Spoofing Face Recognition With 3D Masks

Nesli Erdogmus, *Member, IEEE*, and Sébastien Marcel, *Member, IEEE*

***Abstract*— Spoofing is the act of masquerading as a valid user by falsifying data to gain an illegitimate access. Vulnerability of recognition systems to spoofing attacks (presentation attacks) is still an open security issue in biometrics domain and among all biometric traits, face is exposed to the most serious threat, since it is particularly easy to access and reproduce. In this paper, many different types of face spoofing attacks have been examined and various algorithms have been proposed to detect them. Mainly focusing on 2D attacks forged by displaying printed photos or replaying recorded videos on mobile devices, a significant portion of these studies ground their arguments on the flatness of the spoofing material in front of the sensor. However, with the advancements in 3D reconstruction and printing technologies, this assumption can no longer be maintained. In this paper, we aim to inspect the spoofing potential of subject-specific 3D facial masks for different recognition systems and address the detection problem of this more complex attack type. In order to assess the spoofing performance of 3D masks against 2D, 2.5D, and 3D face recognition and to analyze various texture-based countermeasures using both 2D and 2.5D data, a parallel study with comprehensive experiments is performed on two data sets: the Morpho database which is not publicly available and the newly distributed 3D mask attack database.**

***Index Terms*— Spoofing, presentation attack, face recognition, mask attack.**

I. Introduction

BEING the most commonly used biometric trait by humans, face recognition has become an active research topic for many decades now and it has found great application in consumer electronics and software. Face owes its reputation mainly to being easily and non-intrusively accessible compared to other biometric traits like finger print or iris. However, this advantage becomes a weakness in malicious circumstances, enabling attackers to create copies and spoof face recognition systems without any difficulties.

Spoofing attack is the act of outwitting a biometric system by presenting a fake evidence in order to gain authentication [1]. It is relatively simple to forge such an attack for facial recognition systems, due to the fact that the photographs or videos of a valid user can be easily captured from a distance or obtained via internet, e.g. through social networks. Valid users (simply users or clients) can be defined as the persons that are enrolled in a face recognition system. An attacker can attempt to gain access by simply showing their printed photos or replaying their recorded videos to the sensor.

This apparent vulnerability of face has evoked great interest in the biometric community and many papers have been published on countermeasure studies. Mainly as a result of their simplicity and low-cost, the previously mentioned photo print and video replay attacks [2] constitute the focus of research activities in this domain. Existing anti-spoofing approaches against these type of attacks can be roughly classified into three groups: texture analysis, motion analysis and liveness detection.

Assuming the presence of cues like printing artefacts [3] and/or blurring [4], many anti-spoofing techniques examine the texture of the captured face image. Similarly, in a recent study [5], micro-texture analysis using multi-scale local binary patterns is proposed. It can be argued that this type of approaches highly depends on the quality of the printed image or video display.

The second group of methods aims to detect spoofing attacks by analysing the motion in the scene based on the fact that planar objects like a sheet of paper or a mobile phone screen move in a significantly different way compared to real faces. For example, in [6], the trajectories of small regions of face images are examined to be classified as real or fake. In a similar manner, by computing geometric invariants of a set of automatically located facial points, Marsico *et al.* [7] exploit the same phenomenon.

Finally in the last group of methods, liveness of the face is determined based on live-face specific gestures such as eye blinking [8] or lip movements [9]. However, approaches of this kind are bound to fail in the case of video replay attacks or even more simply, with photographic masks which are actually high resolution facial prints worn on face after the eyes and mouth regions are cut out, as claimed in [10]. Similarly in [11], it is again shown that with eyes cut out from the photos, traditional visible liveness detection method still detects blinking, in other words, cannot distinguish a photo attack.

Recently, several studies have been published that present methodical and reproducible analyses of several of these and some other methods, with a shared purpose of providing comparable results on public databases [12]–[14].

Work on fraud detection capabilities for face is still limited and a substantial part of it is based on the flatness of the captured surface in front of the sensor during an attack. This is also true for approaches that examine the 3D nature of the face by employing additional devices, which is much more realistic now with the introduction of affordable consumer depth cameras like Kinect. For instance, in [15], 3D data acquired with a low-cost sensor is utilized to localize face

Manuscript received November 20, 2013; revised February 6, 2014 and March 28, 2014; accepted May 3, 2014. Date of publication May 6, 2014; date of current version May 22, 2014. This work was supported by the FP7 Programme through the European TABULA RASA Project and the partners of its consortium. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jaihie Kim.

The authors are with Idiap Research Institute, Martigny 1920, Switzerland (e-mail: nesli.erdogmus@idiap.ch; sebastien.marcel@idiap.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2014.2322255

and at the same time to test its authenticity to decrease their system's vulnerability to spoofing attacks.

Unfortunately, methods that depend on the assumption of a planar surface for a fake face are rendered futile in case of 3D facial mask attacks [16]. With the help of the advancements in 3D manufacturing technologies, easily attainable facial masks take the spoofing attacks one step further and introduce new challenges for counter measure studies. To the best of our knowledge, there have been very few studies published addressing this issue and they are detailed in the next section.

II. Related Work

The earliest studies in mask detection aim to distinguish between facial skin and mask materials by exploiting the difference in their reflectance characteristics. This idea can be traced 30 years back to [17], which claims that a face thermogram is not vulnerable to disguises and even plastic surgery can be detected, since it reduces the thermal signature of face. Later, stating that disguises can be detected even better in near-infrared, Pavlidis and Symosek propose to utilize the 1.3-1.7 μm sub-band of the upper band [18]. Simple thresholding is suggested for classification, without reporting any experimental results, but only illustrations.

Two more studies that follow the same way of thinking are published with systematic experiments and results [11], [19]. A multi-spectral analysis is proposed in both, claiming that fake, by its definition, is indistinguishable for human eyes and therefore, using only visual images is not sufficient to detect the attacks. On the other hand, they both handle the mask attack problem in an evasion/disguise scenario rather than spoofing since they don't examine masks that are replicas of valid users to be impersonated.

In [19], the authors conduct experiments on different mask materials such as silicon, latex or skin-jell to see how different they behave in reflectance when compared to facial skin that is sampled from the forehead region. For this purpose, the distribution of albedo values for illumination at various wavelengths are analysed and two best wavelengths, one from visual and one from near-infrared spectrum (685 and 850 nm) are selected. Finally, the resulting 2D vectors that consist of radiance measurements under these illuminations and strictly at a distance of 30cm from the sensor are classified as skin or non-skin via Fisher's linear discriminant. The method is reported to detect fake faces with 97.78% classification rate. However, the possibility of occlusion in the forehead region and the imposed range limitation restricts practical application. Additionally, in this study, masks don't even exist since the analyses are done directly on mask materials.

