IDIAP RESEARCH REPORT

BIOMETRICS EVALUATION UNDER SPOOFING ATTACKS

Ivana Chingovska André Anjos Sébastien Marcel

Idiap-RR-12-2014

AUGUST 2014

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Biometrics Evaluation under Spoofing Attacks

Ivana Chingovska, André Anjos, Sébastien Marcel

***Abstract*—While more accurate and reliable than ever, the trustworthiness of biometric verification systems is compromised by the emergence of spoofing attacks. Responding to this threat, numerous research publications address isolated spoofing detection, resulting in efficient counter-measures for many biometric modes. However, an important, but often overlooked issue regards their engagement into a verification task and how to measure their impact on the verification systems themselves. A novel evaluation framework for verification systems under spoofing attacks, called Expected Performance and Spoofability (EPS) framework, is the major contribution of this paper. Its purpose is to serve for an objective comparison of different verification systems with regards to their verification performance and vulnerability to spoofing, taking into account the system's application-dependent susceptibility to spoofing attacks and cost of the errors. The convenience of the proposed open-source framework is demonstrated for the face mode, by comparing the security guarantee of four baseline face verification systems before and after they are secured with anti-spoofing algorithms.**

***Index Terms*—Attack, Counter-Measures, Counter-Spoofing, Disguise, Dishonest Acts, Biometric Verification, Forgery, Liveness Detection, Replay, Spoofing, Evaluation, Face recognition**

I. Introduction

Automatically recognizing people by their biometric characteristics is a well-established research area. Although some biometric modes already have a wide usage in security systems, novel traits keep on being discovered [1], [2], [3]. The typical way to recognize people by their traits is to create a biometric reference (often referred to as template or model) which allows comparison (matching) to biometric samples [4]. For example, in a face recognition system, models can be created from existing user face photos and matched against new photos or video sequences acquired by a camera. Varying acquisition conditions, noise and poor lighting are some of the problems that the biometric community is facing, but has successfully solved in many cases. A relatively new security threat that these systems have to handle comes from *spoofing* attacks.

Unlike a zero-effort impostor who may positively claim a different identity despite presenting his own biometric traits [5], in the case of spoofing, the attacker (*active impostor*), tries to fake somebody else's identity by presenting fake samples of that person's traits to the acquisition device. The type of sample used in the attack heavily depends on the acquisition system being attacked. In a fingerprint spoofing attempt, attackers may show molds containing a copy of somebody's prints prepared with silicon [6]. For the voice mode, on the other side, it suffices to present a signal which contains the speaker's vocal characteristics [7]. Interestingly enough, the signal does not even need to be understandable by a human, as long as it exhibits the deterministic vocal features of the attacked identity. Unfortunately, information globalization acts in favor for malicious users, making access to biometric data easily accessible: users' photos and possibly videos may be available through various sites on the Internet. Users' voice can be easily recorded and examined at distance. Fingerprint molds can be easily manufactured from latent marks left on cups and door knobs.

After recognizing the problem of spoofing, different counter-measures have been proposed for many biometric modes. One possible approach, relying on the assumption that spoofing two or more modes is more difficult than spoofing a single one, [8], [9], is combining several of them. Another set of options use additional hardware that will verify the presence of a live person in front of the recognition system, referring to the process as *liveness detection*. Examples are temperature sensing or pulsation detection in the case of fingerprint recognition systems [10]. Other systems ask users to correctly respond to a challenge, like repeating a particular phrase in speaker recognition or changing the facial expression in face recognition. The current trend though suggests completely automatic and autonomous software-based anti-spoofing solutions which rely solely on additional processing of the information captured by the system's biometric sensor and which are likely more convenient for deployment and user experience [11].

Up to this point, biometrics researchers have tackled the problems of biometric recognition and anti-spoofing independently. Researchers in biometric verification develop binary classification systems capable of distinguishing two categories of samples: genuine users as a positive class and zero-effort impostors as a negative one. On the other hand, the anti-spoofing community has been focused on the binary classification problem of discriminating real accesses as a positive with respect to spoofing attempts as a negative class. The relation of the anti-spoofing to the biometric systems has been mostly disregarded. Evaluation of the two types of systems is also performed independently, usually following the evaluation conventions for binary classification systems. Note that, besides verification, biometric recognition systems can work in an identification mode, which is more suitable for negative recognition applications [5]. This paper, however, focuses on biometric verification and biometric identification is out of its scope.

A spoofing counter-measure, by definition, needs to protect a biometric verification system and its role comes into play when coupled with the latter. From an application point of view, we are interested not in a system which detects spoofing attacks, but which recognizes identities and accepts them only

Ivana Chingovska is with Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne, Switzerland, e-mail: ivana.chingovska@idiap.ch

André Anjos and Sébastien Marcel are with the Idiap Research Institute, Switzerland, e-mails: {andre.anjos, sebastien.marcel}@idiap.ch

if they are not spoofing attacks. Thus, to build a highly secure environment, we need a system which, one way or the other, performs person verification in a highly reliable and trustworthy way.

These observations emphasize the drawbacks of the independent treatment of verification and anti-spoofing systems for real-world applications. Attempt to ally the two systems together in order to create a spoof-resistant verification system have been already presented in several publications [12], [13], [14]. In such a setup, unless the spoofing counter-measure has perfect discrimination capabilities, a drop in the verification performance can be expected.

The biometric verification system, regardless of whether and how it incorporates mechanism for rejecting spoofing attacks, now has to handle them as an additional input class. We attribute the necessity of this step to the fact that the final system design considerations, like the expected frequency of attacks, can not be known prior to deployment time. Having three classes at input instead of two requires a complete redefinition of the problem of biometric verification. Furthermore, if we want to have a precise analysis of the system performance, we need to have a suitable metrics for measuring its *spoofability*. Up to this point, different systems have performed the evaluation in a different way and, despite the many attempts, there is no golden standard for evaluating biometric verification systems under spoofing attacks.

The main goal of this paper is to emphasize this issue and the necessity to solve it in order to provide real-world applications, as well as to establish an evaluation framework based on the newly proposed Expected Performance and Spoofability (EPS) framework, that considers all the parameters imposed by the new problem domain. To do this, we firstly review the standards for evaluation of biometric systems in their common setup. Then, we inspect the efforts to adapt them to the new problem definition reporting on their drawbacks for deployment in real world conditions.

To demonstrate the capacity of the proposed evaluation framework, we evaluate and compare several state-of-the-art verification systems under spoofing attacks. The verification systems work with the face mode. The analysis of the spoofing vulnerability of these systems, as well as the study of the change in their performance after adding a spoofing counter-measure are additional contributions. The source code for calculating the measurements and plotting the curves is freely available as well.

In the text that follows, Section II provides a survey on the standard evaluation metrics for binary classification problems, as a basis for the widely accepted methodology for evaluation of biometric verification and anti-spoofing systems. The restatement of the problem of biometric verification system under spoofing attacks, together with the commonly used evaluation methodologies are given in Section III. Section IV describes the proposed evaluation framework. Its practical usage is illustrated in Section V via a comparative analysis of several baseline as well as trustworthy systems in the domain of face verification. Section VI gives our final remarks.

II. Summary of evaluation metrics in biometrics

As both biometric verification and anti-spoofing systems by themselves are of binary nature, the overview of the state-of-the-art will firstly cover the standard metrics for evaluation of binary classification systems in Section II-A. The adaptations of the general metrics to the specific tasks of biometric verification and anti-spoofing are given in Sections II-B and II-C, respectively.

A. Evaluation of binary classification systems

Binary classification systems receive two types of input belonging to two classes, usually referred to as positive and negative class. They are trained to assign scores to the input samples. Then, a threshold is calculated to separate the scores of the positive and the negative class and the samples with scores above the threshold are classified as positives, while the ones with scores below the threshold as negatives.

Metrics for evaluation of binary classification systems are associated to the types of errors they commit and how to measure them, as well as to the threshold calculation and evaluation criterion [15]. Binary classification systems are 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.

An objective and unbiased performance evaluation of the binary classification systems requires a database with a specific design and strictly defined protocols. It is recommended that the samples in the database are divided into three subsets: training $\mathcal{D}_{train}$, development (validation) $\mathcal{D}_{dev}$ and test (evaluation) set $\mathcal{D}_{test}$ [16]. Even greater objectivity will be achieved if the identities in separate subsets do not overlap [17]. The training set serves to train the system, while its fine tuning is done using the development set. Since in a real world scenario the final system will be used for data which have not been seen before, the performance measure is normally reported on the test set [16], [18]. An exception from this recommended design may happen if the number of samples in the database is not big enough. In such a case, the samples can be divided only in training and test set, and tuning of the parameters is done with a cross-validation procedure [16].

The decision threshold τ is computed to serve as a boundary between the output scores of the positive and the negative class. 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. One well established criterion is Equal Error Rate (EER) [15], which selects the threshold τ^*_{EER} to ensure that the difference between FPR and FNR is as small as possible (Eq. 1). The optimal threshold, also referred to as *operating point*, is a tuning parameter, and it is usually determined using the development set [16], [18].

$$\tau^*_{\text{EER}} = \arg\min_{\tau} |\text{FPR}(\tau, \mathcal{D}_{dev}) - \text{FNR}(\tau, \mathcal{D}_{dev})| \quad (1)$$

Once the threshold τ^* is determined, the accuracy of the system can be summarized reporting different metrics. For example, the Detection Cost Function (DCF), given in Eq. 2, has been proposed in [19] and is used in the NIST evaluations [20]. The DCF accounts for the cost of the error rates (c_{FPR} and c_{FNR}), as well as for the probability of occurrence of positive and negative samples (p_{pos} and p_{neg}).

$$\begin{aligned}\text{DCF}(\tau^*, \mathcal{D}_{test}) = c_{\text{FPR}} \cdot p_{\text{neg}} \cdot \text{FPR}(\tau^*, \mathcal{D}_{test}) \\ + c_{\text{FNR}} \cdot p_{\text{pos}} \cdot \text{FNR}(\tau^*, \mathcal{D}_{test})\end{aligned} \tag{2}$$

By giving equal priors to the occurrence of positive and negative samples and normalizing the cost values, Weighted Error Rate (WER) is proposed in [18]. In its computation (Eq.3), $\beta \in [0, 1]$ is the parameter balancing between the cost of FPR and FNR. For the special case of $\beta = 0.5$, the Half Total Error Rate (HTER) is reached.

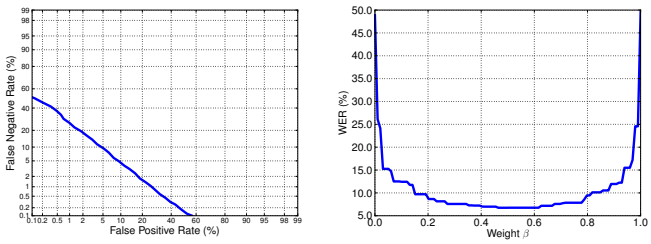
$$\begin{aligned}\text{WER}_\beta(\tau^*, \mathcal{D}_{test}) = \beta \cdot \text{FPR}(\tau^*, \mathcal{D}_{test}) \\ + (1 - \beta) \cdot \text{FNR}(\tau^*, \mathcal{D}_{test})\end{aligned} \tag{3}$$

Important tools in evaluation of classification systems are the different graphical representations of the classification results. For example, to present the trade-off between FPR and FNR depending on the threshold, the performance of the binary classification systems is often visualized using Receiver Operating Characteristic (ROC) curve. Parameterizing over different values for the decision threshold, the ROC curve usually plots FPR versus 1-FNR. Sometimes, when one number is needed to represent the performance of the system in comparison with other systems, the Area Under ROC curve (AUC) may be reported. The higher the AUC the better the system.

