IET Biometrics

Special Issue: Face Recognition and Spoofing Attacks

IET Journals
The Institution of Engineering and Technology

Deeply vulnerable: a study of the robustness of face recognition to presentation attacks

ISSN 2047-4938
Received on 8th May 2017
Revised 14th September 2017
Accepted on 26th September 2017
E-First on 16th November 2017
doi: 10.1049/iet-bmt.2017.0079
www.ietdl.org

Amir Mohammadi[1,2] ✉, Sushil Bhattacharjee[1], Sébastien Marcel[1]

[1]*Idiap Research Institute, Centre du Parc, Rue Marconi 19, P.O. Box 592, CH-1920 Martigny, Switzerland*
[2]*École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*
✉ *E-mail: amir.mohammadi@idiap.ch*

Abstract: The vulnerability of deep-learning-based face-recognition (FR) methods, to presentation attacks (PA), is studied in this study. Recently, proposed FR methods based on deep neural networks (DNN) have been shown to outperform most other methods by a significant margin. In a trustworthy face-verification system, however, maximising recognition-performance alone is not sufficient – the system should also be capable of resisting various kinds of attacks, including PA. Previous experience has shown that the PA vulnerability of FR systems tends to increase with face-verification accuracy. Using several publicly available PA datasets, the authors show that DNN-based FR systems compensate for variability between bona fide and PA samples, and tend to score them similarly, which makes such FR systems extremely vulnerable to PAs. Experiments show the vulnerability of the studied DNN-based FR systems to be consistently higher than 90%, and often higher than 98%.

1 Introduction

Progress in face-recognition (FR) technology, from strongly constrained models to fully unconstrained environments, has been enabled by the adoption of successively complex learning paradigms. The first generation of large-scale appearance-based FR systems, such as Eigenfaces [1] and Fisherfaces [2], attempted to model the face variability in a simple linear sub-space. Subsequently, methods such as joint factor analysis, inter-session variability (ISV) modelling [3] and probabilistic linear discriminant analysis [4, 5] were developed to better model variability in face images. Deep-learning-based methods, notably convolutional neural networks (CNN) [6–9], have become very popular in recent years, due to their near-perfect recognition accuracy on unconstrained datasets such as 'labelled faces in the wild' (LFW) [10].

An FR system may be used in two kinds of applications: *face identification* or *face verification*. Face identification is a one-to-many problem, where the face image to be identified is tested against all previously enrolled identities, to see if it matches any of the known identities. In face-verification systems, a claimed-identity is provided along with the input face image, and the problem is simply to verify that the input image corresponds to the claimed-identity. In this paper, the terms *recognition* and *verification* have been used interchangeably, and refer to the use of an FR system in verification mode.

FR systems should not only have very high accuracy but should also be robust to attacks. For example, in access-control applications, high FR accuracy in unconstrained environments is not a significant asset, when not accompanied with resistance to *presentation attacks* (PA), also known as spoof attacks. A face PA is said to have occurred when a face biometric sample is presented to the camera of an FR system "with the intention of interfering with the operation of biometric recognition" [11]. For example, person A may attack an FR system by claiming to be an enrolled client, B, and presenting a printed photo of the person B to the camera. Other examples of PAs are: digital photos or videos displayed on an electronic screen, face sketches and make-up [12].

Most high-accuracy FR systems today rely on deep-learning methods. Indeed, deep-learning-based FR systems are already being deployed in commercial face-verification applications [8]. In this context their vulnerability to PAs becomes an important factor in the trustworthiness of the entire face-verification process. Several studies [6–9] have explored the capacity of deep-learning-based FR systems to handle variations in pose, illumination and scale, that is challenges related to variability of face-biometric samples of a single client. Karahan *et al.* [13] studied the fragility of CNN-based FR systems when confronted with image degradations such as blurring, occlusion, compression-artefacts and colour distortion. Robustness to such degradations is necessary for FR systems operating in relatively unconstrained environments. For face-verification systems functioning under controlled conditions, the vulnerability to PAs is a much more significant concern.

The European research project TABULA RASA [14] indicated a positive correlation between the efficacy of FR systems and their vulnerability to PAs. In other words, FR methods with higher face-verification accuracy tended to be more vulnerable to PAs. Given that FR systems, especially the popular CNN-based FR methods have not been explicitly trained for presentation attack detection (PAD), one would intuitively expect such systems to be vulnerable to PAs. To our knowledge though, the vulnerability of deep learning-based FR systems to PAs has not been studied in detail.

In this paper, we present a large-scale empirical study of the vulnerability to PAs of five recent FR systems, including three recent CNN-based methods. Specifically, we compare the vulnerability of the following systems:

- three CNN-based FR systems, namely:
  - VGG-Face [7]
  - LightCNN [15]
  - FaceNet [16]
- ISV modelling [3]
- ROC-software development kit (SDK) from rank-one computing (company website: www.rankone.io).

The main contribution of this paper is empirical evidence to support the claim that the CNN-based FR method is extremely vulnerable to PAs. Our experiments, using three recent PA datasets, show that other highly rated FR methods are also very vulnerable to PAs. Another significant contribution of this work is that all the datasets, protocols and source-code for the experiments reported in the paper will be made publicly available on the web. This will allow other researchers to reproduce the results presented here.

A brief review of previous vulnerability studies for FR methods is presented in Section 2. This discussion justifies the present work and helps us to offset our results from those in previous published works. The FR methods included in our study are described in Section 3. The experiments reported in this paper have been conducted on publicly available datasets, with well-defined protocols for FR and face PAD experiments. In Section 4, we describe the datasets, the kinds of PAs represented in each dataset, and the protocols for training and testing FR offered in each dataset. Experimental results are summarised and discussed in Section 5, and concluding remarks are presented in Section 6.

2 Related work

There have been very few studies specifically exploring the vulnerability of FR systems to different kinds of attacks. Duc *et al.* [17] have investigated the vulnerability of FR-based login on computers (specifically Lenovo, Toshiba and Asus products) to PAs. Kose *et al.* [18] present a study of the vulnerability of FR methods to 3D-mask attacks. They evaluate two different FR methods – one designed specifically for FR using 3D data [19], and a generic approach to FR using local binary patterns-based face-image comparison. Hadid [20] discusses the vulnerability of the parts-based Gaussian mixture model (GMM) FR method to PAs. In a recently published work, Scherhag *et al.* [21] have studied the vulnerability of FR methods to morphed-face attacks. Here attacks are constructed by morphing face images from two enrolled identities, and the challenge is to see if the two enrolled identities can both be successfully spoofed using the same morphed image. Both FR methods tested in this work – VeriFace SDK (website: www.neurotechnology.com/verilook.html) [a commercial off-the shelf (COTS) FR product from Neurotechnology] and OpenFace [22] [an open-source, pre-trained deep neural network (DNN)-based FR system] – are shown to be highly susceptible to such attacks [21].

Although deep-learning-based FR systems have attracted considerable attention, both academic and commercial, as far as we are aware, no previous study has investigated the vulnerability of CNN-based FR systems using a variety of PAs. Most of the studies cited above have been performed on FR methods that rely on hand-crafted features. The study by Scherhag *et al.* [21], which included a DNN-based FR system (OpenFace), was done in the restricted context of a single class of attacks based on morphed images. Such attacks are expected to have a very narrow range of application.

In this paper, we focus on the vulnerability of CNN-based FR methods to PAs. This study includes three publicly available pre-trained CNN-FR models. Our experiments demonstrate that all these FR methods are consistently highly vulnerable to several classes of PAs. For comparison, we have also included the ISV modelling method, which relies on hand-crafted features, and ROC-SDK – a COTS FR product – in this study. We use four publicly available FR and PA datasets to estimate the vulnerability of the five FR methods to a variety of PAs. Our study has been motivated by the hypothesis that although CNN-FR systems outperform FR methods based on hand-crafted features, they are also more vulnerable to PAs. This assertion makes intuitive sense to researchers in face biometrics, given that the FR methods have not been explicitly trained to detect PAs. However, this is the first large-scale study to empirically quantify the vulnerability of these FR methods.

3 FR systems included in this study

A typical FR system functions in three phases: training, enrolment and probing. In the training phase, a background model, assumed to broadly represent the space of face images, is constructed using training data. In the enrolment phase, the FR system generates templates for the given enrolment samples, which are then stored in the gallery. In the probing (operational) phase, the FR system is presented with a probe image and a claimed identity. A template is created for the given probe sample, and is compared with the set of enrolled templates associated with the claimed identity. The result of the comparison is a score, which is then thresholded to produce a decision (accept or reject).

To mitigate the influence of variations on the actual FR process, the raw input image is usually preprocessed, to extract sub-images representing individual faces. Geometric and colour transforms may also be applied to the extracted face images, depending on the requirements of the specific FR method. The result of the pre-processing stage is a *normalised* face image, of predefined size and scale, that may be processed by a FR system. Before describing the different FR methods, we explain the pre-processing steps applied to normalise the input face images.

3.1 Face image normalisation

Several FR systems expect the input to be a grey-level image. Therefore, our first pre-processing step is to convert the input RGB image to a grey-level image (essentially, the Y component of the YUV domain representation of the image).

The next step is to locate each face represented in the input image. The ROC-SDK takes a full frame image as input. For the other FR methods in our study (CNN-FR, and ISV modelling), the input is a suitably cropped face image. There are two ways to accomplish the face-localisation step explicitly:

- provide additional information, such as annotations for specific facial landmarks, that the FR system can use to extract a normalised face region or
- provide as input a cropped and normalised face image.

In our experiments, the normalised face region is extracted using annotations identifying the centre of each eye. Imposing the constraints that the straight-line joining the two eye centres should be horizontal, and should have a predefined length, an affine transform can be used to extract a normalised face image of fixed size from the given input image. Some face-biometrics test datasets include annotations for the eye locations. For datasets where this information is not explicitly provided, we use a two-step process to extract the eye positions. First, a boosted classifier, trained to detect faces, is used to localise the face region [23]. Next, facial landmarks are extracted from the face region using the *flandmarks* method [24]. This method provides pixel coordinates for seven facial landmarks, including the two corners of each eye. The pixel location for the centre of each eye can be interpolated from these landmarks.