Similarly in [11], two discriminative wavelengths (850 and 1450 nm) are selected after examining the albedo curves of facial skin and mask materials with varying distances. An SVM classifier is trained to discriminate between genuine and fake attempts and tested on a database of 20 masks of different materials: 4 plastic, 6 silica gel, 4 paper pulp, 4 plaster and 2 sponge. The results show that the method can achieve a classification rate of 89.18%. This work improves the state of the art by eliminating the range limitation and experimenting on real facial masks. On the other hand, no analysis of how well the spoofing attacks perform could be presented, since although the masks are face-like, they do not mimic any real person.

Apart from lacking this analysis on spoofing performances of the masks, another limitation with these two studies [11], [19] is that they are not very convenient due to their special expensive hardware requirements, as stated by the same authors in [20].

Lately, a different line of research in spoofing with masks has been published by Kose *et al.* for which a non-public database composed of printed 3D masks of 16 users is utilized [21]–[24]. To construct the database (referred as Morpho database), the face models of clients are acquired by a 3D laser scanner and the masks are manufactured using a 3D printing service. In addition to texture images, 3D frontal face models are also made available in the Morpho database for both real and attack samples. More details on this database will be given in Section III-A. In their first two papers, the authors present a vulnerability analysis on 2D, 2.5D and 3D face recognition systems against 3D mask attacks and propose a micro-texture analysis based counter measure applied separately on color images and depth maps.

In [22], three baseline face recognition algorithms are implemented to observe the spoofing performances of the masks. In their experiments, a probe sample is compared to the enrolment (gallery) sample of the claimed ID and a binary decision is made based on a similarity metric. In their analysis, the authors do not designate an enrolment set, but instead they employ a method that is referred as *all vs all* and propose two scenarios. In the first scenario, the baseline performance is assessed by only using the real access samples (DB-r) in the database. Each DB-r sample is compared with all other DB-r samples. This results in two types of scores: *real genuine scores* if the compared samples belong to the same user and *real impostor scores*, otherwise. In the second scenario which is referred as the mode under spoofing attacks, mask attack samples (DB-m) are utilized as the probe set. Each sample in DB-m is compared with all DB-r samples, again resulting in two types of scores, that is *mask genuine* and *mask impostor scores*. The results are reported for both identification and verification settings as rank-1 close-set identification and Equal Error Rates (EER), respectively.

Although this analysis gives an idea about the spoofing potential of 3D facial masks, it suffers from two major problems. Firstly, one can strongly argue that spoofing is irrelevant in a close-set identification setting. This is because the probe will always be assigned to an identity in the gallery irrespective of the attack quality. Identity match can be achieved as long as the mask better resembles the target, compared to enrolment samples of other IDs. Secondly, in the verification setting considered for the second scenario, mask impostor scores are obtained by matching DB-m images to DB-r samples of IDs different than the one targeted by the attack. This is irrational since no attacker would produce an attack for a valid user and claim the identity of another. Additionally, the verification setting does not really evaluate the vulnerability of the recognition systems since apart from the algorithms used, their specifications, e.g. operating thresholds, are not determined and fixed using a development set. The correct approach would be to evaluate mask genuine scores against real genuine scores, which congregates the two scenarios in one score space and enables us to calculate false acceptance rates at the same operating point for both real and fake access scenarios.

In [21], a Local Binary Pattern (LBP) based counter measure to detect mask attacks is tested on two modes: color images and depth maps.[1] A depth map is also a grayscale image which contains information relating to the distance of the surfaces of 3D objects from a viewpoint. Multi-scale LBP features are extracted from both 2D and 2.5D images and a linear Support Vector Machine (SVM) classifier is trained to determine whether a feature belongs to a real or an attack sample. A training set is utilized which does not overlap with the testing partition. The results are presented separately for 2D and 2.5D modes as correct classification rates that are calculated at the thresholds giving best performances. Since a development partition does not exist, the thresholds are optimized on the test scores.

Later in [23], two fusion schemes, at feature and score levels, are proposed to combine previously proposed LBP histograms calculated from the 2D and 2.5D images. Classification results are given in the same previous manner; best performances obtained by tuning the decision threshold on the test set. While the 2D and 2.5D modes give 89.4% and 82.4% classification rates separately, the fusion of these two modes increases this rate to 93.5%. Additionally, a proper analysis is included in the paper on the impact of the mask attacks and the proposed counter measure method on two baseline (2D and 3D) face recognition algorithms. The detection error trade-off (DET) plots reveal that without any counter measures 3D masks can be highly detrimental to both 2D and 3D face recognition performances.

The authors explore another type of counter measure technique based on reflectance analysis in [24]. The proposed method utilize the variational retinex algorithm to decompose face texture images into reflectance and illumination components. Lastly, a linear SVM classifier is applied to classify reflectance images as masks or real faces. Again the performance evaluation is done without using a development set and results are reported as best possible classification rate.

With these last additions to the literature in the domain of face spoofing using 3D masks, the studies have definitely gained momentum. However, apart from the previously mentioned shortcomings, they unfortunately comprise a major obstacle to reproducible and comparable research: the utilized database is not available for public use.

Bearing these points in mind, in our paper we pursue three main purposes:

- Describing the first public spoofing database with facial masks, called 3D Mask Attack Database (3DMAD) in detail, along with complete protocols for experimentation.
- Presenting a very detailed baseline analysis on spoofing performances of each mask in this database against state-of-the-art 2D, 2.5D and 3D face recognition systems.
- Reporting comparative experimental results on two databases which will act as the missing link between the previous studies that have been done on Morpho database and the future studies that will be based on 3DMAD.

For reproducibility purposes, in addition to distributing the database, the source codes to generate the reported results are also made freely available to public use.

The rest of the paper is organized as follows: The two mask spoofing databases are described in detail in Section III. Three baseline face recognition algorithms and comparatively studied counter measure techniques are explained in Section IV and V, respectively. In Section VI, experimental results on the two database for both their spoofing capabilities and anti-spoofing performances of the implemented counter measure methods are provided. Finally, the paper is concluded with remarks on the future work in Section VII.

III. Mask Attack Databases

In this section, we will give detailed information about the two spoofing databases for which 3D masks are utilized to generate the attacks.

A. Morpho Database

Morpho database is a non-public database which was collected by MORPHO[2] in the context of the TABULA RASA project[3]. It consists of 207 real access and 199 mask attack samples, in both 2D and 3D (facial images and 3D models).

Subject-specific masks used for spoofing attempts are manufactured by a 3D printer using facial models of 16 different users, acquired with a 3D scanner. The texture of the masks is grayscale and their shapes are accurate replicas of the targeted clients.

For Morpho database, all recordings are done in a single session for both real accesses and mask attacks. A 3D scanner which uses a structured light technology is utilized to take different number of shots (varying between 9-15) for each user. 20 users are recorded in total of which only 16 had printed masks. For each shot, three manually annotated points (two for outer eye corners and one for the nose tip) are also included in the database.