A normal deviate transformation of the ROC curve yields the Detection-Error Tradeoff (DET) curve [21]. Its usage is convenient for comparing systems whose scores follow a Gaussian distribution, since such a transformation guarantees that the curve will become a line. It plots FPR versus FNR. Fig. 1a illustrates the DET curve for a hypothetical binary classification system[1].

Although ROC and DET curves may give an idea about the expected performance of a single system under different thresholds, using them to compare two or more systems can lead to biased conclusions [22]. Usually, when comparing two systems using ROC or DET curves, we select a certain value on the abscissa (most often FPR) as a first step, and then we read the values on the ordinate for the two systems (for example FNR) as a second step. In this way, during the first step, we implicitly choose a threshold *a posteriori*, i.e. on the same data used to read and compare the error rates in the second step. This threshold may not be the optimal one for any of the two systems. However, for an objective comparison, the error rates for the two systems have to be reported at their optimal thresholds, which have to be chosen *a priori*, on a separate data. Unfortunately, by plotting only the error rates on a test set at thresholds not related to the development set, the ROC and DET curves do not give any hint about the optimal thresholds of the two systems. Hence, the conclusions about which one out of two systems is better may be misleading if drawn solely from the ROC or DET curves.

[1]Plots for a hypothetical biometric systems in the figures in this paper are based on a synthetically generated score data. They serve solely to illustrate the concept presented in this paper.



(a) DET curve (b) EPC curve

Fig. 1: Evaluation plots for hypothetical biometric verification system

To solve this issue, the so-called Expected Performance Curve (EPC) is proposed in [22]. It fills in for two main disadvantages of the ROC and DET curves: firstly, it plots the error rate on the test set depending on a threshold selected *a priori* on the development set; and secondly, it accounts for varying relative cost $\beta \in [0; 1]$ of FPR and FNR when calculating the threshold. In the EPC framework, an optimal threshold τ^*_β depending on β is computed based on a certain criteria on the development set. For example, the threshold can be chosen to minimize WER_β for different values of β, which is the variable parameter plotted on the abscissa. The performance for the calculated values of τ^*_β is then computed on the test set. WER_β or any other measure of importance can be plotted on the ordinate axis. The parameter β can be interpreted as the cost of the error rates, but also as the prior of having a positive or a negative sample as an input. One may observe the error rates and compare systems only in the range of values of β which are of interest for a particular application. The EPC curve is illustrated in Fig. 1b for a hypothetical binary classification system.

The performance of a binary system can be summarized in one value by computing the area under the EPC, defined as the expected average of two antagonistic error rates that are being plotted [22].

B. Evaluation of biometric verification systems

The biometrics community has established a common terminology for the samples of the positive and the negative class from the perspective of a biometric verification system [5]:

- *Genuine users* for samples of the positive class,
- *(Zero-effort) impostors* for samples of the negative class.

Hence, in the domain of biometric verification systems, the number of errors known as FP and FN refer to the number of zero-effort impostors incorrectly classified as genuine users and the number of genuine users incorrectly classified as zero-effort impostors, respectively. Since the positives and the negatives are associated with the action of *acceptance* and

rejection by the verification system, a common practice is to replace FPR and FNR with False Acceptance Rate (FAR) and False Rejection Rate (FRR), respectively [4]. Furthermore, due to the process of matching between the samples and the models, FPR and FNR are often reported as False Match Rate (FMR) and False Non-Match Rate (FNMR) [5][2]. More thorough list of synonyms typically used is given in Table II in Appendix A.

An important aspect of a biometric verification database is that part of the samples in the training, development and test set needs to be designated for creating the models for the identities. These samples are usually referred to as *enrollment* [5] (*reference* [23]) data.

C. Evaluation of anti-spoofing systems

In the anti-spoofing community, the terminology to name the samples of the positive and the negative class is as follows:

- *Real accesses* [24] or *live samples* [11], [25] for samples of the positive class,
- *Spoofing* or *presentation attacks* [26] for samples of the negative class.

Anti-spoofing systems work on the principle of acceptance and rejection as well. Hence, in this scope, FAR and FRR are the most commonly used terms for FPR and FNR too. FAR stands for the ratio of incorrectly accepted spoofing attacks and FRR for the ratio of incorrectly rejected real accesses. These error rates are often substituted with different synonyms by different authors. The most common of them are listed in Table I in Appendix A.

When it comes to databases for evaluation of anti-spoofing systems, their primary task is to provide two types of samples: real accesses and spoofing attacks of a number of identities. Additionally, the spoofing database needs to satisfy the requirements of binary classification problems, as the isolated spoofing detection is.

III. Evaluation of Biometric Verification Systems under Spoofing Attacks

While the problem of biometric verification is undoubtedly in the class of binary classification problems, a shift in the concept is required when spoofing attacks are present as a third possible input type. The newly posed system needs a new problem definition, which will be stated in Section III-A. It will help to better understand the metrics which have been used for evaluation of such systems, which, together with their drawbacks, are discussed in Section III-B. We propose a novel evaluation methodology which is better suited to the problem in Section IV.

A. Problem statement and database design

When treating biometric verification as a binary classification system, the designers are interested in determining the capacity of a given system to discriminate between different identities. As explained in Section II-B the systems are assumed to receive two classes of input samples. Depending on the internal algorithm, these systems may or may not have the competence to discover if the input sample comes from a live person present in front of the system, or a spoofing attack.

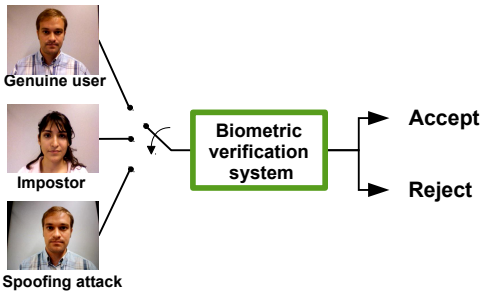


Fig. 2: Biometric verification system under spoofing attack

An accurate representation of the operation of a verification system acknowledging the spoofing attack samples as a possible input type, is given in Fig. 2. It needs to accept only the samples from the class of genuine users, while both zero-effort impostors and spoofing attacks need to be rejected. Consequently, the system is not necessarily required to be able to discriminate between three classes and the problem does not need to be treated as ternary. The system can still operate as a binary classification system, as long as it is able to determine the classes that it needs to reject. Therefore, it is convenient to denote the two classes that need to be rejected as negative.

Despite the comfort of keeping the binary nature of the verification system, it is still of importance to evaluate how vulnerable the system is to spoofing attacks. The evaluation metrics presented in Section II are sufficient to describe only the verification performance of a system. But now, besides FAR and FRR, suitable metric is needed to report on the system spoofability. Additional problem is the way to determine an operation point for such a system. These issues are discussed in Sections III-B and IV.

Before proceeding with the evaluation metrics themselves, a short notice on the design of a database for evaluation of verification systems under spoofing attacks is due. Namely, it has to satisfy the requirements of both a biometric verification (Section II-B) and spoofing II-C database. Typically, the spoofing databases follow the design given in II-C, which poses a major limitation: lack of data to enroll identities in a verification system. Indeed, separate enrollment data within the spoofing database are needed to build models for the identities. In this way, a training and spoofability assessment of a verification system using the spoofing database is enabled.

To formalize the process of training and evaluating a verification system using a spoofing database, let's represent the identity i in the database with the tuple $(\mathbf{x}_i^r, \mathbf{x}_i^s, \mathbf{x}_i^e)$, containing real access $\mathbf{x}_i^r$, spoofing attack $\mathbf{x}_i^s$ and enrollment $\mathbf{x}_i^e$ samples. Then, the spoofing database, providing data for N identities, can be denoted as $\mathcal{D} = \{(\mathbf{x}_i^r, \mathbf{x}_i^s, \mathbf{x}_i^e) : i = 1..N\}$. The process of training a verification system using the spoofing database means creating a set of models $\mathcal{M} = \{\mathcal{M}_i : i = 1..N\}$, where $\mathcal{M}_i = f(\mathbf{x}_i^e)$ and $f(\cdot)$ is a function

[2]In general, the error rates FMR and FNMR are not exactly synonymous with FAR and FRR [5]. However, they are equivalent in the context presented in this paper. Please see Appendix A for further details.

that maps samples to a model. Then, the verification system computes the scores for the classes of real accesses, zero-effort impostors and spoofing attacks. The set of scores for the genuine users may be created by comparing the real access samples of one identity to the model of the same identity: $\mathcal{S}_{genuine} = \{g(\mathbf{x}_i^r, \mathcal{M}_i) : i = 1..N\}$, where $g(\cdot,\cdot)$ is a matching function. A logical way to assemble the set of zero-effort impostor scores is by comparing the real access samples of one identity to the models of the other identities in an exhaustive manner (full cross-comparison [5]), which results in $\mathcal{S}_{impostor} = \{g(\mathbf{x}_i^r, \mathcal{M}_j) : i,j = 1..N, i \neq j\}$. Finally, to assemble the set of spoofing attack scores for the verification system, one needs to compare the spoofing attack samples from one identity to the model of the same identity, which yields $\mathcal{S}_{spoof} = \{g(\mathbf{x}_i^s, \mathcal{M}_i) : i = 1..N\}$.

B. Evaluation methodologies

While the performance metrics for verification systems is well established and widely used, the evaluation for verification systems under spoofing attacks is not unified and is ambiguous in different publications. A detailed overview of all the error rates utilized by various authors is given in Table II in Appendix A.

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

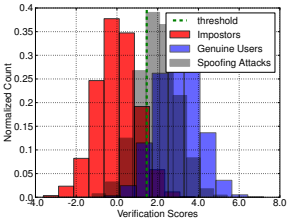
- FRR - ratio of incorrectly rejected genuine users,
- FAR - ratio of incorrectly accepted zero-effort impostors,
- SFAR - ratio of incorrectly accepted spoofing attacks [27].

Fig. 3a shows a plot of the distributions of the scores of the three input classes obtained by a hypothetical verification system. The problem that arises due to the existence of three score distributions is how to determine the decision threshold to discriminate between the samples to accept and reject. A widely accepted strategy to simplify the problem is to decompose it into two sub-problems which resemble the original binary classification problem in biometric verification. The sub-problems correspond to two scenarios the system can operate in:

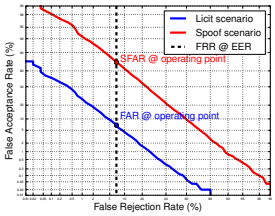
- *Licit* scenario (also called normal operation mode [28]): considers genuine users as positive and only zero-effort impostors as negative class,
- *Spoof* scenario: considers genuine users as positive and only spoofing attacks as negative class.

Researchers generally follow two main evaluation methodologies to obtain the decision threshold and to report the error rates it produces, and they are discussed below.

a) Methodology 1: In the first evaluation methodology, two decision threshold calculations are performed separately for the two scenarios [6], [28], [27], [7]. Analysis of the system in the licit scenario gives values for FRR and FAR, while analysis in the spoof scenario gives values for FRR and SFAR. Since the analysis produces different threshold in the two scenarios, the two values of FRR are not the same. A major weak point of this type of evaluation is that it outputs two decision thresholds for a single verification system, while naturally a single system can have only one operating point



(a) Score distributions and threshold using licit scenario



(b) DET curves

Fig. 3: Graphical tools for evaluation of hypothetical biometric verification system under spoofing attacks

corresponding to one decision threshold. Furthermore, the spoof scenario assumes that all the possible misuses of the system come from spoofing attacks, which in general is not realistic. The threshold calculated in this scenario is not a good discriminating point for a verification system, but rather for an anti-spoofing system and the error rates reported on this way are not a reliable estimate of the system performance under spoofing attacks. The decision threshold and the reported error rates in the spoof scenario are irrelevant in a real-world scenario. Therefore, this type of evaluation is not compliant to a real-world requirements for operation of a verification system.

b) Methodology 2: The second evaluation methodology is adapted for more realistic performance evaluation. The threshold is calculated using various criteria, for example EER, but almost always using the licit scenario, as it is regarded as a normal operation mode for a verification system. Taking advantage of the fact that the licit and spoof scenario share the same positive class, many publications choose a threshold to achieve a particular desired value of FRR [29], [30], [31], [32], [33], [34], [12]. Then, using the obtained threshold, FAR for the licit and SFAR in the spoof scenario are reported and compared.