Note that for the VGG-Face and LightCNN networks, the input images are normalised colour (RGB) images, whereas the FaceNet and ISV modelling method expect normalised grey-level face images of fixed 2D shape.

The various FR methods are discussed in the following sections. Specific implementation details, such as the size of the normalised face image, tunings of the hyperparameters and so on are provided in Section 5.

3.2 CNN-based systems

CNNs [25, 26], a class of DNNs, have been shown to be extremely accurate in FR tasks. For FR applications, CNNs are usually trained for face identification, using face images as input, and the set of identities to be recognised as output. The last few layers of the network, including the output layer, are typically fully connected (*fc* for short) layers. These layers may be seen as the classifier stage of the network, whereas the preceding (convolutional and pooling) layers may be considered to constitute the feature-extraction stage. Although CNNs are typically trained end to end for classification or regression tasks, they are often also used as feature-extraction tools. The terms *representation* and *embedding* are used interchangeably, to denote the outputs of the various layers of a deep network. Representations generated by a specific layer of a pre-trained DNN may be used as templates (feature vectors) representing the corresponding input images. These templates may be subsequently used to train a classifier, or to compare the respective input images using appropriate similarity measures.

Taigman *et al.* [6] have proposed DeepFace, a nine-layer CNN for FR. The input faces are first aligned to have an upright position

using 3D modelling and piecewise affine transforms, before being fed to the network. This network achieves a recognition accuracy of 97.25% on the LFW dataset. Schroff *et al.* [8] report an accuracy of 99.63% on the LFW dataset using FaceNet, a CNN with 7.5 million parameters, trained using a novel *triplet* loss function. Other CNN architectures showing similar FR accuracy on the LFW dataset include the DeepID series by Sun *et al.* [9].

In this work, we analyse the vulnerability to PAs of three CNN-FR methods: the VGG-Face [7], LightCNN [15] and FaceNet [16]. The VGG-Face network has been included in our study because it is a very well known and widely referenced CNN for FR applications. The other two network models have been included because they are newer than the VGG-Face network, and have both shown an FR performance even better than that of the VGG-Face CNN. Another important reason why these three specific CNNs have been studied in this work is that the respective creators of these networks have made publicly available pre-trained models that can be directly used in our experiments. The following sections provide brief summaries of the selected CNNs, and describe how they have been used in our experiments.

3.2.1 VGG-Face CNN: The VGG-Face network model is made publicly available by the Visual Geometry Group (www.robots.ox.ac.uk/vgg/software/vgg_face) at Oxford University. Involving almost 135 million trainable parameters, this network has been shown to achieve an FR accuracy of 98.95% on the LFW unrestricted setting [10]. VGG-Face is a CNN consisting of 16 hidden layers (see Table 3 in [7]). The initial 13 hidden layers are convolution and pooling layers, and the last three layers are fully connected ('fc6', 'fc7' and 'fc8', following the nomenclature used by Parkhi *et al.* [7]). The input to this network is an appropriately cropped colour face image of pre-specified dimensions.

We use the representation produced by the 'fc7' layer of the VGG-Face CNN as a template for the input image. When enrolling a client, the template produced by the VGG-Face network for each enrolment-sample is recorded. For verification, the network is used to generate a template for the probe face image, which is then compared with the enrolled templates of the claimed identity using the Cosine-similarity measure given by (1) (where $\| \boldsymbol{A} \|_2$ represents the L_2-norm of vector $\boldsymbol{A}$)

$$\text{cosine\_similarity}(\boldsymbol{A}, \boldsymbol{B}) = \frac{\boldsymbol{A} \cdot \boldsymbol{B}}{\| \boldsymbol{A} \|_2 \cdot \| \boldsymbol{B} \|_2} \tag{1}$$

The score assigned to the probe is the average cosine similarity of the probe template to all the enrolment templates of the claimed identity. If the score is larger than a predetermined threshold, the probe is accepted as a match for the claimed identity.

3.2.2 LightCNN: Wu *et al.* [15] have proposed a new CNN, called LightCNN, for FR. Their goals in designing this network were to have significantly fewer trainable parameters compared to other state-of-the-art FR-CNNs, as well as to be able to handle noisy labels that are inevitable in datasets mined automatically from the web. Compared to the VGG-Face network, the number of parameters in the LightCNN model is smaller by a factor of 10 (~12 million parameters). This is achieved mainly through the use of a newly introduced max-feature-map (MFM) activation [15], which is a non-linear extension of the maxout activation operation. Although the MFM operator is more expensive to compute than the ReLU unit that it replaces, the large overall reduction of number of units per layer, made possible by the use of the new operator, still leads to smaller computation time for the forward pass of the LightCNN network (by a factor of 5, relative to the VGG-Face CNN [15]).

In our experiments, we use as templates the 256-D representation produced by the 'eltwise_fc1' layer of LightCNN. Note that this template is much smaller than the 4096-D vector produced by VGG-Face. Despite these relative efficiencies, the LightCNN network achieves an FR accuracy of 99.33% on the LFW dataset (in unrestricted setting) – outperforming the VGG-Face network by a small margin. As with the VGG-Face network, the cosine measure (1) is used to compare an input probe template to the relevant enrolment templates.

3.2.3 FaceNet CNN: Very recently, David Sandberg has made publicly available his implementation as well as trained models for a new FR-CNN named *FaceNet* [16]. This is the closest open-source implementation of the FaceNet CNN proposed by Schroff *et al.* [8], for which neither a pre-trained model nor the training-set is publicly available. Sandberg's FaceNet implements an Inception-ResNet V1 DNN architecture [27]. Two pre-trained FaceNet models have been published [16]. In our tests, we have used the 20170512-110547 model, trained on the MS-Celeb-1M dataset [28]. Using this model, FaceNet achieves an FR performance of 99.2% on the LFW dataset [16], which is comparable to the performance of LightCNN. Note that the 128-D representation produced by this network (at the 'embeddings:0' layer) is half the size of that produced by LightCNN. We use this representation to construct enrolment and probe templates, which are compared to each other using the Cosine measure (1).

3.3 GMM-based FR using inter-session variability modelling

ISV modelling is an extension of the GMMs-based method for face verification. We have used the ISV modelling approach proposed by Wallace *et al.* [3]. Among the FR methods included in this study, this is the only method that adopts a *parts-based* approach. The input normalised face image is first decomposed into a set of square sub-images of size (12×12), with an overlap of 11 pixels in each direction. Let N represents the total number of sub-images extracted from the input face image. For each sub image, a predetermined number, D, of low-frequency DCT (discrete cosine transform) coefficients is computed. Thus, a set of N D-dim arrays of DCT coefficients is extracted for each input face image. The DCT coefficients are normalised to zero-mean and unit standard deviation in each of the D dimensions. The resulting set of N normalised D-dim DCT arrays is used to represent the input face image.

To use a GMM for FR, first, a universal background model (UBM) is constructed from the set of training images (each represented by N D-dim DCT arrays). The UBM is a GMM, that is a weighted sum of K Gaussians, where each Gaussian is represented by a D-dim mean vector, and a $(D \times D)$ -dim covariance matrix. The parameters of the Gaussians components of the UBM, as well as their relative weights, are learned from the training data. The UBM describes the probability distribution of face sub-images in the D-dim DCT-coefficient space.

A *supervector*, $\boldsymbol{m}$, is then constructed by concatenating K the mean vectors of all the components of the GMM. To enrol a client i, a supervector $\boldsymbol{s}_i$ is derived as follows:

$$\boldsymbol{s}_i = \boldsymbol{m} + \boldsymbol{d}_i, \tag{2}$$

where $\boldsymbol{d}_i$ is an offset-vector specific to client i, computed using maximum a posteriori adaptation from the UBM [29].

The supervector $\boldsymbol{s}_i$ is assumed to be client specific. In other words, different enrolment images of the same client, i, should, in theory, produce very similar supervectors. For a given client, however, enrolment set typically contains images captured in different sessions, with varying pose, illumination and facial expressions. This within-class variability for client i is also reflected in $\boldsymbol{s}_i$, and can result in diminished recognition accuracy of the GMM-based FR method [3].

ISV-modelling has been developed to enhance the GMM-based approach by explicitly modelling and suppressing the within-class variability of each enrolled client. We assume that each enrolment image, collected in a separate session, results in a unique supervector. For enrolment image (i, j) (the jth enrolment image of client i), let

$$\boldsymbol{\mu}_{i,j} = \boldsymbol{m} + \boldsymbol{u}_{i,j} + \boldsymbol{d}_i, \tag{3}$$

where $\boldsymbol{\mu}_{i,j}$ is the supervector corresponding to the enrolment image (i, j), $\boldsymbol{u}_{i,j}$ is the offset induced by specific session conditions of image (i, j), and $\boldsymbol{d}_i$ is the client-dependent offset. (Note that the client-dependent offset $\boldsymbol{d}_i$ in (3) is free from session variability, unlike the $\boldsymbol{d}_i$ in (2).) The session-dependent offset, $\boldsymbol{u}_{i,j}$ can be expressed as

$$\boldsymbol{u}_{i,j} = \boldsymbol{U}\boldsymbol{x}_{i,j}, \tag{4}$$

where $\boldsymbol{U}$ is a low-dimensional matrix modelling the session variability, and $\boldsymbol{x}_{i,j} \sim \mathcal{N}(0, \mathcal{I})$ is a latent variable corresponding to the session variability. Similarly, the client-dependent offset, $\boldsymbol{d}_i$, can be represented as

$$\boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{z}_i, \tag{5}$$

where $\boldsymbol{D}$ is a diagonal matrix derived from the diagonal variances of the UBM, and $\boldsymbol{z}_i \sim \mathcal{N}(0, \mathcal{I})$ is a client-specific latent random variable. The subspace $\boldsymbol{U}$ is estimated using an expectation-maximisation algorithm. For details of the ISV technique for face verification, refer to the works of Wallace *et al.* [3] and Vogt *et al.* [30].