As the acquisition process with the employed 3D scanner is really sensitive to movement, the cooperation of the clients are required. This requirement hinders the access to facial shape that is crucial to mask manufacture for the attackers.

B. 3D Mask Attack Database

The 3D Mask Attack Database (3DMAD) is a face spoofing database which currently contains 76500 frames of 17 different users, recorded using Microsoft Kinect sensor for both real-access and spoofing attacks using 3D facial masks. Each frame consists of:

[1] By color, it is referred to the texture images obtained by the 3D scanner which are actually in grayscale. In our paper, we will refer to grayscale texture images as 2D and to depth maps as 2.5D.

[2] http://www.morpho.com

[3] http://www.tabularasa-euproject.org/

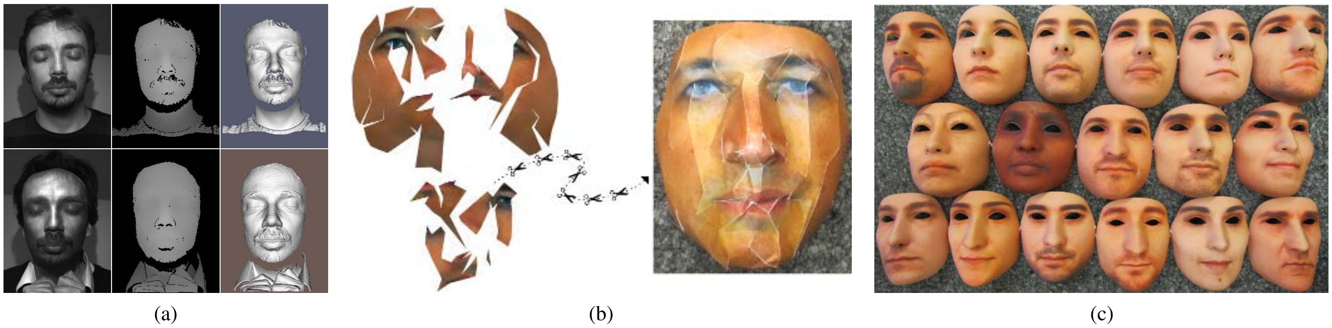


Fig. 1. (a) The top row shows a real access from a user in grayscale texture (2D), depth map (2.5D) and 3D model format while an attacker wearing the same user's mask is displayed in the bottom. (b-c) Facial masks obtained from ThatsMyFace.com. (a) Two example shots from Morpho. (b) Example paper-craft mask from 3DMAD. (c) 17 hard resin masks from 3DMAD.

- a depth image (640 × 480 pixels - 1 × 11 bits)
- the corresponding color image (640 × 480 pixels - 3 × 8 bits)
- manually annotated eye positions (with respect to the color image)

The production of the database can be divided into two stages: manufacturing the 3D masks and recording the videos that will be explained in detail in the following subsections.

1) Manufacturing the 3D Masks: In [25], it is stated that spoofing attacks using 3D facial masks cannot become a common practice in the literature, mainly because of the high cost of client-like masks. However, recently 3D printing services have sprung up and become a rapidly growing market, unfortunately, smoothing the way for different mask attack possibilities to face recognition systems.

The technique used to manufacture the masks in the Morpho database requires the 3D models of the valid users to be captured in order to be constructed using a regular 3D printer. Although 3D scanner technologies are advancing remarkably, range limitations and constraint for user cooperation still exist. For this reason, unconsciously taking hold of a proper 3D face model that belongs to a valid user, possibly from a distance is highly impractical. In order to produce the masks for our database, we used a particular service called ThatsMyFace.com which stands out with its specialization in facial reconstruction and in transforming 2D portraiture into 3D sculptures, among other options.

Using this company, it is possible for a 3D face model constructed using frontal and profile images of a person to be printed and delivered to your mailbox in several forms such as a head on an action figure or a wearable life-size mask in hard resin or a paper-cut file. Its website allows the customers to view and inspect the generated face model before ordering it. Obviously, the main advantage of this service over the others is the possibility of utilizing regular facial images to create a 3D model. Because unlike 3D scans, photographs of the users can be easily captured form a distance or found on the internet.

For 3DMAD, one frontal and two profile face images are taken from 17 different users and uploaded on ThatsMyFace.com. For each user, a **life-size wearable mask** and a **paper-craft mask** is ordered. The uploaded images which were used in the reconstruction of the 3D face models are also available in the database together with the paper-craft mask files (see Fig. 1(a)), for possible future use, making it possible for other researchers to create their own spoofs. However, they are not included in the scope of this study. In Fig. 1(b), the 17 wearable masks made out of a hard resin composite in full 24-bit color with holes at the eyes and the nostrils are displayed with a sample paper mask crafted only for illustration purposes.

Unfortunately, the size of the database could not be larger due to the high cost of the 3D facial masks. On the other hand, more samples can always be collected from the same masks under different conditions.

In conclusion, masks used for 3DMAD differs from the ones for Morpho database in several points. Firstly and most importantly, the 3D shapes of the masks in 3DMAD are not precisely correct as it is the case with Morpho database since the reconstructed models are only approximately computed from 2D images. The resulting mask quality depends on the input images as well as the performance of the reconstruction algorithm. Secondly, for 3DMAD masks, the eyes and the nostrils are cut out while the facial surface is complete in the Morpho database. Lastly, the mask textures are in 24-bit color for 3DMAD whereas they are in grayscale in the other.

2) Recording of the Database: Microsoft Kinect for Xbox 360 is utilized to record all samples in the database both for real accesses and mask attacks. Its sensor captures both color and depth data in the scene at 30 frames per second. The main reason behind the selection of this device over other conventional cameras is the additionally provided depth information, which makes it possible:

- to explore the attacks and devise countermeasures in 3D,
- to analyse the vulnerability of 3D face recognition systems to mask attacks

in addition to their 2D counterparts.

The data is collected in three different sessions for all users and in each session 5 videos of 10 seconds length are captured. The first two sessions are held two weeks apart in which real access samples are collected. Whereas in the third session, mask attacks that are performed by a single operator (attacker) are captured. As a result, 255 color and depth videos of 300 frames are recorded in total. Additionally, for each video,

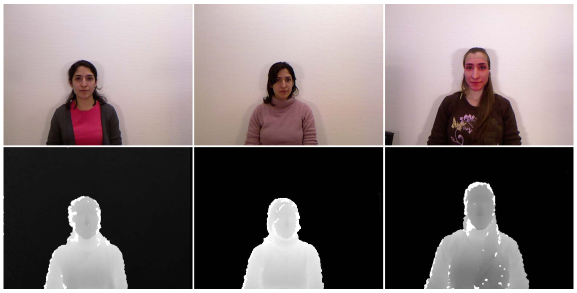


Fig. 2. Example color (top row) and depth (bottom row) images from three different sessions for a user in 3DMAD. The first two are real accesses while in the third, an attacker is wearing the user's mask.

the eye positions are manually labelled for every 60th frame and then, they are linearly interpolated for the rest in between.