On the hypothetical verification system whose score distribution is plotted in Fig. 3a, the threshold is chosen using the EER criteria for the licit scenario. The plotted threshold gives an intuition about how well the system discriminates between genuine users and zero-effort impostors, but also between genuine users and spoofing attacks. Fig. 3b draws two DET curves corresponding to the two scenarios. The vertical line shows the FRR for the chosen threshold. The points where it cuts the DET curves for the two scenarios are the reported error rates.

As an alternative figure delivering similar information as DET for the second evaluation methodology, [32] suggests to plot FAR vs. SFAR. Thresholds are fixed in order to obtain all the possible values of FAR for the licit scenario and SFAR is computed in the spoof scenario and plotted on the ordinate axis. By plotting the curves for different verification systems, the plot enables to compare which one of them is less prone to spoofing given a particular verification performance.

The issue that the second methodology overlooks is that a system whose decision threshold is optimized for one negative

class (usually, the zero-effort impostors), can not be evaluated in a fair manner for another negative class (spoofing attacks). Expectedly, such a threshold will be biased towards the single negative class used for its determination, causing unnecessary larger error rates for the other negative class. If the system is expected to be exposed to two classes of negatives in the test or deployment stage, it would be fair that both of them play a role in the decision of the threshold in the development stage. A novel evaluation methodology to tackle this issue is the subject of Section IV.

IV. Expected Performance and Spoofability Evaluation Framework

Determining the decision threshold for biometric verification systems under spoofing attacks seems to be one of the major issues in the evaluation process. Neither the first, nor the second of the evaluation methodologies explained in Section III-B offer a method that determines an unbiased threshold applicable in a realistic verification scenario. A fair evaluation of a system which needs to reject samples of two different classes is possible only if both of them are considered in the development stage. By neglecting the class of spoofing attacks when deciding on the threshold of the verification system, one deliberately exhibits blindness to the danger of spoofing attacks, thus potentially creating a system more vulnerable to spoofing. Moreover, in some cases a necessity may arise to add a cost to the error rates associated with the positive and the negative class, and this cost has to be considered in the process of computing a decision threshold as well.

The most straight-forward way to involve both negative classes (zero-effort impostors and spoofing attacks) in the threshold decision process, is simply to merge them together into a single negative super-class. However, the number of zero-effort impostors and spoofing attacks is highly dependent on the database and follows the database protocol. Hence, the ratio of the two classes into the super-class is different for different databases and can not be controlled. Furthermore, the super-class tends to be biased towards the component with more samples. For example, in a typical biometric verification database with N identities and M samples per identity, the number of zero-effort impostors will be $N \times (N-1) \times M$. On the other hand, if there is a single spoofing attack for any genuine sample in the database, the number of spoofing attacks will be $N \times M$. The above observations lead to the question of what the correct ratio of zero-effort impostors and spoofing attacks into the super-class of negatives is.

As a matter of fact, there may not be a single answer to that. Any ratio of the two negative classes may be valid depending on the deployment conditions. For example, in highly supervised conditions, like airport control gates, spoofing attacks are more difficult to perform, and hence unlikely. On the other hand, unsupervised verification systems of portable devices are much more exposed to spoofing attacks. Thus, tuning the operating point of any system depends on its expected usage scenario.

The message that the metrics DCF, WER_β and EPC convey sounds with the above reasoning for a biometric verification system. EPC obtains a decision threshold based on a parameter β which balances between FAR and FRR and reports the expected performance for a wide range of values for that parameter. The parameter β can be interpreted as the relative cost or importance of FAR and FRR, or the prior of the negative or the positive class. Using EPC, it is possible to compare algorithms depending on the importance of FAR and FRR in a certain usage scenario.

For evaluating biometric verification systems under spoofing attacks, we develop a method inspired by EPC. Being aware that the prior of zero-effort impostors and spoofing attacks can not be known in advance while developing an algorithm, we design an evaluation framework which measures the expected performance of the system for a range of values of a parameter which balances between FAR and SFAR. Moreover, analogously to EPC, we introduce another parameter which considers the cost of the error rates associated with the positive and the negative classes. As it measures both the verification performance and the vulnerability to spoofing of a system and unifies them into a single value, the adapted evaluation scheme is called Expected Performance and Spoofability (EPS) framework.

The goal of the EPS framework is to analyze and plot error rates regarding the performance and spoofability of a verification system on a test set, with respect to a decision threshold taken on a separate development set. We define two parameters: $\omega \in [0,1]$, which denotes the relative cost of spoofing 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 spoofing attacks) with respect to the positive class. Using these, we introduce a measurement called FAR_ω, which is a weighted error rate for the two negative classes (zero-effort impostors and spoofing attacks). It is calculated as in Eq. 4.

$$\text{FAR}_\omega = \omega \cdot \text{SFAR} + (1-\omega) \cdot \text{FAR} \tag{4}$$

The optimal classification threshold $\tau^*_{\omega,\beta}$ depends on both parameters. It is chosen to minimize the weighted difference between FAR_ω and FRR on the development set, as in Eq. 5.

$$\tau^*_{\omega,\beta} = \arg\min_{\tau} |\beta \cdot \text{FAR}_\omega(\tau, \mathcal{D}_{dev}) - (1-\beta) \cdot \text{FRR}(\tau, \mathcal{D}_{dev})| \tag{5}$$

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 $\text{WER}_{\omega,\beta}$, which can be accounted as a measurement summarizing both the verification performance and the spoofability of the system and which is calculated as in Eq. 6.

$$\begin{aligned}\text{WER}_{\omega,\beta}(\tau^*_{\omega,\beta}, \mathcal{D}_{test}) = \beta \cdot \text{FAR}_\omega(\tau^*_{\omega,\beta}, \mathcal{D}_{test}) \\ + (1-\beta) \cdot \text{FRR}(\tau^*_{\omega,\beta}, \mathcal{D}_{test})\end{aligned} \tag{6}$$

A special case of $\text{WER}_{\omega,\beta}$, obtained by assigning equal cost $\beta = 0.5$ to FAR_w and FRR can be defined as HTER_ω and computed as in Eq. 7. In such a case, the criteria for optimal decision threshold is analogous to the EER criteria given in Section II-A.

$$\mathrm{HTER}_{\omega}(\tau^*_{\omega}, \mathcal{D}_{test}) = \frac{\mathrm{FAR}_{\omega}(\tau^*_{\omega}, \mathcal{D}_{test}) + \mathrm{FRR}(\tau^*_{\omega}, \mathcal{D}_{test})}{2} \tag{7}$$

The parameter ω could be interpreted as relative cost of the error rate related to spoofing attacks. Alternatively, it could be connected to the expected relative number of spoofing 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 under a spoofing attack when it is misused. If it is expected that there is no danger of spoofing attacks for some particular setup, it can be set to 0. In this case, $\mathrm{WER}_{\omega,\beta}$ corresponds to WER_{β} 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 spoofing 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_{β} in Section II-A, 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 spoofing 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.

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 surface, which may not be convenient for evaluation and analysis, especially when one needs to compare two or more systems. Instead, we suggest plotting the Expected Performance and Spoofability Curve (EPSC), 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 FRR 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 SFAR and FAR or the prior of spoofing 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 $\mathrm{WER}_{\omega,\beta}$ on the ordinate, with respect to the varying parameter β on the abscissa.

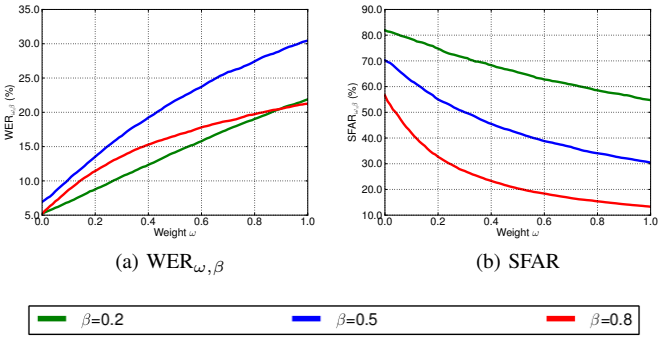
The algorithm on Fig. 4 gives the step-by-step procedure to compute and plot $\mathrm{WER}_{\omega,\beta}$ with regards to ω and β for a given verification system. By fixing one of the parameters ω or β, one can plot EPSC for $\mathrm{WER}_{\omega,\beta}$ with regards to the other parameter.

Besides $\mathrm{WER}_{\omega,\beta}$, EPSC can present other error rates which are of interest. For example, plotting SFAR can show how the system's robustness to spoofing changes with regards to ω or β. Alternatively, to report on all the incorrectly accepted samples, FAR_{ω} can be plotted using EPSC.

Fig. 5 and Fig. 6 give an illustration of the EPSC plotting the

for $\beta \in [0, 1]$ **do**
  for $\omega \in [0, 1]$ **do**
    define $\mathrm{FAR}_{\omega} = \omega \cdot \mathrm{SFAR} + (1 - \omega) \cdot \mathrm{FAR}$
    $\tau^*_{\omega,\beta} = \arg\min_{\tau} |\beta \cdot \mathrm{FAR}_{\omega}(\tau, \mathcal{D}_{dev}) - (1 - \beta) \cdot \mathrm{FRR}(\tau, \mathcal{D}_{dev})|$
    compute $\mathrm{WER}_{\omega,\beta}(\tau^*_{\omega,\beta}, \mathcal{D}_{test})$;
    plot $\mathrm{WER}_{\omega,\beta}(\tau^*_{\omega,\beta}, \mathcal{D}_{test})$ **w.r.t.** ω, β
  end for
end for

Fig. 4: Pseudo code for computing $\mathrm{WER}_{\omega,\beta}$

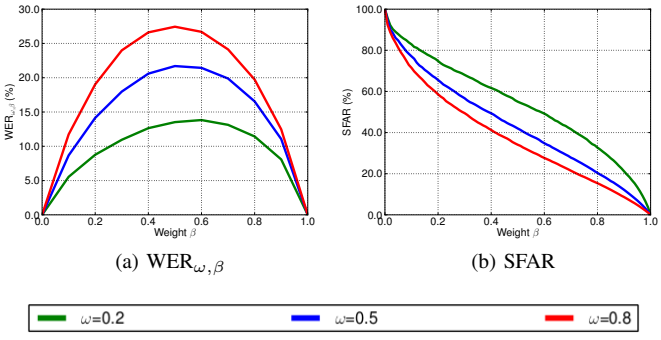


(a) $\mathrm{WER}_{\omega,\beta}$ (b) SFAR

Fig. 5: EPSC of a hypothetical biometric verification system under spoofing attacks, parameterized over ω

error rates $\mathrm{WER}_{\omega,\beta}$ and SFAR as function of the parameters ω and β, respectively. The plots are generated for the hypothetical verification system whose score distribution is given in Fig. 3a.