When enrolling a new client, i, using a set of enrolment images (indexed by j), the latent variables $\boldsymbol{x}_{i,j}$ and $\boldsymbol{z}_i$ are estimated from the enrolment images, and finally, the client-specific supervector, $\boldsymbol{c}_i$, is computed as

$$\boldsymbol{c}_i = \boldsymbol{m} + \boldsymbol{D}\boldsymbol{z}_i\,. \tag{6}$$

Thus, a single supervector, $\boldsymbol{c}_i$, is stored for each enrolled client.

Given a probe image, $\mathcal{P}$, claiming identity, i, the classification score for the probe is computed as the log-likelihood ratio $\mathrm{LLR}(\mathcal{P}, \boldsymbol{c}_i)$ as follows:

$$\mathrm{LLR}(\mathcal{P}, \boldsymbol{c}_i) = Q(\mathcal{P}|\boldsymbol{c}_i) - Q(\mathcal{P}|\mathrm{UBM}), \tag{7}$$

where $Q(\mathcal{P}|\boldsymbol{c}_i)$ gives the log of the likelihood of the probe $\mathcal{P}$ being generated by the model $\boldsymbol{c}_i$, and similarly, $Q(\mathcal{P}|\mathrm{UBM})$ gives the log of the likelihood of the probe $\mathcal{P}$ being generated by the UBM. In practice, we use the linear scoring method [31] to compute the score. This faster scoring method is derived as a first-order Taylor series approximation of the normal log-likelihood ratio. The probe is accepted if the resulting score is higher than a preset threshold.

3.4 ROC-SDK from Rank-One Computing

This FR product, from Rank-One Computing, has been included in our study as it is one of the best performing COTS FR products today [32]. In particular, Rank-One Computing have demonstrated an FR performance of 92% (at false reject rate = 1%) on the LFW dataset [33], and of 98% (at false reject rate = 1%) on the NIST Special Database 32: Multiple Encounter Dataset [34]. The SDK distributed by the company includes two FR methods – ROCFR (a pose-independent FR system), and ROCID (an FR system for images captured under controlled conditions). Here, we have tested the ROCFR method from the SDK (version 1.9). This method represents every face image with a 144 byte template. The SDK provides functions for populating a gallery with enrolment templates, and for comparing probe templates to the gallery of previously enrolled templates.

4 Experiment datasets and protocols

Each FR system is evaluated under two scenarios: *licit* and *attack*. In the licit scenario, where only bona fide samples are considered, the face-verification accuracy of the FR system is evaluated. Identities are enrolled in the system using a set of enrolment samples and the verification performance of the system is determined using the probe samples from various identities. In the attack scenario, we consider PA samples for the identities enrolled using the licit protocol, to evaluate the vulnerability of the system to PAs. That is, when evaluating the vulnerability of an FR system, each enrolled identity is only probed with PAs on that identity. The datasets and protocols used to evaluate the performance of each FR system are described in this section. We start by presenting the metrics used to quantify the performance of each FR system under the two scenarios.

4.1 Performance metrics

During the verification process, the user provides a biometric sample as probe, along with a claimed identity. Samples presented to a biometric verification system fall into three categories:

- *Genuine*: when the biometric sample and the claimed identity both belong to the user being authenticated.
- *Zero-effort impostor (ZEI)*: when the biometric sample belongs to the user, but the user claims the identity of a different enrolled client.
- *Impostor PA*: when the biometric sample matches the claimed identity, but neither correspond to the user attempting to be verified.

In standardised biometrics terminology [11], both genuine and ZEI presentations are considered bona fide presentations. The task of an FR system is to accept only genuine presentations, and to reject both types of impostor presentations.

The verification performance of an FR system is reported using the *false match rate* (FMR) and the *false non-match rate* (FNMR) [35]. The FMR denotes the proportion of ZEI that are wrongly accepted as genuine samples (Type-I error). The FNMR refers to the proportion of genuine samples which are wrongly rejected (Type-II error). The two metrics may be summarised in a single number as the half total error rate (HTER) as $\mathrm{HTER} = (\mathrm{FMR} + \mathrm{FNMR})/2$.

Vulnerability of a biometric system to PAs is reported as the impostor attack presentation match rate (IAPMR) [35], which denotes the proportion of PAs that are accepted by the FR system as genuine presentations.

4.2 Datasets

We have used four publicly available datasets, namely MOBIO [36], REPLAY-ATTACK [37], MSU-MFSD [38] and REPLAY-MOBILE [39].

The MOBIO dataset [36] was collected for bi-modal (voice and face) biometric verification experiments using mobile devices. Therefore, it contains only licit protocols. Biometric samples were collected in the year 2010, using Nokia N93i mobile phones, for 100 male subjects and 52 female subjects, spread over six cities (in five countries). The videos have been recorded at a resolution of (320×240) pixels. Experimental results reported in this paper have been produced using face images from only the male subjects.

The REPLAY-ATTACK [37] face PA dataset, published in 2012, consists of 1300 videos (each, at least 9 s. long) from 50 subjects. The videos have been recorded by using a 13′′ Macbook at 320×240 pixel resolution, under two different lighting conditions. Three kinds of replay attacks are simulated in this dataset, namely printed photo attacks, still face-images displayed on an electronic device and replayed digital videos. Three PA instruments (PAIs) have been used to construct the attacks: photo-quality paper (for the printed photo attacks), iPhone 3GS and iPad (first generation). The two electronic devices have been used for both, photo and video attacks. Videos for each kind of attack have been captured under two conditions, one where the capturing device is hand-held, and the other where the capturing device rests on a stationary support.

The public version of the MSU-MFSD dataset [38] includes real-access and attack videos for 35 subjects. This dataset was published in 2015. Real-access videos (∼12 s. long) have been captured using two devices: a 13′′ MacBook Air (using its built-in camera), and a Google Nexus 5 (Android 4.4.2) phone. Videos captured using the laptop camera have a resolution of 640×480 pixels, and those captured using the Android camera have a

Table 1 Different combinations of PA in each dataset. PA I refers to printed photo attacks in the dataset, PA II refers to low-quality video replay attacks, and PA III corresponds to high-quality video replay attacks. The REPLAY-MOBILE dataset contains only two PA types: PA I and PA III. Low-quality video replay attacks have not been included in this dataset

Dataset		PA I	PA II	PA III
REPLAY-ATTACK	capture device	12.1 mega-pixel Canon PowerShot SX150 IS camera	3.1 mega-pixel iPhone 3GS camera	3.1 mega-pixel iPhone 3GS camera
	PAI	Triumph-Adler DCC 2520 colour laser printer	iPhone 3GS (320 × 480 pixels) (165 ppi pixel density)	iPad (768 × 1024 pixels) (132 ppi pixel density)
	capture conditions	two lighting conditions ('normal' (bright) and 'adverse' (in the shade, but not dark)) hand-held or fixed		
	biometric sensor	Macbook (320 × 240)		
MSU-MFSD [31]	capture device	Canon 550D camera	8 mega-pixel iPhone 5S camera	Canon 550D camera
	PAI	HP Color Laserjet CP6015xh printer	iPhone 5S (640 × 1136 pixels) (326 ppi pixel density)	iPad Air (1536 × 2048 pixels) (264 ppi pixel density)
	capture conditions	Indoors, with artificial lighting		
	biometric sensor	Macbook Air (640 × 480), Nexus 5 (720 × 480)		
REPLAY-MOBILE [32]	capture device	18 mega-pixel Nikon Coolpix P520		LG-G4 (720 × 1280)
	PAI	Konica Minolta ineo + 224e colour laser printer		Philips 227ELH matte monitor (1920 × 1080)
	capture conditions	six lighting conditions (including artificial lights and natural light) Hand-held or fixed		
	biometric sensor	iPad Mini 2 (720 × 1280), LG-G4 (720 × 1280)		

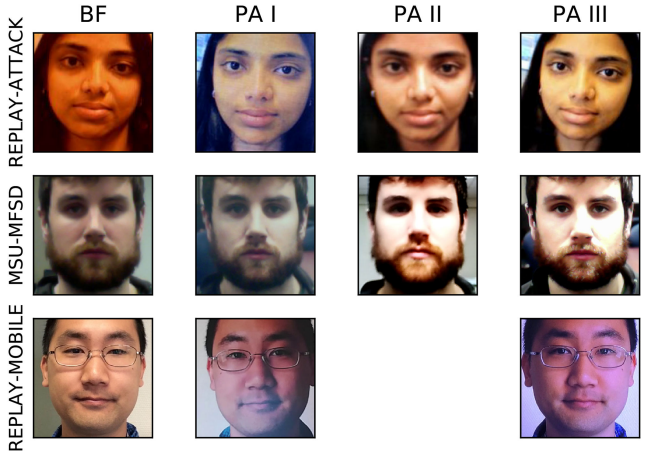

Fig. 1 *Examples of attacks represented in the various PA datasets. The first column shows a bona fide example, and the remaining columns show different types of PAs present in the dataset for the same identity. The various PAs are described in Table 1*

resolution of 720 × 480 pixels. The dataset also includes PA videos representing printed photo attacks, mobile video replay attacks where video captured on an iPhone 5S is played back on an iPhone 5S, and high-definition (HD) video replays (captured on a Canon 550D SLR, and played back on an iPad Air).