For all three sessions, the recording conditions are very well-controlled. The users are recorded directly facing the sensor in front of a uniform background and under good illumination which is adjusted to minimize the shadows cast on their faces. In Fig. 2, three sample from three sessions for the same user is given as an example.

IV. Baseline Face Recognition Algorithms

Before moving on to develope counter measures against mask attacks, it is important to assert the threat they pose on the security of face recognition systems. In other words, it is required to evaluate the vulnerability of commonly employed face recognition algorithms to these type of spoofing attempts.

In our previous paper [26], we have examined the spoofing performances of a subset of masks (that belong to the testing set) in 3DMAD using a 2D face recognition algorithm that is based on Inter Session Variability (ISV) modelling method [27]. Due to the fact that, the dataset was divided into non-overlapping sets for training, development and testing, it was not possible to evaluate every mask.

In this study, we extend the previous analysis in three directions:

- The impact of masks are also evaluated for additional 2.5D and 3D systems.
- Experiments are done in leave-one-out manner so that all masks can be analysed.
- Morpho database is also included so that a connection is established from the current state of the art to possible future studies.

In [22] and [23], the authors also evaluate the spoofing success of the masks in the Morpho database on a 2D, a 2.5D and a 3D face recognition algorithm. For 2D and 2.5D, LBP histograms are extracted and compared using the χ^2 distance metric for both 2D and 2.5D images. For 3D face matching, Thin Plate Spline (TPS) warping parameters are obtained by aligning each face model with a generic one and comparison is done by computing cosine distances between corresponding feature vectors [28].

In a similar manner to these studies, ISV algorithm is selected to be applied on both grayscale texture images and depth maps for 2D and 2.5D face recognition, respectively. As for the 3D, since both databases do not include any facial expressions, Iterative Closest Point (ICP) method is selected to register surface pairs to each other and ICP error is simply taken as a measure of how well they match. Many early studies for 3D face recognition propose this convention [29]–[31]. These algorithms are selected simply to expand the number of different face recognition methods whose vulnerabilities are analysed against 3D mask attacks. Investigation for more recent or more advanced methods is not the point of this work.

A. ISV Method for 2D and 2.5D Face Recognition

Inter Session Variability (ISV) modelling, originally developed for speaker recognition, is applied for face recognition task [27]. Enrolled clients are described with Gaussian Mixture Models (GMM) that are built on set of blocks of pixels extracted from their images in the gallery. ISV aims to make these client models more reliable by eliminating within-client (inter-session) variation.

Initially, 12 × 12 blocks are exhaustively sampled from the facial image by moving the sampling window one pixel at a time. Next, mean and variance normalisation is applied and the first 45 2D DCT coefficients (lowest frequency) are extracted. Based on the distribution of these feature vectors, a GMM is estimated using background model (UBM) adaptation for each client. Finally, ISV modelling is applied to exclude the within-client variation, by assuming it is contained in a linear subspace of the GMM mean super-vector space and estimating subspaces via maximum likelihood and latent variables via maximum *a posteriori* adaptation. More details on the algorithm can be found in [27].

Like most of the existing 2D face recognition methods [32], ISV can also be adopted for facial shape information in 2.5D.

B. ICP Method for 3D Face Recognition

Iterative Closest Point (ICP) algorithm is a well-established technique used for rigid registration of 3D surfaces [33]. In order to minimize the distance between two cloud points (which is the sum of distances calculated for all points in one of the surfaces, finding the closest point on the other), ICP computes and revises the translation and rotation iteratively. This registration is used to establish point-to-point correspondences between two face models. Additionally, the final minimized distance, ICP error, can also be employed directly to compare them [29]–[31]. In fact, this is the chosen approach for our 3D face recognition baseline system.

Two main shortcomings of ICP are that it needs a good initialization for an accurate result and it cannot handle non-rigid transformation which is crucial in the presence of surface deformations, such as occlusions or facial expressions. But then, these issues are irrelevant in our case, since in both databases face samples are neutral and frontal.

As mentioned before, 3D face recognition in particular and face recognition in general is not the focus of this paper. We are aware that there exist many other more powerful methods but we are just interested in providing a baseline study on the vulnerabilities to 3D mask attacks that is open-source and available for the research community to reuse.

V. Anti-Spoofing Algorithms

As explained in the Introduction section, it is more difficult to detect 3D mask attacks with motion analysis and liveness detection methods. For this reason, texture analysis remains as a more reliable approach that can be adopted. Naturally, human skin is different from mask materials with its optical characteristics, such as reflectance or scattering. This fact facilitates utilization of texture properties to discriminate between real faces and masks.

Local Binary Patterns (LBP) is a simple and efficient texture operator which has become a popular approach in various computer vision applications [34]. As a matter of fact, LBP and its variations have been successfully applied in counter measures against 2D spoofing attacks [5], [35]. Moreover, Kose *et al.* evaluated the effectiveness of an LBP based micro texture-analysis technique in 3D face anti-spoofing using Morpho database [21], [23].

In [21], the multi-scale LBP based feature vector proposed in [5] against photo print attacks is utilized to detect 3D masks. For this purpose, it is applied separately on 2D and 2.5D images and classification rates are reported to be 89.4% and 82.4%, respectively. Later in [23], two modes are fused at feature and score levels to achieve higher success rates (93.0% and 93.5%).

In our work, we aim to study and compare discriminative properties of various types of LBP operators including the one proposed in [5] in real face / mask classification, using both 3DMAD and Morpho databases.

A. Extraction of LBP Based Features

The original LBP value for each pixel is calculated by comparing its adjacent pixels in 3 × 3 neighbourhood with the value of that pixel and forming a 8-bit binary number from the results ($\text{LBP}_{8,1}$) [34]. A common extension to the original operator is to eliminate patterns with more that two bitwise transition. This reduces the number of different labels and results in 59 *uniform patterns* ($\text{LBP}^{u2}_{8,1}$). The LBP operator can also be extended to use neighbourhoods of various size (different circle radius (R) and different number of sampling points (P) - $\text{LBP}_{P,R}$) or to change the encoding method.

The occurrences of the LBP labels in the whole image or in blocks are collected into histograms and then considered as feature vectors to be classified.

In [5], three types of LBP histograms are computed and concatenated: $\text{LBP}^{u2}_{8,1}$ from 3 × 3 overlapping blocks of face image (of size $59 \times 9 = 531$), $\text{LBP}^{u2}_{8,2}$ and $\text{LBP}^{u2}_{16,2}$ from the whole image (of size 59 and 234, respectively). It results in an enhanced feature histogram of length 833.