Fig. 5a and Fig. 5b show $\mathrm{WER}_{\omega,\beta}$ and SFAR with respect to ω for three predefined values of β. The blue curve on Fig. 5a, corresponding to $\beta = 0.5$, is equivalent to HTER_{ω}. The left-most points of the curves correspond to $\omega = 0$, meaning that the decision threshold is obtained disregarding the spoofing attacks as possible input. Hence, the threshold at this point corresponds to the threshold plotted in Fig. 3a, calculated for the system when operating in the licit scenario. For the particular hypothetical system and all the three considered



(a) $\mathrm{WER}_{\omega,\beta}$ (b) SFAR

Fig. 6: EPSC of a hypothetical biometric verification system under spoofing attacks, parameterized over β

values of β, this point corresponds to low $\text{WER}_{\omega,\beta}$, which indicates a system with good verification capabilities, but very high SFAR due to the high overlap of the scores of spoofing attacks and genuine users.

As we increase ω, we give weight to the spoofing attacks so that they have a role in the threshold decision process. In the particular example, this results in a shift of the decision threshold to the right of the score distribution plot in Fig. 3a. This decreases the number of spoofing attacks that pass the system, which explains why SFAR decreases with increasing ω. However, the additional caution for the danger of spoofing attacks unavoidably comes with the price of more rejected genuine users and thus higher $\text{WER}_{\omega,\beta}$. A system with high robustness to spoofing attacks will show as mild increase of $\text{WER}_{\omega,\beta}$ as possible, with as steep decrease of SFAR as possible.

Fig. 6a and Fig. 6b show EPSC parameterized over the varying parameter β, for three predefined values of ω. For the extreme cases where $\beta = 0$ and $\beta = 1$, $\text{WER}_{\omega,\beta}$ is 0 because the threshold is determined to minimize the error rate solely associated with the positive or the negative class, respectively. In the case of $\beta = 0$, this results in a successful passing through of all the spoofing attacks.

Considering the spoofing attacks when calculating the decision threshold means taking additional precautions against them. As a result of this, the threshold obtained using EPS framework is better adapted to the input that is expected, contributing to systems with better performance and lower spoofing vulnerability, than systems whose decision threshold has been determined in different way. This is illustrated for a hypothetical biometric verification system in Appendix C.

The EPSC inherits the advantage of unbiased system comparison from the EPC, because it reports the error rates *a priori*. Since the threshold is always determined using the development set, and the error rates are reported using the test set, one can estimate the expected error rates and spoofability of the system in an unbiased way, on data which has not been seen before. The expected error rates can be reported for a particular value or range of values of the parameters ω and β which are of interest in a particular application. Moreover, EPSC allows for easy and unbiased comparison of verification systems with regards to their performance and robustness to spoofing, simply by comparing the EPSC for the two systems on the same plot. Even more, one can compare verification systems range-wise: which one performs better for a range of values of ω or β. Practical examples of such analysis are given in Section V.

Finally, if a single number is needed to describe the performance of a system, we define the Area Under EPSC (AUE) metric, which can be computed for a fixed β or ω. For example, for a fixed β, it represents the average expected $\text{WER}_{\omega,\beta}$ for all values of ω and is computed using Eq. 8. The formula to compute AUE for fixed ω and varying β follows accordingly. Between two systems, better is the one which achieves smaller AUE.

$$\text{AUE} = \int_{\omega \in [0,1]} \text{WER}_{\omega,\beta}(\tau^*_{\omega,\beta}, \mathcal{D}_{test}) \mathrm{d}\omega \tag{8}$$

The AUE can be computed in between certain bounds $a, b \in [0,1]; a < b$, enabling to compare two systems depending on the required range of the varying parameter.

V. Experimental results

Extensive experiments in the domain of face verification and anti-spoofing were conducted in order to evaluate several state-of-the-art systems using the EPS evaluation framework. In particular, we analyzed four baseline face verification systems and their vulnerability to spoofing attacks. Then, we tried to reduce their vulnerability by incorporating three different spoofing counter-measures. While this process naturally increases the robustness to spoofing of the verification systems, it may also significantly affect its verification performance [14]. The EPS framework proves to be very suitable to analyze the trade-off between these two parameters. Note that EPS framework allows evaluation analysis of any biometric system which can perform verification task, regardless whether and how it has an incorporated mechanism to handle spoofing attacks.

In the following analysis, we begin by introducing a general terminology for categorization of spoofing attacks based on their success in deceiving a verification system in Section V-A. Then, in Section V-B we describe the face spoofing database as well as the face verification and anti-spoofing systems used in the experiments. Empirical results using EPS framework are reported in Sections V-C, V-D, V-E and V-F. Through the analysis, we demonstrate how to interpret EPSC and we illustrate its advantages over other evaluation methodologies.

The reported results are easily reproducible, as the experiments are implemented using the free signal-processing and machine-learning toolbox Bob [35][3]. The source code to compute and plot the EPSC is freely available as Bob's satellite package[4].

As the comparison for 3D plots showing the error rates depending on β and ω is difficult, in our further analysis we fix $\beta = 0.5$ and adhere to comparing systems using HTER_{ω}. This is not an unreasonable choice: the evaluation of many biometric verification systems is traditionally done only by using EER nad HTER.

A. Categories of spoofing attacks

As shown in Section III-B, a score distribution plot as in Fig. 3a may be a good indicator of the discriminability the system demonstrates. Not only it suggests how well the system performs in verification of identities, but it also gives an intuition how vulnerable the system is to spoofing attacks. Depending on the position of the spoofing attack scores on the abscissa, the spoofing attacks can be clustered in 4 distinct categories with regards to a particular verification system.

[3]http://www.idiap.ch/software/bob

[4]http://pypi.python.org/pypi/antispoofing.evaluation

- *Insufficient* attacks: attack scores are distributed inside the histogram area spawned by zero-effort impostors or to the left of it,
- *Sub-optimal* attacks: attack scores are situated between impostors and genuine users,
- *Optimal* attacks: attack scores are contained within the range of the scores of genuine users,
- *Super-Optimal* attacks: attack scores are mostly situated to the right of the scores of genuine users.

Using the previously defined terminology, to make a system more robust to spoofing means bringing the spoofing attacks from optimal and super-optimal to sub-optimal, or, if possible, insufficient level. A straight-forward way to achieve this is by fusing several modes to be verified, like in [32], [34]. Another approach is to blend together the outputs of two separate systems: a verification and an anti-spoofing one. Significant publications covering this problem include [12], [13], [14].

A visualization of the score distributions for the four categories of spoofing attacks is given in Appendix D. Furthermore, we give case studies for EPSC for the four categories of spoofing attacks. They should give an understanding about the differences in the EPSC appearance for a system highly vulnerable and a system highly robust to spoofing.

B. Database and systems

All of the experiments were conducted using the Replay-Attack database [36][5], which is specifically designed for face spoofing. Unlike the other face spoofing databases (NUAA [37] and CASIA-FASD [38]), Replay-Attack satisfies the requirement for training a verification system by providing separate enrollment samples. It contains video sequences of real accesses and attacks to 50 identities. The types of attacks present in this database are printed and digital photographs, as well as videos displayed on a screen.

The experimental setup includes four baseline face verification systems which have proven to be state-of-the-art on several face verification databases. The first one is a Gaussian Mixture Model (GMM) based system which extracts Discrete Cosine Transform (DCT) features from the input images [39]. The second one, called Local Gabor Binary Pattern Histogram Sequences (LGBPHS) [40], calculates Local Binary Patterns (LBP) histograms over the input images convoluted with Gabor wavelets, and computes the similarity scores using χ^2 measure. The third considered system is based on [41] and compares Gabor jets extracted from different positions and put into a single rectangular grid graph (GJet) [42]. Finally, DCT features are used once again in the fourth system, to create Universal Background Model and to estimate a linear subspace of the within-class variability [43]. We will refer to this system as Inter-Session Variability modeling (ISV). The verification scores of these systems on the Replay-Attack database are obtained using the open-source face verification framework from [44][6].

Concerning the face anti-spoofing systems, they can be categorized in three groups with respect to the cues they use to detect the spoofing attack [45]. The first group of systems tries to detect signs of vitality on the scene, like eye-blinking or mouth movements. The second group evaluates the differences in motion patterns, while the third one compares the texture properties for real accesses and attacks. In this work we used three different face anti-spoofing systems whose implementation is published as open-source. The first one uses (LBP) [36][7], while the second one an LBP variant capturing dynamic texture properties in three orthogonal planes (LBP-TOP) [46][8]. The third system estimates the correlation in the movements of the face with regards to the background and detects higher correlation in the case of spoofing attacks [47][9]. These systems show different capacity in detecting the spoofing attacks in Replay-Attack, which consequently affects the performance of the verification system they are fused with.

With a goal to achieve greater robustness to spoofing of the verification systems, we fuse their output with the output of the anti-spoofing systems at score level. In particular, three of the fusion strategies presented in [14][10] are examined: SUM of scores, Logistic Regression (LR) and Polynomial Logistic Regression (PLR).

In the following experiments, we firstly examine the performance of the verification systems at disposal (GMM, LGBPHS, GJet and ISV) and their vulnerability to spoofing attacks in Section V-C. In our second experiment in Section V-D, we compare the fusion methods when employed to fuse the baseline systems with the simplest LBP based anti-spoofing system. In Section V-E, we fix the fusion rule and we perform the comparison with respect to the anti-spoofing systems. Finally, in Section V-F, we compare all the face verification systems fused with the best performing fusion method and anti-spoofing system.

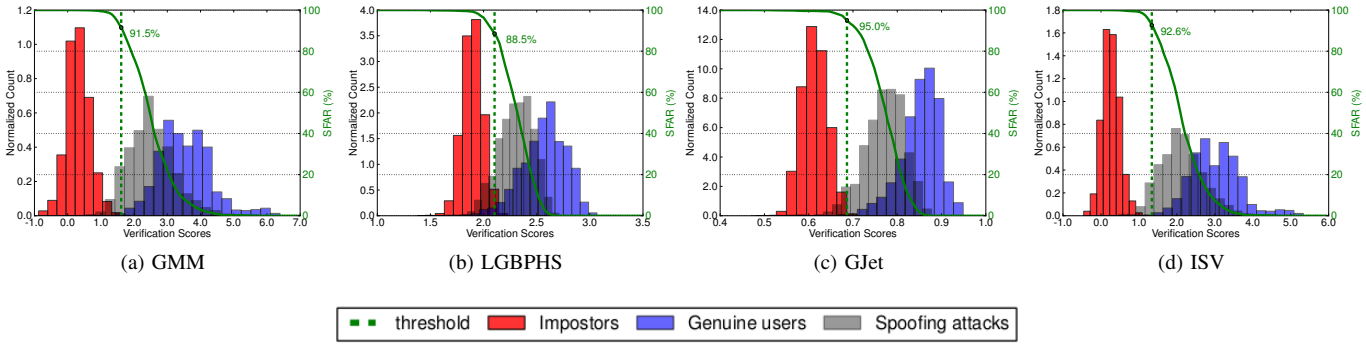
The primary goal of the experiments is to demonstrate the advantages of the EPS framework over other evaluation methodologies and its usefulness in analyzing the performance of biometric verification systems. As an additional result, they provide insights about how fusion affects the systems verification performance and robustness to spoofing and demonstrates which of the fused systems performs the best.

C. Performance of baseline face verification systems

The goal of the first experiment is to assess the performance of the four considered face verification systems in recognizing the identities in Replay-Attack, as well as to estimate their vulnerability to spoofing. In this experiment, they are operating independently, without any protection with an anti-spoofing system. In our analysis, we will compare the conclusions obtained using the evaluation Methodology 2 described in Section III-B, and the ones delivered by EPS framework and EPSC. The score distribution of the four systems are given in Fig. 7.