Published in 2016, The REPLAY-MOBILE dataset [39] is the newest of the three PA datasets used here. This dataset contains short (∼10 s. long) HD (720 × 1280) resolution videos corresponding to 40 identities, recorded using two mobile devices: an iPad Mini 2 tablet and an LG-G4 smartphone. The videos have been collected under six different lighting conditions, involving artificial as well as natural illumination. PAs represented in this database have been constructed using two PAIs: matte-paper for print attacks and matte screen monitor for digital-replay attacks. For each PAI, two kinds of attacks have been recorded: one where the user holds the recording device in hand, and the second where the recording device is stably supported on a stand. Thus, four kinds of attacks are represented in the database.

The various PAs represented in the three PA datasets are summarised in Table 1.

In this table, PA I refers to printed photo attacks, PA II refers to low-quality replay attacks, and PA III denotes high-quality replay attacks. Fig. 1 shows some example PAs from each PA dataset. The first column shows examples of bona fide presentations. The remaining three columns show examples of selected PAs represented in the corresponding dataset.

4.3 Protocols

As mentioned before, the MOBIO dataset comes with only a licit protocol, whereas the PA datasets include both licit and attack protocols. For our experiments with the MSU-MFSD, we have constructed new licit and attack protocols to analyse the vulnerability of the FR methods to PAs. The protocols in the various datasets are described in this section.

The set of clients in the licit protocol of each dataset is divided into three mutually exclusive *groups*: *training*, *development* and *evaluation*. The training group is reserved for producing the background models, when necessary. In our experiments, the UBM required for the ISV modelling method has been trained using the training group in the licit protocol of the MOBIO dataset. The training groups in the licit protocols of the PA datasets are ignored here. The development and evaluation groups, each consists of two *sets* of samples: an *enrolment* set and a *probe* set. All clients assigned to the group are represented in both sets. That is, for each client in the development group, a certain number of enrolment samples as well as probe samples are available (and likewise, for the evaluation group).

Attack protocols of the various PA datasets also consist of three non-overlapping groups: training, development and evaluation. The training group is intended to be used for training a PAD method. The development group can be used for tuning the parameters of the PAD method being tested, and the final performance metrics are reported for the evaluation group. In the vulnerability analysis experiments using a given PA dataset, attack videos in the evaluation group of the corresponding PAD protocol are used to test the vulnerability of the FR system in question.

The three groups in the male protocol of the MOBIO dataset are constructed as follows:

- *training*: 7881 sample images, from 37 subjects
- *development*: 2760 samples, from 24 subjects
- *evaluation*: 4370 samples, from 38 subjects.

As mentioned before, samples for the MOBIO dataset were collected at six different sites. Each group contains data exclusively from two sites.

The REPLAY-ATTACK dataset comes with a licit protocol consisting of 100 videos (one video per subject, per illumination

Table 2 Composition of the licit and attack protocols of the four datasets used in our experiments. All protocols consist of three mutually exclusive groups of clients: training, development and evaluation. For each dataset, the number of clients assigned to each group is listed here. The number in parentheses shows the total number of samples for the corresponding dataset and group

Dataset	Licit protocol			Attack protocol		
	Train	Development	Evaluation	Train	Development	Evaluation
MOBIO	37 (7881)	24 (2760)	38 (4370)			
REPLAY-ATTACK	15 (90)	15 (90)	20 (120)	15 (300)	15 (300)	20 (400)
MSU-MFSD	10 (200)	10 (200)	15 (300)	10 (600)	10 (600)	15 (900)
REPLAY-MOBILE	12 (1680)	16 (2240)	12 (1580)	12 (1920)	16 (2560)	12 (1920)

Table 3 Summary of the parameters of the various FR systems used in this study. All methods, apart from ROC-SDK, expect input images of fixed dimensions, where the image has been cropped to the face-region appropriately

FR method	Input	Features	Hyperparameters	Comparison metric
VGG-Face	224 × 224 colour image	output of 'fc7' layer length 4096	none (pre-trained CNN)	cosine distance measure
LightCNN	128 × 128 grey image	output of 'eltwise_fc1' layer length 256	none (pre-trained CNN)	cosine distance measure
FaceNet model: 20170512-110547	160 × 160 colour image	output of 'embeddings:0' layer length 128	none (pre-trained CNN)	cosine distance measure
ISV	80 × 64 grey image	session-variability-compensated supervector, length ∼45K	#GMMs: 512; Dim. of session-variability subspace, U: 160	log likelihood ratio
ROC-SDK (v1.9)	arbitrary size	undisclosed, proprietary size ∼140 bytes	using FRONTAL, FR, PARTIAL, and ROLL filters	undisclosed, proprietary

condition), and several PAD protocols. The licit protocol is partitioned thus:

- training: two videos each, from 15 subjects
- development: two videos each, from 15 subjects
- evaluation: two videos each, from 20 subjects.

The dataset also provides several PAD protocols [37]. In our experiments, we have considered the *grandtest* protocol, which includes all PAs. This protocol consists of 200 bona fide presentations, and 1000 PAs (for a total of 1200 videos).

The public version of MSU-MFSD [38] offers 70 bona fide presentation videos and 280 attack videos. As mentioned before, no licit protocol has been published for this dataset. For our experiments, therefore, we have devised a new licit protocol using some frames from the bona fide presentation videos. For each subject, the dataset contains two bona fide presentation videos, one captured using a Macbook Air (labelled 'laptop') and the other using a smart-phone (labelled 'android'). To construct the licit protocol for this dataset, we have used 10 frames from each bona fide video, sampled evenly, between the first and 180th frame. Frames from the 'laptop' videos form the enrolment set, and those from the 'android' videos form the probe set.

The licit protocol of the REPLAY-MOBILE dataset consists of 160 videos (four videos for each of 40 subjects). They are grouped as follows: 48 videos (12 subjects) for training, 64 videos (16 subjects) for development, and the remaining 48 videos for the test group. The PAD protocol of the dataset consists of 1030 videos (390 bona fide presentations and 640 PAs of different kinds).

For ease of reference, the number of clients and samples used in the licit and attack protocols of the various datasets is summarised in Table 2.

5 Experiments

The procedure used here to evaluate the vulnerability of an FR system is as follows. First the face-verification performance of the system is evaluated using only the MOBIO dataset (i.e. in the licit scenario). The hyper-parameters of the FR system are tuned to achieve optimal discrimination between genuine presentations and ZEI presentations. The vulnerability of the FR system to PAs is then evaluated, for the different PA datasets, using the same hyper-parameter settings. In this section, we first describe the methodology adopted for the experiments. Then, we present the results of FR and vulnerability analysis for the five selected FR systems.

5.1 Methodology

The FR and vulnerability-analysis experiments have been performed using the Python based Bob (www.idiap.ch/software/bob/) signal-processing and machine-learning toolkit [40, 41]. Besides a wide selection of machine-learning and signal processing tools relevant for biometics experiments, the Bob toolkit also includes interfaces for accessing the various datasets and associated protocols. All the FR systems studied here have been tested within the same software framework.

The inputs for the LightCNN, ISV modelling and ROC-SDK FR methods are grey-scale images, and the input to the VGG-Face and FaceNet FR methods is a colour (RGB) image. The ROC-SDK takes the entire image (one frame of video) as input. All other FR methods expect the input image to be cropped appropriately, so that it contains only the face region, with as little of the background as possible (see Section 3.1). The specific input image dimensions required by the various FR systems are listed in Table 3. The table also summarises the hyperparameters and the scoring method used in each FR system. For the ISV modelling method, we have used the hyperparameter values recommended by Günther *et al.* [42].

The ROC-SDK as well as the three CNN-FR methods (VGG-Face, LightCNN and FaceNet) come pretrained, and can be used out of the box. The UBM and the session-variability matrix ($\boldsymbol{U}$ in (4)) for the ISV modelling method are trained using training set in the licit male-protocol of the MOBIO dataset.

For experiments with each dataset, the hyperparameters of each FR method are tuned using data in the development group of the licit protocol (see Section 4). The score threshold may be chosen arbitrarily, but usually some heuristics are applied when choosing the threshold. Two common approaches for selecting the score threshold are:

- Select the score threshold, $\mathcal{T}_{\text{EER}}$, corresponding to the equal error rate (EER). That is, $\mathcal{T}_{\text{EER}}$ is the threshold for which FMR $\simeq$ FNMR.
- Select the score threshold, $\mathcal{T}_{\theta}$, that corresponds to a specific FMR $= \theta$ (e.g. $\mathcal{T}_{0.1}$ denotes the score threshold leading to an FMR of 0.1% over the development group).

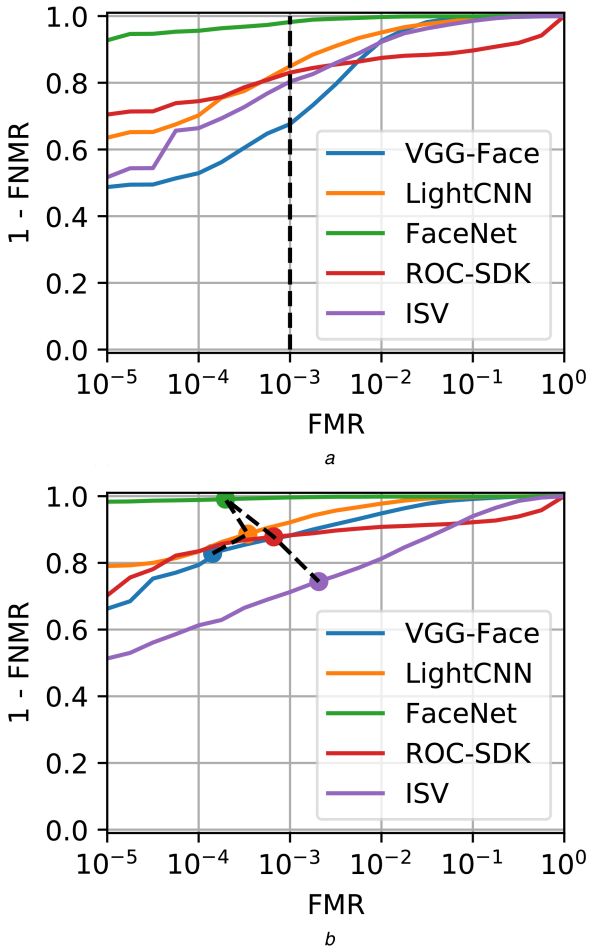

Fig. 2 *Performances of FR systems in licit scenario of the MOBIO dataset* ***(a)*** *ROC curves for the development group of the licit protocol of MOBIO. The vertical-dashed line marks the point where FMR is 0.1% for every method (corresponding to the score threshold $\mathcal{T}_{0.1}$),* ***(b)*** *ROC curves for the evaluation group of the licit protocol of MOBIO. The coloured circular markers (connected by a dashed line) indicate the FMR and FNMR of each FR method, for the $\mathcal{T}_{0.1}$ score threshold*