In this manuscript, three more extensions from [36] are included and analyzed: modified (mLBP), transitional (tLBP) and direction-coded (dLBP). Instead of the pixel value, the mLBP uses the average of the neighbouring pixels for comparison. In tLBP, two consecutive neighbour pixels are compared circularly in clock-wise direction and in dLBP, intensity variation is encoded for only four base directions into two bits, again resulting in 8-bit value (see Fig. 3).

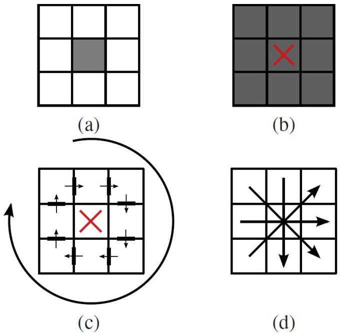


Fig. 3. Extended set of LBPs: (a) conventional LBP; (b) 8-bit coded modified LBP (mLBP); (c) transition coded LBP (tLBP); (d) direction coded LBP [36].

Additionally, the influence of dividing face images into blocks is assessed for each extended LBP type. The face image is broken into 3 × 3 non-overlapping blocks after the LBP values are computed. Histograms are calculated separately for each block and the final feature vector is formed by concatenation. In [14], block processing methodology is concluded to be ineffectual for detection of 2D spoofing attacks.

B. Feature Classification

Since the extracted LBP codes are collected into histograms, the classification can be simply performed by computing histogram similarities. So firstly, two reference histograms are calculated as the average of real access and mask attack samples in the training set and the features extracted from test samples are compared with these two using χ^2 metric, resulting in two distances: D_{real} and D_{mask}. The final score is computed as $D_{mask} - D_{real}$.

In [21] and [23], an SVM classifier is used with a linear kernel. SVM classification is also applied in our previous work [26], but the kernel type is chosen to be the radial basis function. In this study, a comparison between the two kernels is made with respect to their mask attack detection capabilities.

Additionally, Linear Discriminant Analysis (LDA) is tested in addition to χ^2 and SVM. Prior to LDA classification, Principal Component Analysis (PCA) is applied to reduce dimensionality while preserving 99% of the energy.

VI. Experiments and Results

Two types of experiments are conducted on both databases:

- Face verification experiments in which success rates of spoofing attacks with 3D masks are assessed using baseline face recognition algorithms.
- Anti-spoofing experiments in which mask attack/real face classification accuracies of aforementioned counter measure methods are measured.

The experimental setup is summarized in the diagram in Fig. 4, listing the algorithms used (both from previous studies and newly introduced) in face recognition and anti-spoofing modules. It also displays the result categories with respect to the probe image types and the decisions taken.

All of these algorithms are implemented using the free signal-processing and machine learning toolbox Bob.[4]

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

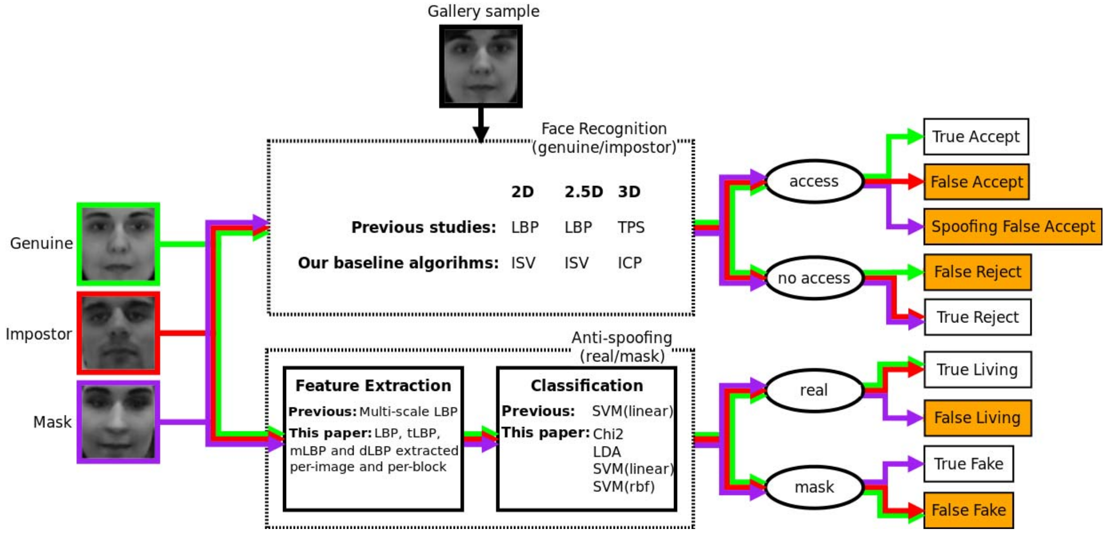


Fig. 4. A diagram showing face recognition and anti-spoofing modules with previously proposed algorithms for 3D mask attacks and the ones additionally studied in this paper. It also presents the result categories based on the probe image type and the decision taken by each module.

The source code for the experiments is available as one of its satellite packages.[5]

A. Vulnerability to Spoofing With 3D Masks

Before going into details of the experimental protocols and the obtained results for each database, the utilized performance measurements for these experiments will be briefly explained.

1) Performance Metrics: As shown in Fig. 4, a binary decision is made in a face verification system whether to accept or reject the input face. Hence, simply two types of errors exist: False rejection occurs when a genuine user is rejected while being compared to its own template and false acceptance occurs when an impostor is accepted despite claiming a false identity. With the introduction of the spoofing attacks, another error emerges referred as spoofing false acceptance which arises when an attack is successful and gains access. The rate of false rejection errors to the over all attempts from genuine users are measured as False Rejection Rate (FRR). Similarly, False Acceptance Rate (FAR) and Spoofing False Acceptance Rate (SFAR) are calculated as the percentage of false acceptance and spoofing false acceptance occurrences in all zero-effort (impostor) and mask attacks, respectively.

In our baseline verification experiments, the decision thresholds are selected as the operating points at which FRR is equal to FAR. The Equal Error Rates (EER) and SFAR values at this threshold are computed and reported.

2) Verification Experiments on the Morpho Database: Face verification experiments on the Morpho database are included in this study with the purpose of reproducing previously published results for comparison. Towards this end, LBP-2D, LBP-2.5D and TPS-3D algorithms are implemented and tested via the same *all vs all* protocol used in [22] and [23] in order to establish baseline systems' verification performance and vulnerability to 3D mask attacks. For real accesses, each sample is matched against all other samples in the real subset of the database and the generated scores are grouped as genuine if the samples belong to the same client or impostor, otherwise. Contrarily to what was proposed in [22], for mask attacks, each sample is only matched against real samples that belong to the target identity, resulting in a third group of scores, namely attack scores.