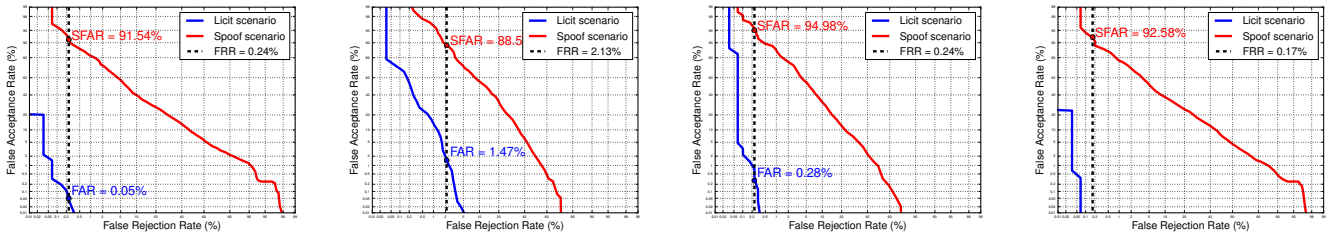
To assess the verification performance of a system using Methodology 2, we consider only the licit scenario. The

[5]http://www.idiap.ch/dataset/replayattack
[6]http://pypi.python.org/pypi/facereclib
[7]http://pypi.python.org/pypi/antispoofing.lbp
[8]http://pypi.python.org/pypi/antispoofing.lbptop
[9]http://pypi.python.org/pypi/antispoofing.motion
[10]http://pypi.python.org/pypi/antispoofing.fusion_faceverif



(a) GMM (b) LGBPHS (c) GJet (d) ISV

Fig. 7: Score distributions of baseline face verification systems. The full green line shows the SFAR as the threshold changes.



(a) GMM (b) LGBPHS (c) GJet (d) ISV

Fig. 8: DET curves for licit and spoof scenario of baseline face verification systems.

vertical lines in Fig. 7 correspond to the thresholds determined in the licit scenario. Using this scenario, we can plot a DET curve, showing the trade-off between FAR and FRR when no spoofing attacks are present. Then, we can consider the spoofing scenario only, and plot an additional DET curve, which shows the trade-off between SFAR and FRR and ignores to the existence of zero-effort impostors. These plots for the four baseline systems are given in Fig. 8.

A decision threshold for such a system is taken at EER on the development set of the licit scenario. Using this threshold, we can compute and report FRR, FAR and SFAR. These values for the four baseline systems are given in Table I.

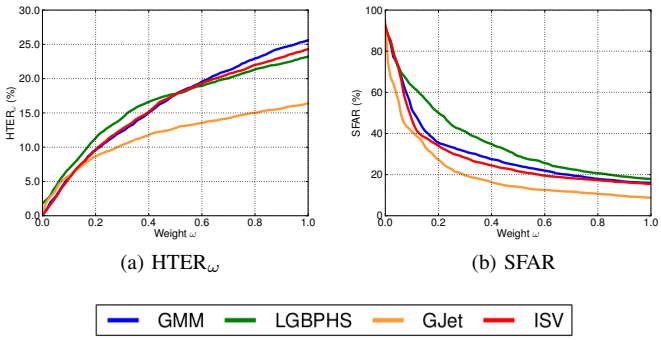
TABLE I: Verification error rates and spoofing vulnerability of baseline face verification systems (in %)

system	FAR	FRR	HTER	SFAR
GMM	0.05	0.24	0.14	91.5
LGBPHS	1.47	2.13	1.8	88.5
GJet	0.28	0.24	0.26	95.0
ISV	0.00	0.17	0.08	92.6

The results show that all the four systems perform well in the verification task. Fig. 7 justifies the results: the score distributions for the genuine users and impostors are almost perfectly separated. However, if we keep the decision threshold selected at EER on the development set for the licit protocol, the systems exhibit a great vulnerability to spoofing of around 90%. The results come with no surprise: as suggested by Fig. 7, the attacks of Replay-Attack appear to be sub-optimal to optimal. Using this evaluation methodology, ISV, with 0.08% of HTER seems to perform the best in the verification task. At the same time, GJet, with 95% of SFAR, appears to be the most vulnerable to spoofing among all the systems. These values are obtained only for a threshold which does not assume any spoofing attacks to be possible.

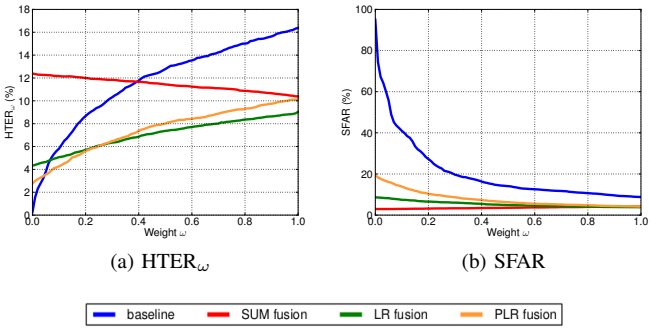
We now proceed with EPS evaluation of the systems. The EPSC given in Fig. 9, report HTER_ω and SFAR for a threshold which considers the relative probability of spoofing attacks, encoded in the parameter ω. Analyzing the EPSC for the four baseline systems, we come to different conclusions. Comparing the HTER_ω values in Fig. 9a, we observe that ISV is best performing in verification only as long as the spoofing attacks appear with a very small probability. After a certain value of ω, GJet shows the best verification performance. The same applies to the vulnerability to spoofing (Fig. 9b): while being the most vulnerable when $\omega \approx 0$, GJet displays the smallest values of SFAR for larger values of ω.

Hence, we can discuss two advantages of EPSC over Methodology 2. Firstly, it overcomes the exclusiveness in analyzing only zero-effort impostors or spoofing attacks at a time of Methodology 2. The HTER_ω summarizes all the three error rates (FRR, FAR and SFAR) into a single value, combining them based on the prior of each of the input classes. Secondly, it rectifies the bias that Methodology 2 demonstrates by neglecting the spoofing attacks that may appear. Although this may increase the value of HTER_ω (EPSC is usually ascending for HTER_ω), it is going to greatly improve the systems vulnerability to spoofing (EPSC is descending for SFAR), especially in condition where spoofing attacks are



(a) HTER$_\omega$ (b) SFAR

Fig. 9: EPSC to compare baseline face verification systems



(a) HTER$_\omega$ (b) SFAR

Fig. 10: EPSC to compare fusion methods: GJet baseline fused with LBP-based anti-spoofing system

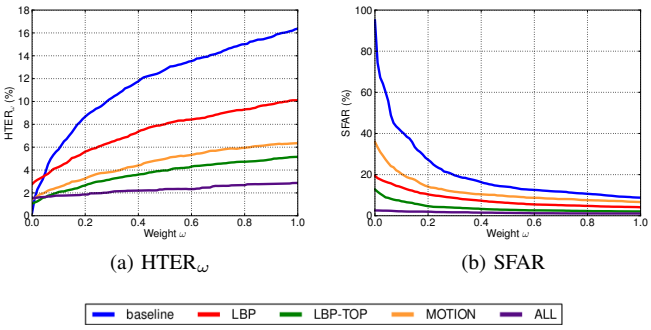
highly probable. Finally, by selecting an *a priori* threshold, EPSC allows to objectively compare several systems on the same figure.

D. Comparison of fusion methods

In our second experiment, we employ EPSC to compare different methods for fusion of verification and anti-spoofing systems and how they affect the performance of the baseline face verification systems. The reported EPSC in Fig. 10 corresponds to the best performing system in the experiment in Section V-C, GJet, when fused with the simplest anti-spoofing system based on LBP. Detailed results covering all the other baseline verification systems is given in Appendix E.

The EPSC helps us to choose which system to use depending on the prior of spoofing attacks we expect at input. As can be observed from Figure 10a, when the prior of spoofing attacks is very small ($\omega \approx 0$), the baseline system not fused with an anti-spoofing system performs the best. As the prior for spoofing attacks is small, any of the fusion schemes only undesirably increases HTER$_\omega$. However, if the prior of spoofing attacks is higher, then fusion is necessary to avoid high vulnerability to spoofing. Expectedly, SFAR and HTER$_\omega$ have a trade-off relationship, and the fusion algorithm that reduces SFAR the most, deteriorates HTER$_\omega$ the most as well. For example, SUM fusion notes the most significant drop of SFAR, but also degrades the verification performance the most, leading to highest HTER$_\omega$.

With respect to the overall performance, LR and PLR perform on similar scale. While SUM fusion helps the baseline only for high values of ω, LR and PLR improve the baseline already for low values of ω. If we need to choose a single algorithm based on HTER$_\omega$, then PLR will be the recommended choice for applications where $\omega < 0.2$, and LR otherwise.



(a) HTER$_\omega$ (b) SFAR

Fig. 11: EPSC to compare anti-spoofing systems: GJet baseline fused using PLR fusion

E. Comparison of anti-spoofing systems

The goal of the third experiment is to employ EPSC to compare the different anti-spoofing systems (LBP, LBP-TOP and MOTION) when fused with baseline face verification systems. Led by the observations of [48] and [49] that using multiple complementary spoofing counter-measures is more effective than a single one, we also attempted to fuse the verification systems with ALL the available anti-spoofing systems at once. We present the results on GJet using PLR fusion, as one of the best performing fusion methods in the experiment in Section V-D. The results for the rest of the baseline systems are given in Appendix E.

Fig. 11 shows that, similarly as in the experiment presented in Section V-D, fusion brings better overall system performance than the isolated baseline, unless spoofing attacks are highly improbable ($\omega \approx 0$). When considering only one anti-spoofing system, the presented results are in favor of the LBP-TOP for all the verification systems along the full range of ω. Yet, fusing several anti-spoofing systems further improves the system robustness to spoofing, as well as its HTER$_\omega$.

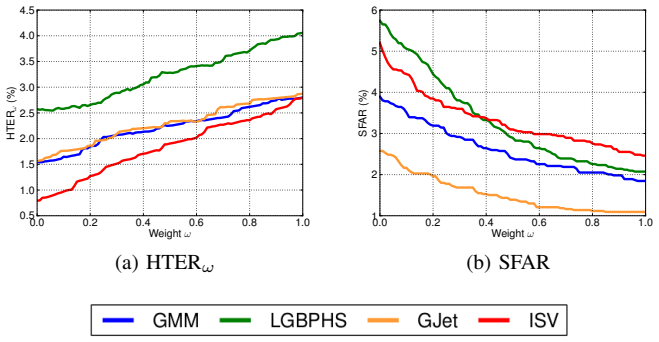
F. Performance of fused systems

In our last experiment, we utilize EPSC to compare the four face verification systems when fused with ALL counter-measures using the PLR fusion scheme. The results are presented in Fig. 12.

The comparison between the EPSC for the baseline (Fig. 9a) and the fused systems (Fig. 12a), confirms that fusion is highly beneficial to the systems' robustness to spoofing. While for some of the baseline systems HTER$_\omega$ increases rapidly with ω and reaches up to 25%, for the fused systems it increases very mildly and does not exceed 4.1%. The major augmentation of robustness to spoofing for the systems after

fusion can be observed by comparing Fig. 9b and Fig. 12b: while unacceptable for the baseline systems for any value of ω, SFAR does not exceed 6% for the fused systems even in the case when spoofing attacks are not considered in the threshold decision process i.e. $\omega = 0$. The benefits of fusing can be also illustrated by the score distribution plots, which are available in Appendix E.

If we summarize both the verification performance and spoofability of the systems into HTER_ω, Fig. 12a suggests that ISV baseline fused with ALL the available anti-spoofing systems performs the best. With AUE value of 0.0184 and HTER_ω varying between 0.8% and 2.7%, ISV is superior over the full range of ω.