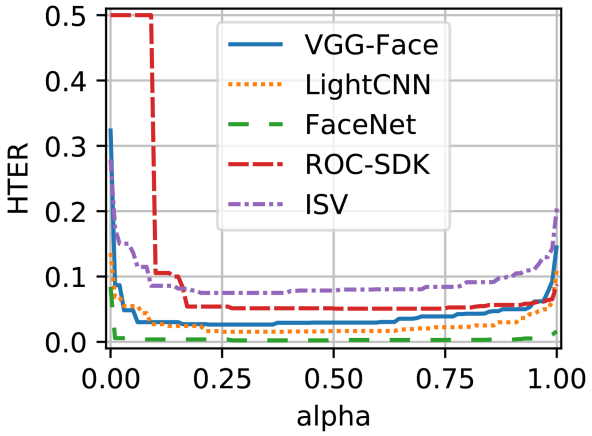

Fig. 3 *Comparison of FR systems in licit scenario. The plot shows EPC curves for the five FR systems. The licit protocol of the MOBIO dataset was used to generate these curves. This plot confirms the conclusion drawn from Fig. 2, that the FaceNet CNN significantly outperforms the other studied FR methods*

For a given score threshold obtained on the development group, the FMR, FNMR and HTER can be reported on the evaluation group.

5.2 Face-verification performance

First, we establish a baseline face-verification performance for each FR system, based on the licit protocol of the MOBIO dataset. The face-verification results of the various FR systems are presented in two ways: using receiver–operating characteristics (ROC) curves and using expected performance curves (EPC) [43].

The ROC curves shown in Fig. 2 illustrate the face-verification performances of the various FR systems. They have been evaluated using the licit male protocol of the MOBIO dataset. Figs. 2*a* and *b* show the ROC plots for the development group and evaluation group, respectively. A score threshold, $\mathcal{T}_{0.1}$, is selected so as to achieve an FMR of 0.1% over the development group. This is indicated in Fig. 2*a* by a dashed black vertical line. In Fig. 2*b*, the circular markers in the various colours (connected by a dashed trace line) indicate the FMR of the corresponding FR methods for the score threshold $\mathcal{T}_{0.1}$. To compare the performances of the five FR methods on the evaluation group at score threshold $\mathcal{T}_{0.1}$, we should consider the distances of these five points from the top left corner of the plot (where (FMR, FNMR) = (0, 0), representing the ideal outcome). This way of comparing different FR systems follows the method used by NIST [32]. It is clear from the ROC plots that the FaceNet CNN significantly outperforms the other FR methods in this study. Indeed, it is noteworthy that the verification accuracy of the FaceNet CNN remains extremely high even for very small values of FMR.

EPC [43] provide an unbiased comparison of FR systems. Fig. 3 shows the EPC plots of the five FR systems. These curves have been computed using the licit protocol of the MOBIO dataset, and show how the performance of each FR method evolves as a function of the trade-off between FMR and FNMR. This trade-off is controlled by the parameter $0 \leq \alpha \leq 1$ as follows:

$$\text{WER} = \alpha \times (\text{FMR}) + (1 - \alpha) \times (\text{FNMR}), \tag{8}$$

where weighted error rate (WER) is the weighted combination of FMR and FNMR. As always, the WER is computed from the classification results of the development group. Any particular value of WER corresponds to specific values of α, FMR and FNMR. Therefore, to achieve a specific WER, for a given value of α, the score threshold should be chosen appropriately. Specifically, when constructing EPC, we consider a discrete set of values for α. For each value of α, the score threshold that minimises the WER for the development group is chosen. The selected score threshold is then used to compute the HTER over the evaluation group, shown in the EPC plots (Fig. 3). Thus, the α values (on the horizontal axis) and the HTER values of the evaluation group (on the vertical axis of the plot) are related via the corresponding score threshold computed from the development group. Fig. 3 confirms that the three CNN-based FR methods perform better than the other two methods (ISV modelling and ROC-SDK). In particular, the figure also shows that the FaceNet CNN promises significantly better performance than all the other FR methods, including the popular VGG-Face network.

5.3 Discussion of face verification results

Comparing the two plots in Fig. 2, we note that (for the MOBIO male dataset) the performance of the VGG-Face CNN is not consistent over the two groups. At the operating point corresponding to score threshold $\mathcal{T}_{0.1}$, on the development group, the ROC-SDK and the ISV modelling method, both outperform the VGG-Face CNN. For the evaluation group, however, the VGG-Face network performs much better. Fig. 2*b* shows that the FMR for the VGG-Face CNN is at least one order of magnitude smaller than the FMR for both the ROC-SDK and the ISV modelling method. By contrast, the LightCNN and FaceNet networks consistently perform better than the other FR methods on both groups.

The variable α in (8) is a cost variable that can be adjusted depending the use case. If minimising FMR (Type I error) is critical, then high values of α are used to determine the score threshold from the development group. In the less likely scenario

where minimising FNMR (Type II error) is of utmost importance, very low values of α should be used. When it is important to optimise both types of errors, α is set to 0.5, which corresponds to determining the score threshold $\mathcal{T}_{\text{EER}}$.

The EPC plots also indicate that the FaceNet CNN shows near perfect FR performance over almost the entire range of α values.

Note the initial horizontal segment of the EPC curve for ROC-SDK, indicating a very high HTER for values of $\alpha \leq 0.1$. This results from the peculiarity of the score-distribution produced by the ROC-SDK method. To explain this behaviour, let us look at the score distributions as shown in Fig. 4. The score-distributions produced by the ROC-SDK for the two classes (genuine presentations in green and ZEI-presentations in blue) are shown in Fig. 4*a*. The score threshold for optimal separation between the two classes increases (along the horizontal axis) with α. The red line shows this evolution of the score threshold. Note that the histograms shown in the plot represent score distributions of the evaluation group, whereas the red line shows the evolution of the score threshold as computed over the development group. Fig. 4*b* shows the score distributions and score-threshold evolution produced by the VGG-Face CNN.

We can see in Fig. 4*a* that for some genuine presentations the ROC-SDK produces very low scores, which are similar to scores of ZEI presentations. Therefore, for low values of α, where the goal is to minimise FNMR even at the cost of high FMR, the selected score threshold is so low that almost every presentation is accepted. This leads to the near-50% HTER. We do not see this behaviour in the EPC curves of the other FR methods because these methods do not produce abnormally low scores for genuine presentations. Fig. 4*b* shows that the two score distributions produced by the VGG-Face network do not have significant overlap, and therefore, even for low values of α, the selected score threshold is not very low (relative to the score distribution of the ZEI class). Therefore, we see a gradual change at the beginning of the EPC curve for the VGG-Face method, and not a sudden drop as seen for the ROC-SDK method. Of course, this analysis applies only to the MOBIO male dataset. For other datasets, the score distributions produced by the ROC-SDK for the two classes need not have significant overlap.

5.4 Vulnerability analysis

The trained FR systems are next evaluated with respect to their vulnerability to PAs, using the PA datasets: REPLAY-ATTACK, MSU-MFSD and REPLAY-MOBILE, as well as a combined PA dataset, created by merging the three PA datasets. In the combined dataset, the licit protocol consists of the licit protocols of the individual PA datasets, and similarly, the attack protocol includes the attack protocols of the three PA datasets. In each protocol of the combined dataset, the development group is composed of the development groups of the three constituent PA datasets, and likewise for the evaluation group. First, the development group of the licit protocol in each PA dataset is used to determine a score threshold for classifying genuine and ZEI presentations. Presentations in the evaluation group of the licit protocol are then classified using this score threshold. These classification results are reported using the FMR and FNMR metrics. The presentations in the attack protocol of the dataset are also classified using the same score threshold, and the results are used to compute the IAPMR.

Vulnerability-analysis results for the five FR systems are summarised in Table 4. For each FR system, the table shows the results for the three PA datasets individually, as well as the combined PA dataset. The table shows both the face-verification accuracy and the vulnerability of each FR system. Values of the different metrics are reported for two score thresholds – $\mathcal{T}_{\text{EER}}$, and $\mathcal{T}_{0.1}$ – determined using the development group. The values of the performance metrics shown in the table have been computed over the evaluation groups of the various datasets. On the combined dataset, the LightCNN shows the same face-verification performance as the VGG-Face system. However, the general conclusion drawn from the HTER values in the table is that the FaceNet CNN achieves the best face-verification performance for all three PA datasets, as well as for the combined dataset.