Due to the fact that the database is utilized as a whole and no divisions are proposed for training, development or testing in the previous analyses [22], the decision (EER) thresholds are determined a posteriori using all genuine and impostor scores obtained. Finally, at the same threshold, the rate of successful mask attacks (SFAR) are computed.

Two sets of experiments are conducted with the three algorithms implemented. Firstly, thanks to the collaboration of the authors of [22] who gave us access to their preprocessed data, the baseline face recognition algorithms are applied on their 2D and 2.5D images. It could not be done for 3D due to large data size. This practice enables us to compare different implementations of same algorithms, stripped from the effects of the pre-processing step.

In the second set of experiments, preprocessed images from [22] are replaced with images obtained by our preprocessing procedure. Similar to prior work, our preprocessing in 2D consists of cropping and geometrically normalizing face images according to eye positions. On the other hand, for 3D face models, only cropping is applied using a sphere centred at the nose tip and the hole filling and smoothing steps employed in [22] are omitted.

Setting the central axis of view of the camera in the direction of the camera's Z axis, the depth maps are obtained by measuring the distances of 3D facial points from the camera's XY plane. For this reason, the preprocessing applied on the 3D data also affects the resulting 2.5D depth maps. No additional preprocessing is employed.

3) Verification Experiments on 3DMAD Database: For 3DMAD database, ISV method on 2D and 2.5D images and

[5] Code available at: http://pypi.python.org/pypi/maskattack.study

TABLE I

Verification and Spoofing Rates: (1) Reported Results in [22] on Morpho; (2) Preprocessing From [22] and Our Implementation on Morpho; (3) Our Preprocessing and Implementation on Morpho; (4) Our Preprocessing and Implementation on 3DMAD

	LBP (2D)				LBP (2.5D)				TPS (3D)			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
EER	5.90%	6.16%	6.54%	4.92 %	7.27%	7.69%	17.63%	18.81%	3.85%	-	9.58%	13.28%
SFAR	-	51.27%	59.94%	28.04%	-	76.93%	41.98%	11.64%	-	-	54.09%	16.61%

ICP method on 3D models are also tested, in addition to the algorithms employed for Morpho database.

Contrary to 3D scanner outputs which are precise measurements of the facial shape, the 3D reconstruction of face models from 2D images is prone to error and noise. Due to this nature of the production process, the masks can have different levels of resemblance to the target persons. For this reason, in order to fully analyse the spoofing capabilities of the masks, a different protocol is adopted in which they are handled separately using leave-one-out cross-validation (LOOCV). For this purpose, the experiments are done in 17-folds, one for each mask. In each fold, the identity of the mask under analysis is taken out of the client set and the samples that belong to this user constitute the test set. From the remaining 16 clients, the first 8 is assigned to the training and the rest to the development set. In this way, training and development sets are kept maximally similar for all folds.

Among all algorithms, the only method that requires training is ISV. Hence, the real samples in the training set (from sessions 1 and 2) are used to train the UBM-GMM and to estimate the within-client variation of ISV.

The real access samples in the development set are utilized to obtain genuine and impostor score distributions, by using the session 1 samples for enrolment and session 2 samples for probing. With these scores, operating thresholds are determined at EER to assess baseline verification performances.

Lastly, attack scores are calculated for the mask samples in the test set, by again using the session 1 for enrolment, but this time session 3 for probing. All mask attack samples are matched with all enrolment samples since they all share a single identity.

Prior to verification tests, the data samples are preprocessed. Similar to preprocessing employed in 2D for Morpho database, 2D texture images in 3DMAD are also cropped, geometrically normalized using the eye positions supplied in the database and converted to grayscale.

For the 3D mode, a more complex preprocessing technique is utilized in order to obtain denser and more complete face models from the raw Kinect output. For this purpose, for each depth video in the database, every 30 frames are accumulated into one model, resulting in 10 face models per video. The process is illustrated in Fig. 5. Again for 2.5D, 3D models obtained after preprocessing are utilized to obtain the depth maps and no further action is required.

Before the accumulation of frames, firstly the raw depth data is converted to real world coordinates using the internal calibration parameters that are shipped with the Kinect device used for recording. Next, the obtained point clouds are cropped with a sphere of radius 8cm, centered at the nose tip.

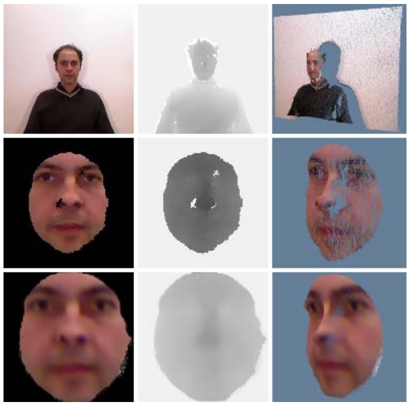


Fig. 5. Top: Raw color and depth images obtained with Kinect and the point cloud computed by mapping the depth data to the world coordinates. Middle: Texture map, depth map and point cloud from a single frame after cropping. Bottom: Texture map, depth map and point cloud after 30 frames are accumulated and post-processed. Texture maps are not used in the experiments, they are included here only for visualization purposes.

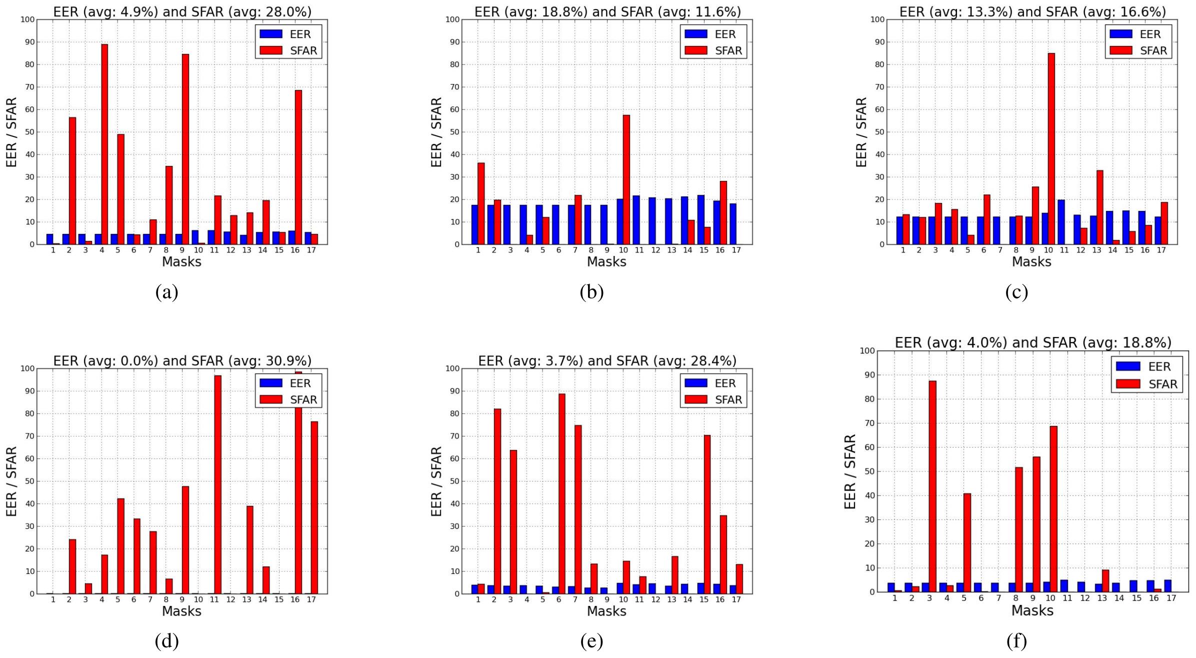
The nose tip is identified as the point with minimum depth value in the proximity of the eyes (using the eye positions included in the database). Then, each cropped face model is rigidly transformed and aligned with the first frame of that video using ICP method. Finally, all points in all 30 frames are collected into bins in a 300×300 grid of size $15cm \times 15cm$ which is centred at the nose tip. The accumulated model is obtained by taking the average of each bin. The grid also provides us the depth map, which is created by taking the z value at each bin (pixel) as intensity.