(a) HTER_ω (b) SFAR

Fig. 12: EPSC to compare fused systems: PLR fusion with ALL anti-spoofing systems

VI. Conclusions

The spoofing attacks have proven to be a security threat for the biometric verification systems in many modes and the problem of anti-spoofing has been significantly treated in the past few years. However, to apply anti-spoofing in a real-world scenario, it is of importance to make a link between anti-spoofing and biometric verification systems. The alliance of the two will result in a verification system which will hopefully demonstrate higher robustness to spoofing, but probably for the price of modified verification accuracy.

In the traditional setup, the verification systems are evaluated using the well-established metrics for binary classification systems. Their vulnerability to spoofing is rarely reported. When the spoofing attacks are acknowledged as a possible danger, the verification system loses its binary nature and has to cope with three input classes: genuine users, zero-effort impostors and spoofing attacks. Inevitably, this introduces a new definition for the verification systems and a necessity for adjusted evaluation methodology.

The main concern of this paper is to find an appropriate way to evaluate verification systems under spoofing attacks. Several attempts already exist and are thoroughly covered in this paper. Among their most crucial disadvantages is their biased behavior of ignoring the spoofing attacks in the threshold decision process. This leads to unnecessary high vulnerability to spoofing.

This paper proposes a novel evaluation methodology, which objectively assumes that both the zero-effort impostors and spoofing attacks need to be considered in the threshold decision process with a part that reflects the prior probability among all the misuses of the system. Furthermore, the methodology accounts for the application-dependent cost of the error rates associated with the positive and the negative classes. The proposed framework, EPS, and the corresponding curve report on the verification performance and the spoofability of the verification systems using a single measure, called $\text{WER}_{\omega,\beta}$. It does so *a priori*, setting the threshold with no knowledge on the test set in the development phase.

The power of the EPS framework and EPSC is demonstrated by evaluating four state-of-the-art verification systems in the face mode, before and after they are fused with an anti-spoofing system. The EPSC allows for objective comparison of the systems depending on the prior probability of the spoofing attacks or the cost of the error rates and helps decide which combination of verification system, anti-spoofing system and fusion method to use for a given application.

The evaluation concepts covered in this paper are general and could be employed for other verification systems and modes. For this purpose, the implementation of the proposed evaluation framework is available as free software and can be downloaded at http://pypi.python.org/pypi/antispoofing.evaluation.

Acknowledgments

The authors would like to thank the FP7 European TABULA RASA (257289) and BEAT (284989) projects for their financial support.

References

[1] S. Sarkar and Z. Liu, *Handbook of Biometrics*. Springer-Verlag, 2008, ch. Gait recognition. 1

[2] P. Kasprowski and J. Ober, *Lecture Notes in Computer Science*. Springer-Verlag, 2004, ch. Eye Movements in Biometrics. 1

[3] S. Marcel and J. d. R. Millán, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation," *IEEE TPAMI, Special Issue on Biometrics*, 2007. 1

[4] A. K. Jain and A. Ross, *Handbook of Biometrics*. Springer-Verlag, 2008, ch. Introduction to Biometrics. 1, 4

[5] A. J. Mansfield, J. L. Wayman, A. Dr, D. Rayner, and J. L. Wayman, "Best practices in testing and reporting performance," 2002. 1, 3, 4, 5

[6] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artifical "gummy" fingers on fingerprint systems," in *SPIE Proceedings: Optical Security and Counterfeit Deterrence Techniques*, vol. 4677, 2002. 1, 5

[7] F. Alegre, R. Vipperla, N. Evans, and B. Fauve, "On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 36–40. 1, 5

[8] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., 2006. 1

[9] A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Biometrics*. Springer-Verlag, 2008, ch. Introduction to multibiometrics. 1

[10] D. Osten, H. M. Carim, M. R. Arneson, and B. L. Blan, "Biometric, personal authentication system," Patent US Patent #5,719,950, 02 17, 1998. 1

[11] S. Schuckers, *Encyclopedia of Biometrics*. Springer-Verlag, 2009, ch. Liveness Detection: Fingerprint, pp. 924–931. 1, 4

[12] E. Marasco, P. Johnson, C. Sansone, and S. Schuckers, "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*, 2011, pp. 309–318. 2, 5, 9

[13] E. Marasco, Y. Ding, and A. Ross, "Combining match scores with liveness values in a fingerprint verification system," in *5th IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2012. 2, 9

[14] I. Chingovska, A. Anjos, and S. Marcel, "Anti-spoofing in action: joint operation with a verification system," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Biometrics*, Jun. 2013. 2, 8, 9

[15] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recognition Journal*, vol. 39. 2

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York Inc., 2001. 2

[17] Y. M. Lui, D. Bolme, P. Phillips, J. Beveridge, and B. Draper, "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*, 2012, pp. 9–16. 2

[18] E. Bailly-Baillire, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Marithoz, J. Matas, K. Messer, F. Pore, and B. Ruiz, "The BANCA database and evaluation protocol," in *In Proc. Int. Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA03*, 2003. 2, 3

[19] A. Martin and M. Przybocki, "The NIST 1999 speaker recognition evaluation - an overview," 2000. 3

[20] M. Przybocki, A. Martin, and A. Le, "NIST speaker recognition evaluation chronicles - part 2," in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, June 2006. 3

[21] A. Martin, G. Doddington, T. Kamm, and M. M. Ordowski, "The DET curve in assessment of detection task performance," in *Eurospeech*, 1997, pp. 1895–1898. 3

[22] S. Bengio, J. Mariéthoz, and M. Keller, "The expected performance curve," in *International Conference on Machine Learning, ICML, Workshop on ROC Analysis in Machine Learning*, 2005. 3

[23] "Information Technology Vocabulary Biometrics," 2012. 4

[24] M. Wagner and G. Chetty, *Encyclopedia of Biometrics*. Springer-Verlag, 2009, ch. Liveness Assurance in Face Authentication, pp. 924–931. 4

[25] P.Johnson, R. Lazarick, E. Marasco, E. Newton, A. Ross, and S. Schuckers, "Biometric liveness detection: Framework and metrics," in *International Biometric Performance Conference*, 2012. 4

[26] A. Adler and S. Schuckers, *Encyclopedia of Biometrics*. Springer-Verlag, 2009, ch. Security and Liveness, Overview, pp. 1146–1152. 4

[27] P. A. Johnson, B. Tan, and S. Schuckers, "Multimodal fusion vulnerability to non-zero (spoof) imposters," in *IEEE International Workshop on Information Forensics and Security*, 2010. 5

[28] J. Galbally-Herrero, J. Fierrez-Aguilar, J. D. Rodriguez-Gonzalez, F. Alonso-Fernandez, J. Ortega-Garcia, and M. Tapiador, "On the vulnerability of fingerprint verification systems to fake fingerprints attacks," in *IEEE International Carnahan Conference on Security Technology*, 2006, pp. 169–179. 5

[29] J. Galbally, R. Cappelli, A. Lumini, G. G. de Rivera, D. Maltoni, J. Fiérrez, J. Ortega-Garcia, and D. Maio, "An evaluation of direct attacks using fake fingers generated from iso templates," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 725–732, 2010. 5

[30] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, 2011, pp. 1–8. 5

[31] V. Ruiz-Albacete, P. Tome-Gonzalez, F. Alonso-Fernandez, J. Galbally, J. Fierrez, and J. Ortega-Garcia, "Direct attacks using fake images in iris verification," in *Proc. COST 2101 Workshop on Biometrics and Identity Management, BIOID*. Springer, May 2008, pp. 181–190. 5

[32] R. N. Rodrigues, L. L. Ling, and V. Govindaraju, "Robustness of multimodal biometric fusion methods against spoofing attacks," *Journal of Visual Languages and Computing*, vol. 20, no. 3, pp. 169–179, 2009. 5, 9

[33] Z. Akhtar, G. Fumera, G.-L. Marcialis, and F. Roli, "Robustness evaluation of biometric systems under spoof attacks," in *16th International Conference on Image Analysis and Processing*, pp. 159–168. 5

[34] ——, "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. 5, 9

[35] A. Anjos *et al.*, "Bob: a free signal processing and machine learning toolbox for researchers," in *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. ACM Press, Oct. 2012. 8

[36] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proceedings of the 11th International Conference of the Biometrics Special Interes Group*, 2012. 9

[37] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proc. European Conference on Computer Vision (ECCV)*, ser. LNCS 6316. Springer, 2010, pp. 504–517. 9

[38] Z. Zhiwei, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. IAPR Int. Conf. on Biometrics (ICB)*, 2012, pp. 26–31. 9

[39] F. Cardinaux, C. Sanderson, and S. Marcel, "Comparison of mlp and gmm classifiers for face verification on xm2vts," in *Proceedings of the 4th International Conference on AVBPA*, University of Surrey, Guildford, UK, 2003. 9

[40] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01*, ser. ICCV '05. IEEE Computer Society, 2005, pp. 786–791. 9

[41] L. Wiskott, J.-M. Fellous, N. Krger, and C. V. D. Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis And Machine Inteligence*, vol. 19, pp. 775–779, 1997. 9

[42] M. Günther, D. Haufe, and R. P. Würtz, "Face recognition with disparity corrected Gabor phase differences," in *Artificial Neural Networks and Machine Learning*, ser. Lecture Notes in Computer Science, vol. 7552. Springer Berlin, 2012, pp. 411–418. 9

[43] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *International Joint Conference on Biometrics*, 2011. 9

[44] M. Günther, R. Wallace, and S. Marcel, "An open source framework for standardized comparisons of face recognition algorithms," in *Computer Vision - ECCV 2012. Workshops and Demonstrations*, ser. Lecture Notes in Computer Science, vol. 7585. Springer Berlin, 2012, pp. 547–556. 9

[45] M. M. Chakka *et al.*, "Competition on counter measures to 2-d facial spoofing attacks," in *Proceedings of IAPR IEEE International Joint Conference on Biometrics (IJCB), Washington DC, USA*, 2011. 9

[46] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Lbp-top based countermeasure against face spoofing attacks," in *International Workshop on Computer Vision With Local Binary Pattern Variants - ACCV*, 2012, p. 12. 9

[47] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *International Joint Conference on Biometrics 2011*, 2011. 9

[48] J. Komulainen, A. Anjos, A. Hadid, M. Pietikainen, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," 2013. 11

[49] I. Chingovska *et al.*, "The 2nd competition on counter measures to 2d face spoofing attacks," in *International Conference of Biometrics 2013*, 2013. 11

Biometrics Evaluation under Spoofing Attacks

Ivana Chingovska, André Anjos, Sébastien Marcel

APPENDIX A
NOTES ON COMMON TERMINOLOGY FOR EVALUATION METRICS IN BIOMETRICS AND ANTI-SPOOFING

a) Error rates for evaluation of biometric systems: In the context of a binary classification system, we introduce False Negative Rate (FNR) and False Positive Rate (FPR) as error rates associated with number of wrongly classified positive and negative samples respectively. In the context of a biometric verification system, the typically used terms are False Match Rate (FMR) and False Non-Match Rate (FNMR), as well as False Acceptance Rate (FAR) and False Rejection Rate (FRR). However, as suggested in [1], FRR and FAR are not synonymous with False Non-Match Rate (FNMR) and False Match Rate (FMR). FNMR and FMR are used at a level of a single sample-to-model comparison, whereas FRR and FAR are used at a transaction level, where a transaction includes all the allowed attempts of a user to be recognized by the system. Hence, in general, FAR and FRR depend on FMR and FNMR, but also on error rates like Failure to Acquire (FTA), Binning Error Rate (BER) and Penetration Rate (PR). Furthermore, FAR and FRR refer to the claim of the user, and this claim is different for a biometric verification and biometric identification system. However, in the scope of our work, we are considering only biometric verification systems and we do our evaluation in a pre-collected database, thus precluding error rates like FTA, BER and PR. In such circumstances, which, as stated in [1], are typical for technology evaluation, FAR and FRR are equivalent to FMR and FNMR. Therefore, in our manuscript, we adhere to the terms FAR and FRR. This terminology is also accepted, for example, in [2].

b) Error rates for evaluation of anti-spoofing systems: Table I gives the most common terminology and synonyms for error rates in evaluating anti-spoofing systems.