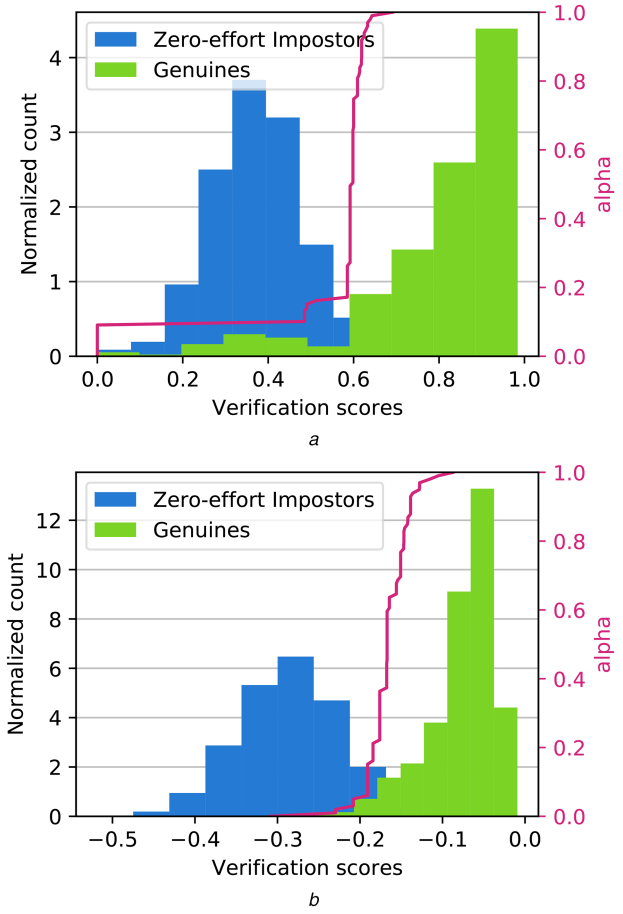

Fig. 4 *Score distributions of two FR systems in the licit scenario of MOBIO dataset. (a) Histograms of scores produced by the ROC-SDK method. (b) Score histograms for the VGG-Face network. In each plot, the green histogram represents the distribution of scores for genuine presentations, and the blue histogram shows the scores for the ZEI presentations. These histograms are based on scores computed for the evaluation group. In each plot, the red line shows the evolution (from left to right) of the score threshold computed from the development group, as a function of α (right vertical axis)*

(a) Score-distribution produced by the ROC-SDK, ***(b)*** Score-distribution produced the VGG-Face CNN

In Table 4, the IAPMR values for the all three CNN-FR systems are consistently >90% for all PA datasets (using both score thresholds). In fact, when we consider the lower-bounds of the respective confidence intervals, in a majority of experiments, the CNN-FR systems show vulnerabilities higher than 95%. These experiments clearly demonstrate that the CNN-based FR systems are highly vulnerable to PAs. Comparing the IAPMR values in the table, we note that both, the ISV modelling and the ROC-SDK methods, show levels of vulnerability significantly lower than those for the three CNN-based systems. For each dataset, the highest IAMPR values are shown in bold font in the table.

In the experiments where the vulnerability of an FR system is relatively low, the corresponding FNMR is also unacceptably high. For example, the ISV modelling method shows an IAPMR of 14.8% for the REPLAY-MOBILE dataset, at score threshold $\mathcal{T}_{0.1}$. The corresponding FNMR is 50.1%. A system with such a high FNMR would not be useful in practical applications.

This can also be observed from the score distributions of the FR systems. Fig. 5 shows the score distributions of the FR systems for the combined PA dataset. For every FR system, the score distribution of the PAs (grey histogram) significantly overlaps the genuine score distribution (green histogram). This indicates that none of the FR systems is able to adequately distinguish between genuine presentations and attack presentations. We note, however, that both, the overlap between the PA-score-histogram and the genuine-score-histogram as well as the separation between genuine

Table 4 FR accuracy and vulnerability of each FR system are shown for four datasets (the three PA datasets, as well as the combined dataset). The FR accuracies of five FR methods are reported in terms of FMR and FNMR. For convenience, the HTER is also reported for each experiment. The table also shows the vulnerability of each FR method to presentation attacks (IAPMR). The values in the table have been computed on the evaluation group, based on two different score thresholds: $\mathcal{T}_{\text{EER}}$ (corresponding to the EER on the development group) and $\mathcal{T}_{0.1}$ (leading to an FMR of 0.1% on the development group). 95% confidence intervals are shown for the IAPMR values in brackets. The highest IAPMR and the lowest HTER values for each dataset and score threshold are highlighted in bold

FR method	Dataset	Score threshold at EER ($\mathcal{T}_{\text{EER}}$)				Score threshold at FMR = 0.1% ($\mathcal{T}_{0.1}$)			
		FMR	FNMR	HTER	IAPMR	FMR	FNMR	HTER	IAPMR
VGG-Face	REPLAY-ATTACK	0.0	0.0	**0.0**	**98.2** [96.4, 99.3]	0.6	0.0	0.3	**99.8** [98.6, 100]
	MSU-MFSD	0.0	4.0	2.0	92.4 [90.5, 94.1]	0.0	1.3	0.7	**93.1** [91.3, 94.7]
	REPLAY-MOBILE	0.1	2.5	1.3	95.4 [94.3, 96.3]	0.0	4.4	2.2	90.7 [89.3, 92.0]
	combined	0.3	2.0	1.2	97.0 [96.4, 97.6]	0.0	3.8	1.9	92.7 [91.7, 93.5]
LightCNN	REPLAY-ATTACK	0.1	0.0	**0.0**	95.0 [92.4, 96.9]	0.1	0.0	**0.1**	98.0 [96.1, 99.1]
	MSU-MFSD	0.0	0.0	**0.0**	93.4 [91.6, 95.0]	0.4	0.0	0.2	**99.9** [99.4, 100]
	REPLAY-MOBILE	0.9	1.5	1.2	**99.9** [99.6, 100]	0.3	2.1	1.2	**99.7** [99.3, 99.9]
	combined	1.0	1.4	1.2	**99.8** [99.6, 99.9]	0.5	1.7	1.1	**99.6** [99.3, 99.8]
FaceNet	REPLAY-ATTACK	0.0	0.0	**0.0**	**99.5** [98.2, 99.9]	0.5	**0.0**	0.2	**99.5** [98.2, 99.9]
	MSU-MFSD	0.0	0.0	**0.0**	**99.0** [98.1, 99.5]	0.0	0.0	**0.0**	**100** [99.6, 100]
	REPLAY-MOBILE	0.0	0.5	**0.3**	**99.9** [99.6, 100]	0.0	0.5	**0.3**	**99.8** [99.5, 100]
	combined	0.1	0.5	**0.3**	**99.9** [99.7, 100]	0.0	0.5	**0.2**	**99.8** [99.6, 99.9]
ROC-SDK	REPLAY-ATTACK	2.4	1.2	1.8	89.5 [86.1, 92.3]	0.3	3.8	2.0	87.0 [83.3, 90.1]
	MSU-MFSD	14.5	10.7	12.6	83.3 [80.7, 85.7]	0.0	16.0	8.0	77.3 [74.5, 80.0]
	REPLAY-MOBILE	3.9	3.6	3.8	91.5 [90.1, 92.7]	0.0	5.2	2.6	87.5 [86.0, 89.0]
	combined	5.0	4.8	4.9	87.8 [86.6, 88.9]	0.1	6.7	3.4	84.5 [83.2, 85.7]
ISV	REPLAY-ATTACK	0.4	0.0	0.2	92.2 [89.2, 94.7]	0.1	1.2	0.7	82.0 [77.9, 85.6]
	MSU-MFSD	0.2	12.7	6.5	90.8 [88.7, 92.6]	0.9	6.7	3.8	95.0 [93.4, 96.3]
	REPLAY-MOBILE	10.2	5.5	7.9	92.0 [90.7, 93.2]	0.6	50.1	25.3	14.8 [13.3, 16.5]
	combined	9.9	4.7	7.3	94.4 [93.6, 95.2]	0.5	45.0	22.8	36.1 [34.4, 37.7]

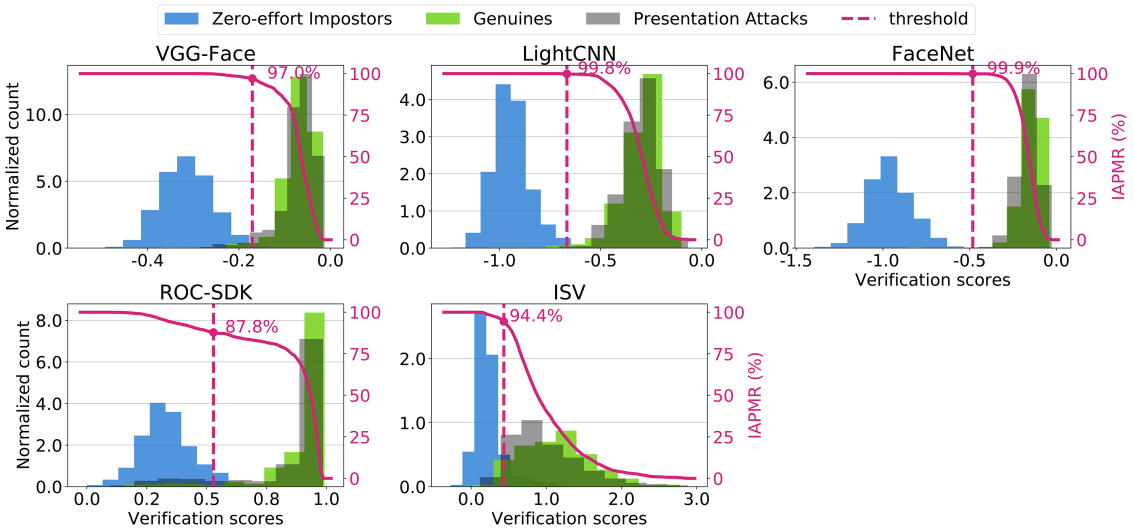

Fig. 5 *FR score distributions for all PA datasets combined. Each plot corresponds to one FR method (shown above the plot), and shows three histograms of scores – for genuine presentations (green), ZEI presentations (blue) and attack presentations (grey). The red vertical-dashed line marks the score threshold $\mathcal{T}_{EER}$, corresponding the EER of the development group (of the licit protocol), for the FR method in question. Samples with scores lower than $\mathcal{T}_{EER}$ are classified as ZEI presentations. In the ideal scenario the blue histogram would lie entirely to the left of this threshold, and the green histogram would lie entirely to its right. The solid red line shows the IAPMR for different thresholds. Given a specific score threshold, the IAPMR of resulting the FR system can be read-off at the point where the IAPMR curve (solid red curve) intersects the threshold (dashed vertical line)*

and ZEI score distributions are stronger for all the CNN-based FR systems, compared to the ISV modelling and ROC-SDK methods. The IAPMR values at $\mathcal{T}_{\text{EER}}$ shown in the plot (repeated in Table 4) are higher for the CNN-based methods than the other two FR methods.