This method results in a denser and smoother facial shape due to accumulation and averaging, respectively. But, it can still be noisy and include holes. To overcome these issues, holes are filled via linear interpolation and then bilateral smoothing is applied.

In order to have equal number of texture and shape samples, every 30th frame is used for the experiments in 2D. For each fold of the LOOCV, 800 real access samples from two non-overlapping sets of 8 users are used for training and development, whereas for testing, 100 mask samples are matched against 100 enrolment samples. In total, 2550 samples are obtained and utilized for each mode (2D, 2.5D and 3D).

4) Results: The verification and spoofing results in terms of EER and SFAR for the Morpho and 3DMAD databases using LBP-2D, LBP-2.5D and TPS-3D algorithms are given in Table I. As explained before, SFAR values are obtained as the percentage of successful attacks that achieve higher similarity score than the EER threshold overall spoofing attempts.

EER rates reported in [22] are given in column (1) of Table I. SFAR values are not provided in [22] so they



(a) (b) (c) (d) (e) (f)

Fig. 6. EER and SFAR values obtained separately for each mask using the algorithms indicated in the sub-captions. (EER bars in (d) are too low to be visible.) (a) LBP method - 2D. (b) LBP method - 2.5D. (c) TPS method - 3D. (d) ISV method - 2D. (e) ISV method - 2.5D. (f) ICP method - 3D.

are omitted. Columns (2) and (3) give the results for our two sets of experiments on Morpho database: first using the preprocessed images from [22] and second using our preprocessing. Lastly, in column (4), EER and SFAR rates are given for 3DMAD for comparison.

Contrarily to Morpho database, the performance measurements on 3DMAD are done separately for each mask (as explained previously) and the average values of all folds are computed. More detailed results for each mask separately is reported in Fig. 6 (first row). Fig. 6 also provides the results of verification and spoofing experiments conducted on 3DMAD (second row) with new face recognition algorithms (ISV-2D, ISV-2.5D and ICP) as a comparison.

In Fig. 6, ISV is observed to have much lower verification errors in both 2D and 2.5D compared to its LBP counterpart. The average EER values for ISV are found to be 0.05% and 3.70% in 2D and 2.5D, respectively, whereas they are 4.92% and 18.81% for LBP. Similarly, the ICP method (EER of 3.99%) is found to perform better than TPS (EER of 13.28%) in 3D mode. It indicates that the proposed baseline algorithms are more advanced than the previous ones.

5) Discussion: The obtained EER values presented in column (2) of Table I show minimal differences compared to (1), indicating a successful reproduction of previously reported results.

Since the pre-processing for 2D images is very similar to the one employed in [22], EER values for 2D face recognition algorithms do not change significantly between columns (2) and (3) of Table I. On the other hand, an increase is observed in SFAR which implies sensitivity to cropping in mask attack detection. This phenomenon can be explained by the varying mask appearance according to whether the mask edges are visible or not in the facial image.

On the other hand, in 2.5D and 3D modes, the ommision of the hole filling and smoothing steps employed in [22] causes a deterioration in verification rates. Additionally, the comparison between SFAR values of (2) and (3) in the 2.5D experiments shows that spoofing performances are also adversely affected. Similar to real accesses, mask attacks also fail to be recognized by the system without appropriate preprocessing.

When the two databases are compared in columns (3) and (4) of Table I, both baseline verification and spoofing results for 3DMAD are found to be worse than Morpho when 2.5D or 3D data is used. This is expected partly because Kinect acquisition of depth is of much lesser quality compared to the laser scanner used for Morpho database. For SFAR values, another reason is that 3D shape accuracy of the masks in 3DMAD are not as high as the ones in the Morpho database, since they are approximately reconstructed from a couple of 2D images instead of being printed from real face models. This degradation in the mask quality is also reflected in spoofing performances in 2D.

If we look at SFAR values for each mask in 3DMAD (Fig. 6), three immediate conclusions can be drawn:

- The baseline algorithms proposed in this paper have much higher verification rates with respect to the ones used in previous studies.
- The spoofing performances differ greatly not only between masks but also between modes and algorithms.
- The vulnerability to mask attacks (i.e. average SFAR of all masks or number of masks that achieve SFAR of 20% or higher) is greater for more successful face verification algorithms, namely ISV-2D, ISV-2.5D and ICP.

For all methods tested, the baseline success rates for verification are observed not to fluctuate too much between folds.

In fact, the average of standard deviations for all 6 experiments is only 0.89%. This suggests that similar score distributions are achieved among folds, allowing more reliable comparisons for the masks.

The results show that spoofing capabilities vary substantially between masks, as well as modes and algorithms. The variation among different masks occurs because of the production quality and resemblance differences. For each method and mode, quite diverse SFAR values can be observed in Fig. 6, varying between 0% and 98.4%.

Additionally, for a fixed mask-algorithm pair the results also vary with respect to the utilized mode. For instance, let us look at the mask #11. With the ISV method in 2D, 96.68% of the spoofing attempts using this mask are successful, whereas in 2.5D, this ratio is only 7.52%. Likewise for the mask #10 with LBP method, SFARs in 2D and 2.5D are found to be 0.56% and 57.4%, respectively. These variations comes from the fact that 3D shape accuracies and texture qualities of the masks are independent from each other. For this reason, it is possible for a mask to present a high threat to a 2D face recognition system while failing in 2.5D or 3D.

