurErgunay_IEEEBTAS_2015.pdf`
16. Galbally, J., Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J.: A high performance fingerprint liveness detection method based on quality related features. Future Gener. Comput. Syst. **28**(1), 311–321 (2012)
17. Galbally, J., Cappelli, R., Lumini, A., de Rivera, G.G., Maltoni, D., Fiérrez, J., Ortega-Garcia, J., Maio, D.: An evaluation of direct attacks using fake fingers generated from iso templates. Pattern Recognition Letters **31**(8), 725–732 (2010)

18. Galbally, J., Fierrez, J., Alonso-Fernandez, F., Martinez-Diaz, M.: Evaluation of direct attacks to fingerprint verification systems. Telecommunication Systems, Special Issue on Biometrics **47**(3), 243–254 (2011)
19. Galbally-Herrero, J., Fierrez-Aguilar, J., Rodriguez-Gonzalez, J.D., Alonso-Fernandez, F., Ortega-Garcia, J., Tapiador, M.: On the vulnerability of fingerprint verification systems to fake fingerprints attacks. In: IEEE International Carnahan Conference on Security Technology, pp. 169–179 (2006)
20. Gao, X., Tsong Ng, T., Qiu, B., Chang, S.F.: Single-view recaptured image detection based on physics-based features. In: IEEE International Conference on Multimedia & Expo (ICME). Singapore (2010)
21. Ghiani, L., Yambay, D., Mura, V., Tocco, S., Marcialis, G., Roli, F., Schuckers, S.: Livdet 2013 - fingerprint liveness detection competition. In: IEEE Int. Conf. on Biometrics (ICB) (2013)
22. Hastie, T., Tibshirani, R., Friedman, J.H.: The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations. New York: Springer-Verlag (2001)
23. Jain, A.K., Flynn, P., Ross, A.A. (eds.): Handbook of Biometrics. Springer-Verlag (2008)
24. Johnson, P.A., Tan, B., Schuckers, S.: Multimodal fusion vulnerability to non-zero (spoof) imposters. In: IEEE International Workshop on Information Forensics and Security (2010)
25. Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N., Yamagishi, J., Lee, K.A.: The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017)
26. Korshunov, P., Goncalves, A.R., Violato, R.P., Simões, F.O., Marcel, S.: On the Use of Convolutional Neural Networks for Speech Presentation Attack Detection. In: International Conference on Identity, Security and Behavior Analysis (2018)
27. Lui, Y.M., Bolme, D., Phillips, P., Beveridge, J., Draper, B.: Preliminary studies on the good, the bad, and the ugly face recognition challenge problem. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on, pp. 9–16 (2012)
28. Mansfield, A.J., Wayman, J.L., Dr, A., Rayner, D., Wayman, J.L.: Best practices in testing and reporting performance (2002)
29. Marasco, E., Ding, Y., Ross, A.: Combining match scores with liveness values in a fingerprint verification system. In: 5th IEEE International Conference on Biometrics: Theory, Applications and Systems (2012)
30. Marasco, E., Johnson, P., Sansone, C., Schuckers, S.: Increase the security of multibiometric systems by incorporating a spoofing detection algorithm in the fusion mechanism. In: Proceedings of the 10th international conference on Multiple classifier systems, pp. 309–318 (2011)
31. Marcialis, G.L., Lewicke, A., Tan, B., Coli, P., Grimberg, D., Congiu, A., Tidu, A., Roli, F., Schuckers, S.: First international fingerprint liveness detection competition – livdet 2009. In: Proc. IAPR Int. Conf. on Image Analysis and Processing (ICIAP), pp. 12–23. LNCS-5716 (2009)
32. Martin, A., Doddington, G., Kamm, T., M. Ordowski, M.: The det curve in assessment of detection task performance. In: Eurospeech, pp. 1895–1898 (1997)
33. Matsumoto, T., Matsumoto, H., Yamada, K., Hoshino, S.: Impact of artifical ”gummy” fingers on fingerprint systems. In: SPIE Proceedings: Optical Security and Counterfeit Deterrence Techniques, vol. 4677 (2002)
34. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pp. 1–8 (2007)
35. Patrick, P., Aversano, G., Blouet, R., Charbit, M., Chollet, G.: Voice forgery using alisp: Indexation in a client memory. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP ’05). IEEE International Conference on, vol. 1, pp. 17–20 (2005)
36. Peixoto, B., Michelassi, C., Rocha, A.: Face liveness detection under bad illumination conditions. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, pp. 3557–3560 (2011)