The ROC-SDK method failed to generate templates for a certain number of input images, probably be due to some mechanism for rejecting low-quality input samples. This may explain the relatively lower vulnerability of this method as some PA (as well as genuine) samples may have been rejected based on their quality. We are not able to verify this hypothesis, as implementation details of ROC-SDK are not public knowledge.

Table 5 shows the IAPMR values for each PA type (see Table 1 for an explanation of the various PA types in each dataset). The values in the table show the average success rate for each type of PA, over the five FR methods. Clearly, all PA types are highly successful in spoofing all the five FR methods. The table indicates that the FaceNet CNN was successfully spoofed more often than the other FR methods.

Table 5 IAPMR (%) of each of the three PA types listed in Table 1

FR method	Dataset	PA I	PA II	PA III
VGG-Face	REPLAY-ATTACK	98.8	97.5	98.8
	MSU-MFSD	93.3	90.7	93.3
	REPLAY-MOBILE	91		99.7
LightCNN	REPLAY-ATTACK	96.2	95.0	94.4
	MSU-MFSD	93.3	93.3	93.7
	REPLAY-MOBILE	99.8	100	
FaceNet	REPLAY-ATTACK	100	98.8	100
	MSU-MFSD	100	97.0	100
	REPLAY-MOBILE	99.8		100
ROC-SDK	REPLAY-ATTACK	93.8	85	91.9
	MSU-MFSD	81.3	86.7	82.0
	REPLAY-MOBILE	89.6		93.4
ISV	REPLAY-ATTACK	95	91.2	91.9
	MSU-MFSD	95.7	78.3	98.3
	REPLAY-MOBILE	87.2		96.8

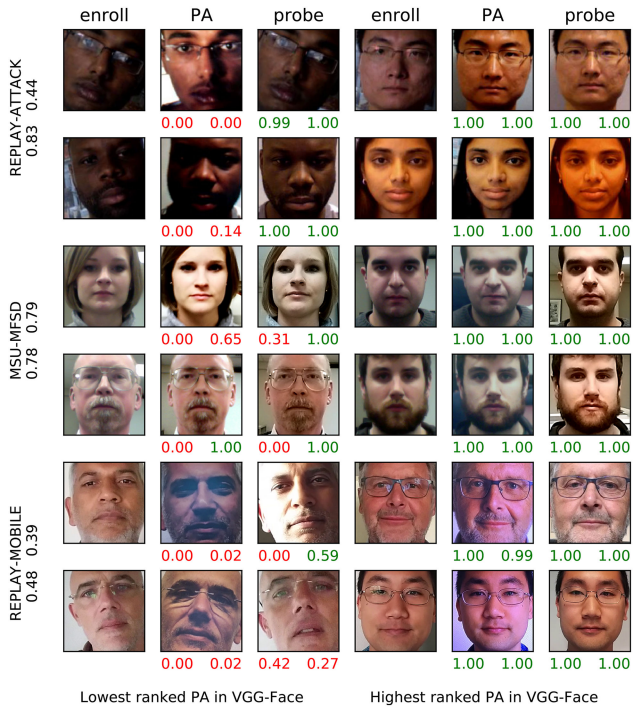

Fig. 6 *Examples of unsuccessful and most successful PAs in attacking the VGG-Face system. The three images on the left are associated with the unsuccessful PAs and the three images on the right are associated with the most successful PAs. Two sets of examples (two rows) is shown from each dataset. Three images are shown for each client: (left) an enrolment sample, (middle) example PA and (right) the genuine probe image that has the closest score to that of the corresponding PA sample. For each dataset, the EER probability thresholds are shown on the y-axis label for the VGG-Face and ISV systems, respectively. For each PA sample and probe sample, verification probabilities assigned by the VGG-Face system (left) and the ISV modelling system (right) are indicated below the sample. The verification probability is red if a sample is rejected using the EER probability threshold and green otherwise*

5.5 Discussion of vulnerability analysis results

In Table 5, we note that PA-III attacks in the REPLAY-MOBILE dataset are generally more successful in spoofing the FR systems than PA-III presentations in other datasets. (As described in Table 1, PA-III refers to high-quality video replay attacks.) This observation correlates with the fact that the PAI used (for video-replay attack presentations) in the REPLAY-MOBILE dataset is of higher quality than the PAIs used for video-replay attacks in the other two, older, datasets. That is, the digital screens used as PAIs for PA-III videos in the REPLAY-MOBILE dataset are more recent than those used for creating the older datasets.

Let us look at some specific cases of successful and unsuccessful PAs in detail. Fig. 6 shows the most successful and the least successful PAs in attacking the VGG-Face CNN. For each PA dataset, sample images are shown for four different clients:

- on the left, the two sets corresponding to the two clients with least successful attacks
- on the right, two sets of images corresponding to the two most successful PAs.

The labels on the vertical axis indicate the dataset from which the examples have been taken. For some images in the figure, the associated scores from two FR systems – the VGG-Face CNN and the ISV modelling method – are also shown. These scores have been calibrated to a standard scale using Platt scaling [44]. Therefore, the score values shown in Fig. 6 can be seen as face-verification probabilities.

For each dataset, the label on the vertical axis of Fig. 6 also shows the calibrated score thresholds (i.e. probability thresholds) for the two FR systems considered in this figure. The first number gives the EER probability threshold for the VGG-Face CNN, and the second number gives the EER probability threshold for the ISV modelling method. (The probability thresholds are computed over the development group.) For example, the probability threshold for the VGG-Face system, for the REPLAY-ATTACK dataset is 0.83, and the probability threshold for the ISV modelling method, for the same dataset, is 0.44.

The figure shows three images for each client – (from left to right) an enrolment sample, a PA sample and a genuine probe sample. For the PA and probe samples, the verification probabilities are shown in the figure, for the two FR systems (the left number is the probability assigned by the VGG-Face system for the image in question, and the number on the right is the probability estimated by the ISV modelling system). When the probability is equal to or higher than the corresponding probability threshold, the score is shown in green (i.e. the presentation was accepted as genuine). Otherwise, the score is shown in red (i.e. the sample was rejected by the corresponding FR system).

For all three datasets, the clients in the right half of the figure are those that received the highest scores from the VGG-Face system. For these cases, both presentations (PA and genuine probe) were also scored highly by the ISV modelling method. Visual inspection of the images indicates that for all these clients, the probe and PA samples are well-cropped images, with good frontal pose, captured under reasonably good illumination conditions.

In the left half of the figure, we show the two clients, per dataset, that were assigned the lowest score by the VGG-Face system. As explained earlier, the probability values in red indicate that the sample was rejected by the corresponding FR system. For example, the PA sample of the woman in the MSU-MFSD dataset was rejected by both FR systems. The genuine probe sample of the same client, however, was rejected by the VGG-Face system, but accepted by the ISV modelling system. This result is difficult to explain. The eyes in the three images for this client look well aligned, which indicates that the eye localisation was correct in all three images. In this case, illumination variations could be to blame for the low scores by the VGG-Face CNN. In some of the other cases, we note that either the enrolment image or the probe image is not correctly aligned to the image axes. Note that the images shown in this figure are normalised face images. Therefore, a misaligned normalised face indicates that the eye locations have been incorrectly estimated in the original image, during the normalisation process. This error may be due to insufficient illumination, shadows, or reflections from the eye wear.

6 Conclusions

For trustworthy face-biometrics systems, high face-verification accuracy and robustness to PA are equally important. In recent years, deep learning-based FR methods have captured the interest of biometrics researchers, because they have been shown to

outperform preceding methods by a wide margin. In this paper, we have empirically verified the hypothesis that the higher the face-verification accuracy of an FR system, the higher is its vulnerability to PA. This is the first study to empirically explore the vulnerability of CNN-based FR systems to a variety of PA. For comparison, we have also included two other FR methods not explicitly based on CNNs. Specifically, we have studied the robustness of five FR methods to PA, namely three CNN-based FR methods (VGG-Face, LightCNN and FaceNet), ISV modelling method, which relies on hand-crafted features, and the ROC-SDK, a commercial FR product. The three CNN-based FR methods as well as the ISV modelling method are not explicitly designed to detect PAs. (No claim can be made about the ROC-SDK as implementation details for this product are not publicly available.) Therefore, one would intuitively expect these methods to be vulnerable to PAs.

The face-verification performance and vulnerability to PA of the various FR systems are reported using recently standardised ISO metrics. Our experiments, conducted on three publicly available PA datasets, show that, while all the studied FR systems are highly vulnerable to PAs, the three CNN-based FR methods are all more vulnerable to PAs than the other two methods (based on the IAPMR values and confidence intervals shown in Table 4). Among the three CNN-based methods, the FaceNet-based FR method, which shows the best FR performance, is also most vulnerable to PA.

The experimental results presented in this paper are noteworthy for the following reasons:

- Face-verification performances of the various CNN-FR systems, previously evaluated only on the LFW dataset, are confirmed using a widely used face-verification dataset (MOBIO) as well as several recent publicly available PA datasets.
- Although the FR methods studied here have always been intuitively considered to be vulnerable to PAs, these are the first experiments to quantify this vulnerability empirically, based on large-scale study using several PA datasets. The experiments also demonstrate that raw FR performance alone does not make an FR method suitable in face-verification applications.
- Our experiments clearly demonstrate that FR methods with higher FR accuracy also show higher levels of vulnerability to PA.
- Since these experiments have involved publicly available PA datasets, the numerical results presented here can serve as baseline performance values for researchers in the field of face PAD.