TABLE I: Typically used error rates for anti-spoofing systems and their synonyms.

Error rate	Acronym	Synonyms
False Positive Rate	FPR	False Acceptance Rate (FAR), False Spoof Acceptance Rate [3], False Living Rate (FLR) [4]
False Negative Rate	FNR	False Rejection Rate (FRR), False Alarm Rate [5], False Live Rejection Rate [3], False Fake Rate (FFR) [4]
True Positive Rate	TPR	True Acceptance Rate
True Negative Rate	TNR	True Rejection Rate, detection rate [5], [6], [7], detection accuracy [8]
Half Total Error Rate	HTER	Average Classification Error (ACE) [4]

c) Error rates for evaluation of biometric verification systems under spoofing attacks: Table II gives the most common error rates in evaluation of biometric verification systems under spoofing attacks. It contains error rates reported when the system is evaluated only considering one negative class (either zero-effort impostors or spoofing attacks, resulting in licit or spoof scenario, respectively), or both of them.

TABLE II: Typically used error rates for biometric verification systems under spoofing attacks and their synonyms.

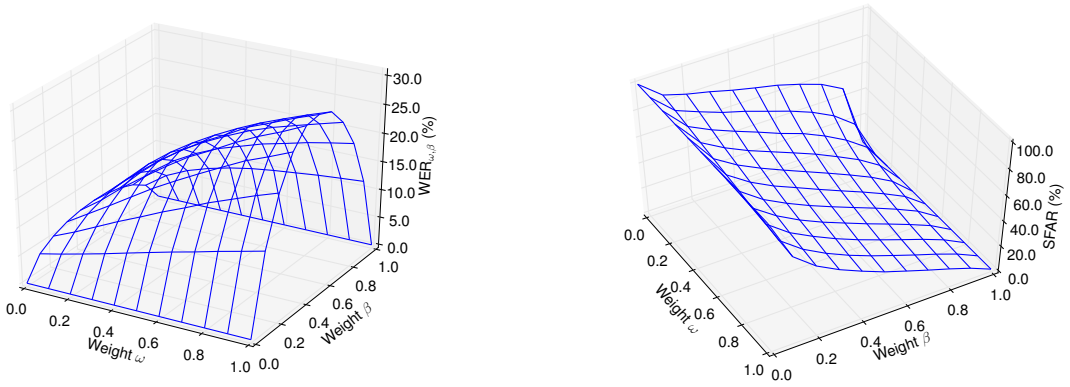
Error rate	Acronym	Negative class	Synonyms
False Negative Rate	FNR	any	False Rejection Rate (FRR), False Non-Match Rate [9], [3], Pmiss [10])
		both	Global False Rejection Rate (GFRR) [3]
True Positive Rate	TPR	any	True Acceptance Rate, Genuine Acceptance Rate [11], [12]
False Positive Rate	FPR	zero-effort impostors	False Acceptance Rate (FAR), False Match Rate [9], [3], Pfa [10]
		spoofing attacks	False Acceptance Rate (FAR) [13], Spoof False Acceptance Rate [14], Liveness False Acceptance Rate [15], Success Rate [16], Attack Success Rate [9]
		both	System False Acceptance Rate (SFAR) [15], Global False Acceptance Rate (GFAR) [3]

For a more general framework, where the system is specialized to detect any kind of suspicious or subversive presentation of samples, be it a spoofing attack, altered sample or artifact, [11] 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.

Ivana Chingovska is with Idiap Research Institute and Ecole Polytechnique Fédéerale de Lausanne, Switzerland, e-mail: ivana.chingovska@idiap.ch
André Anjos and Sébastien Marcel are with the Idiap Research Institute, Switzerland, e-mails: {andre.anjos, sebastien.marcel}@idiap.ch

APPENDIX B
EPS FRAMEWORK: 3D PLOT OF ERROR RATES WITH RESPECT TO THE PARAMETERS

If we parameterize $\text{WER}_{\omega,\beta}$ by the two parameters, we are going to obtain a 3D surface, which, for a hypothetical biometric verification system is shown in Fig. 1. Using this plot, we can clearly infer on the expected error rates depending on the parameters' values or range of values which are of interest. However, the visualization of two or more 3D plots on the same figure is difficult and not convenient for comparative analysis of systems.



(a) $\text{WER}_{\omega,\beta}$ (b) SFAR

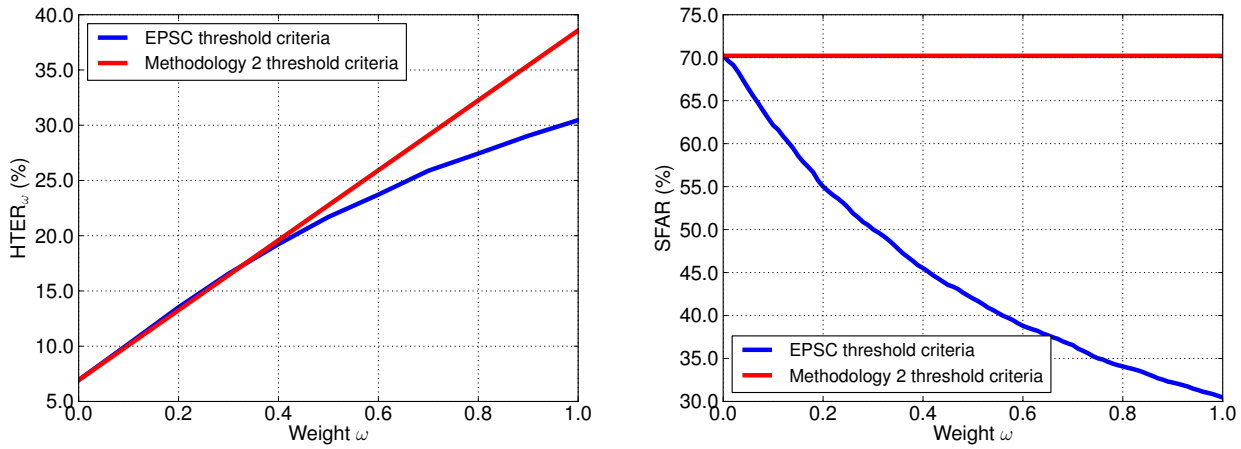
Fig. 1: 3D plot of $\text{WER}_{\omega,\beta}$ and SFAR computed using EPS framework for a hypothetical biometric verification system

APPENDIX C
COMPARISON OF EPSC WITH METHODOLOGY 2

To support the assertion that consideration of the spoofing attacks is necessary when determining the decision threshold, we compare EPSC with Methodology 2 described in Section III-B. For a hypothetical verification system, Fig. 2 plots the error rates HTER_ω and SFAR as they are defined in Section IV. For EPSC, the decision threshold is determined using the criteria given in Eq.5 of the manuscript. For Methodology 2, it does not depend on the parameter ω and is determined using the licit scenario only. In both cases we fix the parameter $\beta = 0.5$.

Fig. 2: Comparison of error rates for EPSC and Methodology 2 (hypothetical biometric verification system)

(a) HTER_ω (b) SFAR

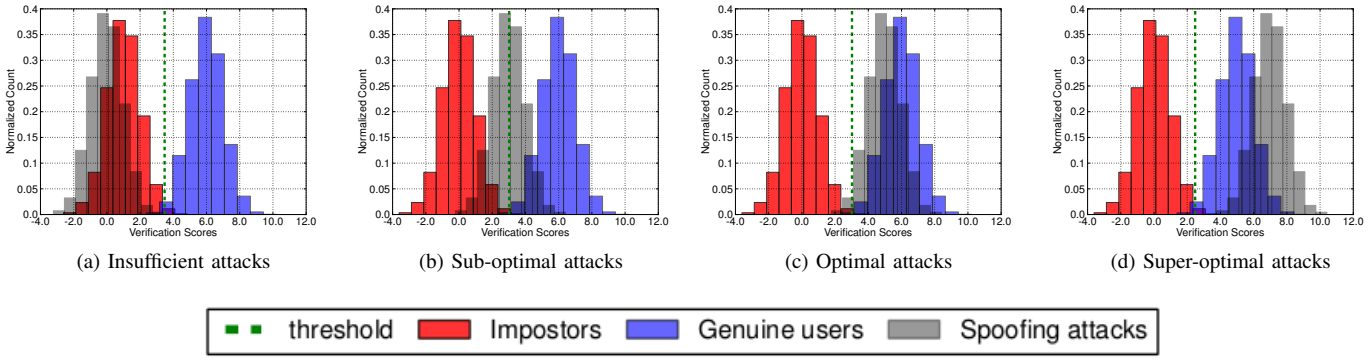


Both EPSC and Methodology 2 give the same results when $\omega = 0$ i.e. when the verification system is not under spoofing attacks. However, as soon as the spoofing attacks get even a small weight $\omega > 0$, the vulnerability of the system under Methodology 2 remains very high, while EPSC quickly adapts the threshold and achieves much better robustness to spoofing (Fig. 2b). This is also reflected to the HTER_ω: EPSC notes more mild increase of HTER_ω as the weight of the spoofing attacks increases (Fig. 2a).

APPENDIX D
CATEGORIES OF SPOOFING ATTACKS: EPSC CASE STUDY

When reporting on the performance and spoofability of a verification system, it is usually done with respect to a certain dataset. To be accounted for robust to spoofing with respect to a dataset, the system needs to gives score distributions as illustrated in Fig. 3a. This means that, with respect to this system, the attacks are in the insufficient category. To be accounted as vulnerable to spoofing, the system needs to give score distributions as in Fig. 3c of Fig. 3d. In such a case, the attacks are in the optimal or super-optimal category with respect to that system.

The success of the attacks in spoofing the system primarily depends on two factors: their quality and the system design. Spoofing attacks of low quality, which do not look realistic and which contain a lot of noise and artifacts may be insufficient and fail to pass the verification system. Sub-optimal attacks are probably the most common: they are realistic enough to be verified as the claimed identity, but their score is low due to the presence of artifacts. Optimal and super-optimal attacks look more realistically and contain less artifacts, and hence their production may require user cooperation, expensive materials and high-level skills. Hence, they are usually difficult to create.



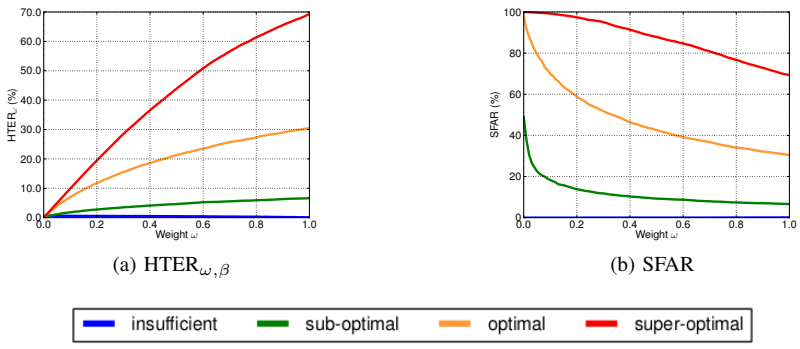
(a) Insufficient attacks (b) Sub-optimal attacks (c) Optimal attacks (d) Super-optimal attacks

Fig. 3: Score distributions of 4 categories of spoofing attacks (hypothetical biometric verification system)

Fig. 4 illustrates the appearance of EPSC for the four hypothetical verification systems in Fig. 3 giving the four different categories of spoofing attacks. The parameter $\beta = 0.5$ is fixed, while the parameter ω varies.