37. Pinto, A., Schwartz, W.R., Pedrini, H., de Rezende Rocha, A.: Using visual rhythms for detecting video-based facial spoof attacks. IEEE Transactions on Information Forensics and Security **10**(5), 1025–1038 (2015)
38. P.Johnson, Lazarick, R., Marasco, E., Newton, E., Ross, A., Schuckers, S.: Biometric liveness detection: Framework and metrics. In: International Biometric Performance Conference (2012)
39. Poh, N., Bengio, S.: Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication. Pattern Recognition Journal **39**
40. Rodrigues, R., Kamat, N., Govindaraju, V.: Evaluation of biometric spoofing in a multimodal system. In: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on (2010)
41. Rodrigues, R.N., Ling, L.L., Govindaraju, V.: Robustness of multimodal biometric fusion methods against spoofing attacks. Journal of Visual Languages and Computing **20**(3), 169–179 (2009)
42. Ruiz-Albacete, V., Tome-Gonzalez, P., Alonso-Fernandez, F., Galbally, J., Fierrez, J., Ortega-Garcia, J.: Direct attacks using fake images in iris verification. In: Proc. COST 2101 Workshop on Biometrics and Identity Management, BIOID, pp. 181–190. Springer (2008)
43. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: Proc. European Conference on Computer Vision (ECCV), LNCS 6316, pp. 504–517. Springer (2010)
44. Tome, P., Marcel, S.: On the vulnerability of palm vein recognition to spoofing attacks. In: The 8th IAPR International Conference on Biometrics (ICB) (2015). URL `http://publications.idiap.ch/index.php/publications/show/3096`
45. Tome, P., Raghavendra, R., Busch, C., Tirunagari, S., Poh, N., Shekar, B.H., Gragnaniello, D., Sansone, C., Verdoliva, L., Marcel, S.: The 1st competition on counter measures to finger vein spoofing attacks. In: The 8th IAPR International Conference on Biometrics (ICB) (2015). URL `http://publications.idiap.ch/index.php/publications/show/3095`
46. Tome, P., Vanoni, M., Marcel, S.: On the vulnerability of finger vein recognition to spoofing. In: IEEE International Conference of the Biometrics Special Interest Group (BIOSIG) (2014). URL `http://publications.idiap.ch/index.php/publications/show/2910`
47. Tronci, R., Muntoni, D., Fadda, G., Pili, M., Sirena, N., Murgia, G., Ristori, M., Ricerche, S., Roli, F.: Fusion of multiple clues for photo-attack detection in face recognition systems. In: Proceedings of the 2011 International Joint Conference on Biometrics, IJCB '11, pp. 1–6. IEEE Computer Society (2011)
48. Vanoni, M., Tome, P., El Shafey, L., Marcel, S.: Cross-database evaluation with an open finger vein sensor. In: IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BioMS) (2014). URL `http://publications.idiap.ch/index.php/publications/show/2928`
49. Villalba, J., Lleida, E.: Preventing replay attacks on speaker verification systems. In: Security Technology (ICCST), 2011 IEEE International Carnahan Conference on, pp. 1–8 (2011)
50. Wang, L., Ding, X., Fang, C.: Face live detection method based on physiological motion analysis. Tsinghua Science and Technology **14**(6), 685–690 (2009)
51. Wen, D., Han, H., Jain, A.K.: Face Spoof Detection With Image Distortion Analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015). DOI 10.1109/TIFS.2015.2400395
52. Yambay, D., Ghiani, L., Denti, P., Marcialis, G., Roli, F., Schuckers, S.: LivDet 2011 - fingerprint liveness detection competition 2011. In: Biometrics (ICB), 2012 5th IAPR International Conference on, pp. 208–215 (2012)
53. yan, J., Zhang, Z., Lei, Z., Yi, D., Li, S.Z.: Face liveness detection by exploring multiple scenic clues. In: 12th International Conference on Control, Automation, robotics and Vision (ICARCV 2012). China (2012)

54. Zhang, Z., Yi, D., Lei, Z., Li, S.: Face liveness detection by learning multispectral reflectance distributions. In: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 436–441 (2011)
55. Zhiwei, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: Proc. IAPR Int. Conf. on Biometrics (ICB), pp. 26–31 (2012)