These results make a clear case for incorporating explicit countermeasures to make FR systems significantly robust to PAs. In future work, we will study the impact of combining these FR methods with a PAD system.

The FR systems studied in this work have all been trained only for FR. The next logical step is to expressly take PAs into account when training the FR systems. Such experiments would have greatly enlarged the scope of the current work. In a subsequent work, we plan to show results of training a CNN for FR, when PAs are explicitly identified as a separate class.

The datasets used in our experiments are already publicly available. Python code for all the experiments and test protocols discussed in this paper is publicly available (https://gitlab.idiap.ch/bob/bob.paper.deepfacevuln2017).

7 Acknowledgments

We thank Mr. Tiago de Freitas Pereira for his assistance in interfacing the Bob machine-learning toolkit with the VGG-Face CNN. We also thank him and the anonymous reviewers for their comments, which have helped improve this paper. We gratefully acknowledge the support of the Norwegian SWAN project, the EU H2020 project TeSLA and the Swiss Center for Biometrics Research and Testing.

8 References

[1] Turk, M., Pentland, A.: 'Eigenfaces for recognition', *J. Cogn. Neurosci.*, 1991, **3**, (1), pp. 71–86

[2] Belhumeur, N., Hespanha, J.P., J., K.D.: 'Eigenfaces vs. Fisherfaces: recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, pp. 711–720

[3] Wallace, R., McLaren, M., McCool, C., *et al.*: 'Inter-session variability modelling and joint factor analysis for face authentication'. Int. Joint Conf. Biometrics (IJCB), 2011, pp. 1–8

[4] Prince, S.J., Elder, J.H.: 'Probabilistic linear discriminant analysis for inferences about identity'. Proc. IEEE Int. Conf. Computer Vision (ICCV), 2007, pp. 1–8

[5] El-Shafey, L., McCool, C., Wallace, R., *et al.*: 'A scalable formulation of probabilistic linear discriminant analysis: applied to face recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (7), pp. 1788–1794. Available at https://pypi.python.org/pypi/xbob.paper.tpami2013

[6] Taigman, Y., Yang, M., Ranzato, M., *et al.*: 'Deepface: closing the gap to human-level performance in face verification'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2014

[7] Parkhi, O.M., Vedaldi, A., Zisserman, A.: 'Deep face recognition'. British Machine Vision Conf., 2015, vol. **1**, pp. 41.1–41.12

[8] Schroff, F., Kalenichenko, D., Philbin, J.: 'FaceNet: a unified embedding for face recognition and clustering'. Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823. 00297

[9] Sun, Y., Liang, D., Wang, X., *et al.*: 'DeepID3: face recognition with very deep neural networks', arXiv preprint arXiv:150200873, 2015, **abs/1502.00873**, pp. 1–5, Available at https://arxiv.org/abs/1502.00873

[10] Huang, G.B., Ramesh, M., Berg, T., *et al.*: 'Labeled faces in the wild: A database for studying face recognition in unconstrained environments'. Technical Report 07-49, University of Massachusetts, Amherst, 2007

[11] '*ISO/IEC DIS 30107-1. information technology – biometric presentation attack detection – part 1: framework*' (International Organization for Standardization, Geneva, CH, 2016)

[12] Marcel, S., Nixon, M.S., Li, S.Z.: '*Handbook of biometric anti-spoofing*' (Springer, 2014)

[13] Karahan, Ş., Yildirim, M.K., Kirtaç, K., *et al.*: 'How image degradations affect deep CNN-based face recognition'. IEEE Int. Conf. Biometrics Special Interest Group (BIOSIG), 2016, pp. 313–320

[14] 'Trusted biometrics under spoofing attacks (TABULA RASA)'. Available at http://www.tabularasa-euproject.org/, accessed 12 September 2017

[15] Wu, X., He, R., Sun, Z., *et al.*: 'A light CNN for deep face representation with noisy labels', arXiv preprint arXiv:151102683, 2015, **abs/1511.02683**, pp. 1–13, Available at https://arxiv.org/abs/1511.02683

[16] Sandberg, D.: 'FaceNet: face recognition using tensorflow'. Available at https://github.com/davidsandberg/facenet, accessed 1 August 2017

[17] Duc, N.M., Minh, B.Q.: 'Your face is not your password face authentication bypassing Lenovo–As us–Toshiba'. Black Hat Briefings, 2009

[18] Kose, N., Dugelay, J.L.: 'On the vulnerability of face recognition systems to spoofing mask attacks'. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, 2013. Available at http://www.eurecom.fr/publication/3944

[19] Erdogmus, N., Dugelay, J.L.: 'On discriminative properties of TPS warping parameters for 3D face recognition'. Proc. Int. Conf. Informatics, Electronics and Vision (ICIEV), 2012

[20] Hadid, A.: 'Face biometrics under spoofing attacks: vulnerabilities, countermeasures, open issues, and research directions'. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2014, pp. 113–118

[21] Scherhag, U., Raghavendra, R., Raja, K.B., *et al.*: 'On the vulnerability of face recognition systems towards morphed face attacks'. Proc. 5th Int. Biometrics and Forensics Workshop (IWBF), 2017

[22] Amos, B., Ludwiczuk, B., Satyanarayanan, M.: 'OpenFace: A general-purpose face recognition library with mobile applications'. CMU-CS-16-118, CMU School of Computer Science, 2016

[23] Atanasoaei, C.: '*Multivariate boosting with look-up tables for face processing*' (EPFL, 2012)

[24] Uřičář, M., Franc, V., Hlaváč, V.: 'Detector of facial landmarks learned by the structured output SVM'. Proc. 7th Int. Conf. Computer Vision Theory and Applications (VISAPP), Portugal, 2012, vol. **1**, pp. 547–556

[25] LeCun, Y., Bottou, L., Bengio, Y., *et al.*: 'Gradient-based learning applied to document recognition', *Proc. IEEE*, 1998, **86**, (11), pp. 2278–2324

[26] Goodfellow, I., Bengio, Y., Courville, A.: '*Deep learning*' (MIT Press, 2016). Available at http://www.deeplearningbook.org

[27] Szegedy, C., Ioffe, S., Vanhoucke, V., *et al.*: 'Inception-v4, inception-ResNet and the impact of residual connections on learning', Proc. 31st AAAI Conf. On Artificial Intelligence (AAAI-17), San Fransisco, 2016, pp. 4278–4284

[28] Guo, Y., Zhang, L., Hu, Y., *et al.*: 'MS-Celeb-1M: A dataset and benchmark for large-scale face recognition', in Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.) '*Computer Vision-ECCV 2016*' Lecture Notes in Computer Science, vol **9907**, Springer, Cham, pp. 87–102

[29] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: 'Speaker verification using adapted Gaussian mixture models', *Digit. Signal Process.*, 2000, **10**, (1), pp. 19–41

[30] Vogt, R., Sridharan, S.: 'Explicit modelling of session variability for speaker verification', *Comput. Speech Lang.*, 2008, **22**, (1), pp. 17–38

[31] Glembek, O., Burget, L., Dehak, N., *et al.*: 'Comparison of scoring methods used in speaker recognition with joint factor analysis'. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2009, pp. 4057–4060

[32] Grother, P., Ngan, M., Hanaoka, K.: '*Face recognition vendor test (FRVT) part 1: verification*' (NIST, 2017). Available at https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing

[33] Huang, G.B., Ramesh, M., Berg, T., *et al.*: 'Labeled faces in the wild: a database for studying face recognition in unconstrained environments'. 07-49, University of Massachusetts, Amherst, 2007
[34] Watson, C.I.: 'Multiple encounter dataset I (MEDS-I)'. 7679, NIST, 2010
[35] Ramachandra, R., Busch, C.: 'Presentation attack detection methods for face recognition systems: a comprehensive survey', *ACM Comput. Surv.*, 2017, **50**, (1), Article 8; pp. 1–37
[36] McCool, C., Marcel, S., Hadid, A., *et al.*: 'Bi-modal person recognition on a mobile phone: using mobile phone data'. Int. Conf. Multimedia and Expo Workshops (ICMEW), 2012, pp. 635–640
[37] Chingovska, I., Anjos, A., Marcel, S.: 'On the effectiveness of local binary patterns in face anti-spoofing'. Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), 2012, pp. 1–7
[38] Wen, D., Han, H., Jain, A.K.: 'Face spoof detection with image distortion analysis', *IEEE Trans. Inf. Forensics Security*, 2015, **10**, (4), pp. 746–761
[39] Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., *et al.*: 'The REPLAY-MOBILE face presentation-attack database'. Int. Conf. Biometrics Special Interest Group (BIOSIG), 2016, pp. 209–216
[40] Anjos, A., El-Shafey, L., Wallace, R., *et al.*: 'Bob: a free signal processing and machine learning toolbox for researchers'. Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 1449–1452
[41] Anjos, A., Günther, M., de Freitas-Pereira, T., *et al.*: 'Continuously reproducing toolchains in pattern recognition and machine learning experiments'. Int. Conf. Machine Learning (ICML), Workshop on Reproducibility in Machine Learning Research, Sydney, Australia, 2017
[42] Günther, M., El-Shafey, L., Marcel, S.: 'Face recognition in challenging environments: an experimental and reproducible research survey', in Bourlai, T. (Ed.): '*Face recognition across the imaging spectrum*' (Springer, 2016, 1st edn.)
[43] Bengio, S., Mariéthoz, J., Keller, M.: 'The expected performance curve'. Int. Conf. Machine Learning, (ICML), Workshop on ROC Analysis in Machine Learning, 2005
[44] Platt, J.: 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods', *Adv. Large Margin Classif.*, 1999, **10**, (3), pp. 61–74