The general trend for all the cases is increasing HTER_ω as ω increases, but at the same time decreasing SFAR. Certainly, this is a result of the security cautions taken by EPS framework by accounting on the spoofing attacks when deciding on the decision threshold. However, there are significant differences in the appearance of EPSC for the systems with different categories of attacks. For a system which is already robust to spoofing, i.e. puts the attacks in the insufficient category, both HTER_ω and SFAR are relatively constant (blue curves). For systems relatively robust to spoofing, i.e. putting the attacks in the sub-optimal category, the increase of HTER_ω is mild, while the decrease of SFAR is sharp (green curves). On the other hand, for systems vulnerable to spoofing, the increase of HTER_ω is sharp, while the decrease of SFAR is mild (yellow and red curves).

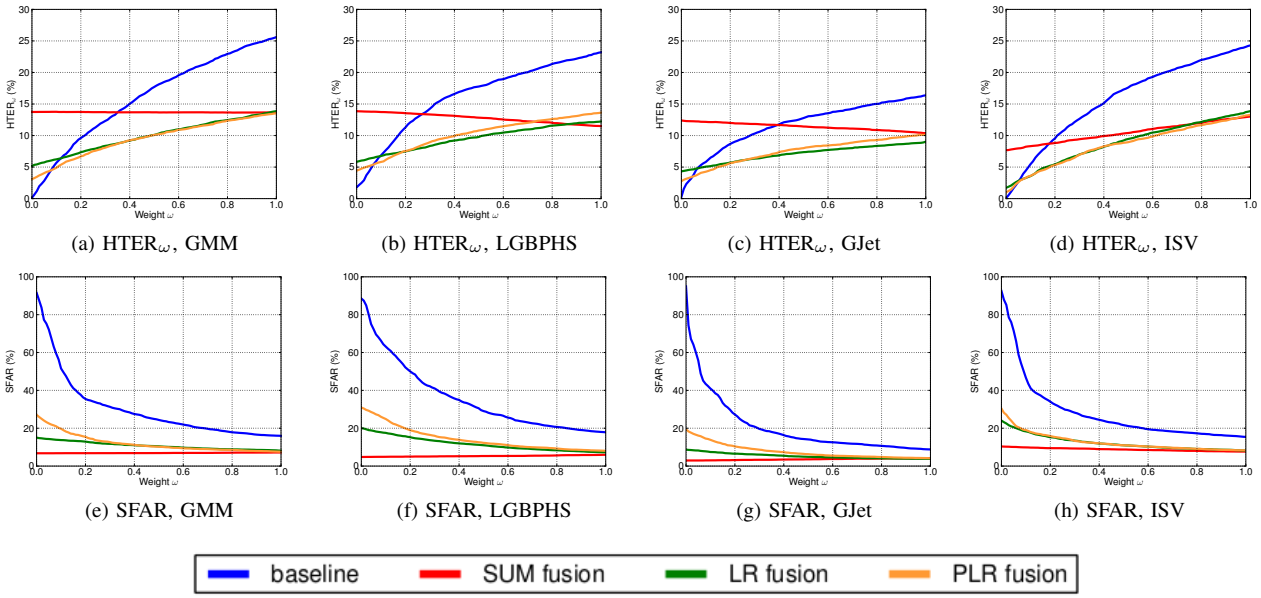
By visually analyzing and comparing the incline of the EPSC curves for two systems, we can infer which one has higher robustness to spoofing.



(a) $\text{HTER}_{\omega,\beta}$ (b) SFAR

Fig. 4: EPSC for different categories of spoofing attacks

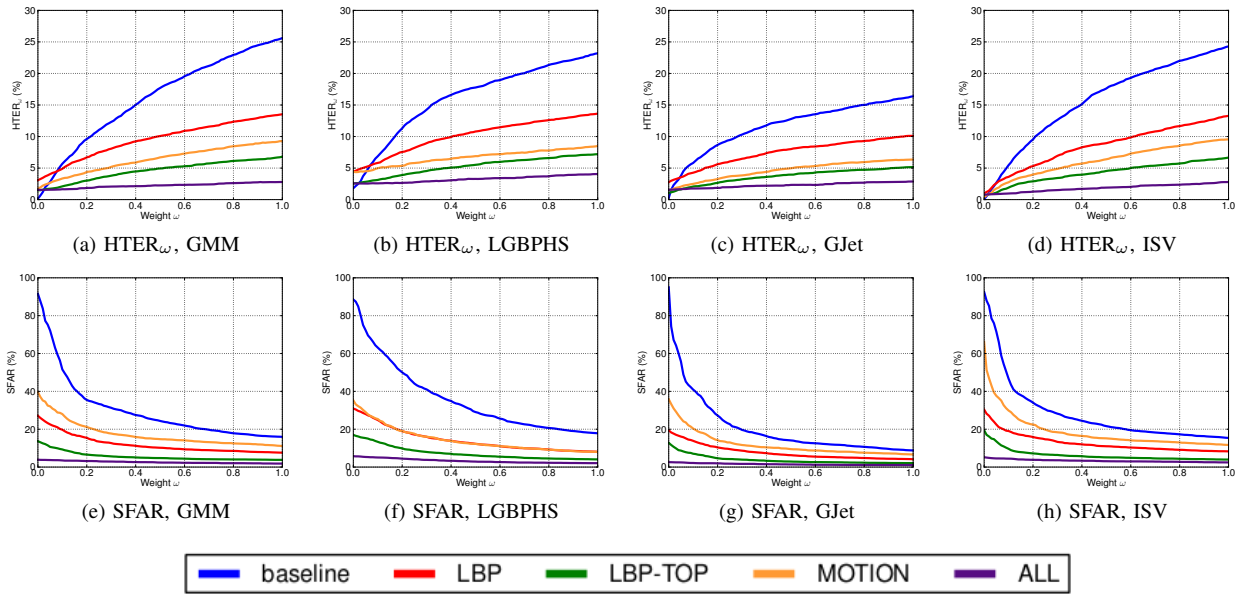
APPENDIX E
EXPERIMENTAL RESULTS



(a) HTER$_\omega$, GMM (b) HTER$_\omega$, LGBPHS (c) HTER$_\omega$, GJet (d) HTER$_\omega$, ISV

(e) SFAR, GMM (f) SFAR, LGBPHS (g) SFAR, GJet (h) SFAR, ISV

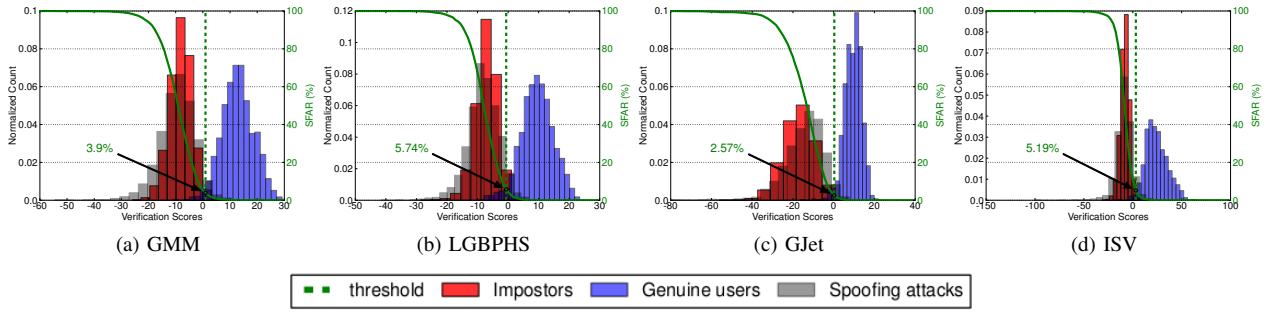
Fig. 5: EPSC to compare fusion methods: the four baseline face verification systems fused with LBP-based anti-spoofing system. The four columns correspond to the four baselines: GMM, LGBPHS, GJet and ISV, respectively. The top row gives the EPSC for HTER$_\omega$, while the bottom row the EPSC for SFAR.



(a) HTER$_\omega$, GMM (b) HTER$_\omega$, LGBPHS (c) HTER$_\omega$, GJet (d) HTER$_\omega$, ISV

(e) SFAR, GMM (f) SFAR, LGBPHS (g) SFAR, GJet (h) SFAR, ISV

Fig. 6: EPSC to compare anti-spoofing systems: the four baseline face verification systems fused using PLR fusion. The four columns correspond to the four baselines: GMM, LGBPHS, GJet and ISV, respectively. The top row gives the EPSC for HTER$_\omega$, while the bottom row the EPSC for SFAR.



(a) GMM (b) LGBPHS (c) GJet (d) ISV

Fig. 7: Score distributions of fused systems: PLR fusion with ALL anti-spoofing systems. The full green line shows the SFAR as the threshold changes.

REFERENCES

[1] A. J. Mansfield, J. L. Wayman, A. Dr, D. Rayner, and J. L. Wayman, "Best practices in testing and reporting performance," 2002. 1

[2] A. K. Jain, P. Flynn, and A. A. Ross, Eds., *Handbook of Biometrics*. Springer-Verlag, 2008. 1

[3] E. Marasco, Y. Ding, and A. Ross, "Combining match scores with liveness values in a fingerprint verification system," in *5th IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2012. 1

[4] J. Galbally, F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "A high performance fingerprint liveness detection method based on quality related features," *Future Gener. Comput. Syst.*, vol. 28, no. 1, pp. 311–321, 2012. 1

[5] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8. 1

[6] W. Bao, H. Li, N. Li, and W. Jiang, "A liveness detection method for face recognition based on optical flow field," in *Image Analysis and Signal Processing, 2009. IASP 2009. International Conference on*, 2009, pp. 233–236. 1

[7] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *Tsinghua Science and Technology*, vol. 14, no. 6, pp. 685–690, 2009. 1

[8] Z. Zhang, D. Yi, Z. Lei, and S. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 436–441. 1

[9] J. Galbally, R. Cappelli, A. Lumini, G. G. de Rivera, D. Maltoni, J. Fiérrez, J. Ortega-Garcia, and D. Maio, "An evaluation of direct attacks using fake fingers generated from iso templates," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 725–732, 2010. 1

[10] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, 2011, pp. 1–8. 1

[11] P.Johnson, R. Lazarick, E. Marasco, E. Newton, A. Ross, and S. Schuckers, "Biometric liveness detection: Framework and metrics," in *International Biometric Performance Conference*, 2012. 1

[12] R. Rodrigues, N. Kamat, and V. Govindaraju, "Evaluation of biometric spoofing in a multimodal system," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, 2010. 1

[13] J. Galbally-Herrero, J. Fierrez-Aguilar, J. D. Rodriguez-Gonzalez, F. Alonso-Fernandez, J. Ortega-Garcia, and M. Tapiador, "On the vulnerability of fingerprint verification systems to fake fingerprints attacks," in *IEEE International Carnahan Conference on Security Technology*, 2006, pp. 169–179. 1

[14] P. A. Johnson, B. Tan, and S. Schuckers, "Multimodal fusion vulnerability to non-zero (spoof) imposters," in *IEEE International Workshop on Information Forensics and Security*, 2010. 1

[15] A. Adler and S. Schuckers, *Encyclopedia of Biometrics*. Springer-Verlag, 2009, ch. Security and Liveness, Overview, pp. 1146–1152. 1

[16] V. Ruiz-Albacete, P. Tome-Gonzalez, F. Alonso-Fernandez, J. Galbally, J. Fierrez, and J. Ortega-Garcia, "Direct attacks using fake images in iris verification," in *Proc. COST 2101 Workshop on Biometrics and Identity Management, BIOID*. Springer, May 2008, pp. 181–190. 1