Even more interestingly, the spoofing performance of a mask in one mode can also change with respect to the employed verification algorithm. Same kind of data from the same mask can be accepted using one algorithm and rejected in another. For example, in 3D mode if we compare the sFARs for the mask #5, two fifth of the attacks are successful with ICP (40.76%), but a very small portion is above the verification threshold with TPS (4.04%). A similar relationship exists also for the mask #17, but in the opposite direction. As another example, cropped 2D samples for the mask #17 achieve 76.32% SFAR with ISV and 4.44% with LBP. When we look deeper, we realize that this variation is partially related to the propensity of the users to contribute to FRR, which are classified as *goats* by Doddington *et al.* [37]. For the mask #5 given as an example above, the FRR for real access samples which share the same identity as the mask is much higher (32.52%) for TPS method, compared to ICP (8.84%). The reluctance to recognition for this user with TPS technique persists also for his mask. On the other hand, of course there are other variant factors between the algorithms that cannot be properly measured, such as the discriminative power or generalization ability of the extracted features in different specific cases.

Considering the verification rates vs spoofing performances, one would expect a direct proportion relation because a system which easily accepts the zero-effort (impostor) attacks should logically be even more susceptible to spoofing attacks. However, the experimental results show the opposite; that the systems with lower EER have more serious vulnerabilities and these two aspects are in fact inversely correlated. This phenomenon can be explained by the generalization ability of the advanced face recognition algorithms to a wide range of within-class variations. For instance, ISV suppresses these detrimental variations and hence it can generalize well to new unseen test samples, as can be seen from the excellent scores obtained. But this also results in an "ability" to recognize a mask, even if it is distorted and not exactly the same as the real person, making the ISV method the most vulnerable one among the others.

B. Anti-Spoofing for 3D Mask Attacks

In order to assess mask attack/real access classification rates of different LBP-based features extracted from 2D, 2.5D and 3D data and various classification methods, extensive tests are conducted.

1) Performance Metrics: As depicted in Fig. 4, being another binary (real/mask) classification problem, anti-spoofing also contains two types of errors: False Fake where the real accesses are classified as mask attacks (FFR) and False Living where the mask attacks are classified as real accesses (FLR). The performance measurements are done differently for Morpho and 3DMAD databases, due to their dissimilar experimental protocols.

In the prior studies, the Morpho database is divided into two non-overlapping partitions for training and testing. Since no development set is used, the accuracies are given as the best possible performance calculated on the test set and in terms of the area under the ROC curve (AUC) which is a plot of FLR vs 1-FFR (i.e. True Fake Rate). For the accuracy calculation, Equation 1 is utilized where τ is the decision threshold and N_r and N_m are number of real access and mask attack attempts.

$$\mathrm{Acc} = \max_{\tau} \left(1 - \frac{\mathrm{FFR}(\tau) \cdot N_m + \mathrm{FLR}(\tau) \cdot N_r}{N_m + N_r} \right) \quad (1)$$

On the other hand, the LOOCV protocol employed for 3DMAD allows us to determine the decision threshold on the development set. For this purpose, the operating point is obtained where FFR is equal to FLR, using the scores obtained from development samples. At the same threshold, FFR and FLR rates are also calculated for the test set. For both development and test sets, Half Total Error Rates (HTER) are computed according to Equation 2 as the final performance metric.

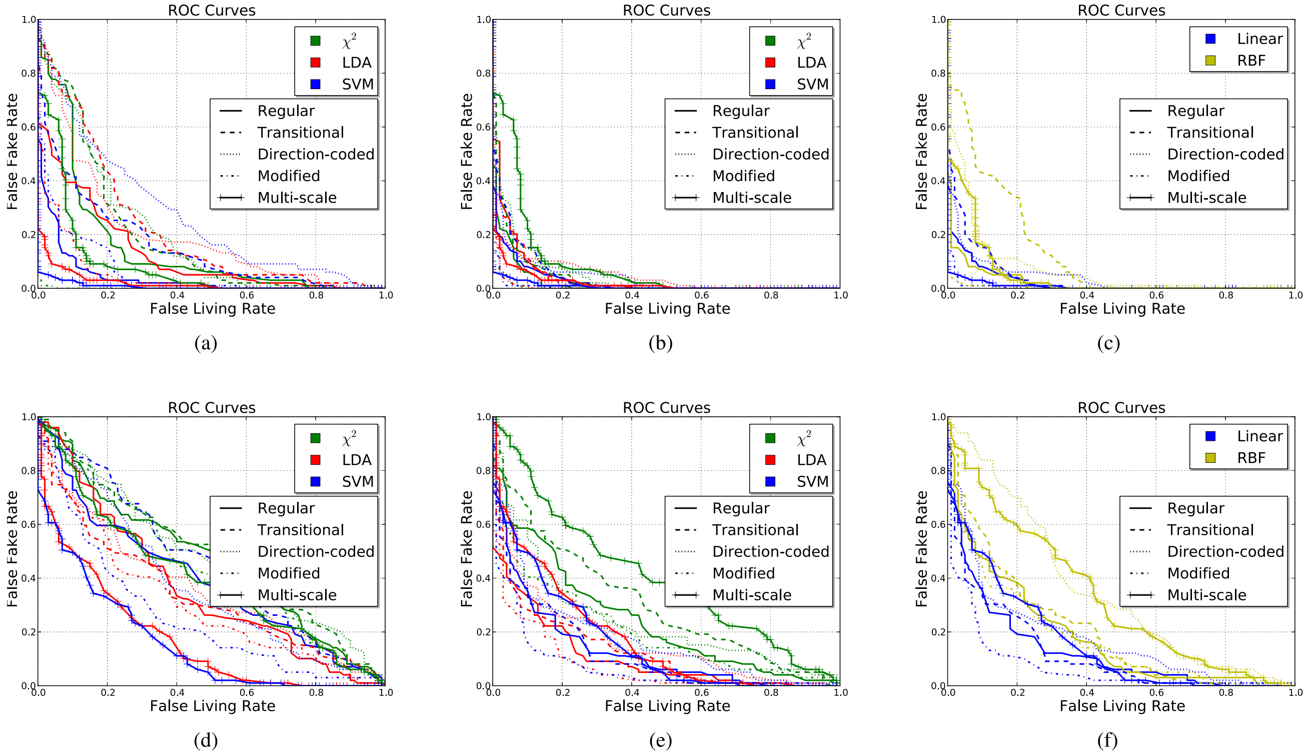
$$\mathrm{HTER}(\tau^*) = \frac{\mathrm{FFR}(\tau^*) + \mathrm{FLR}(\tau^*)}{2} \quad (2)$$

2) Anti-Spoofing Experiments on the Morpho Database: With the purpose of creating a connection to prior studies and reproducing the reported results, we implemented micro-texture analysis algorithm employed in both [21] and [23] and conduct classification experiments with the same training and test partitions used in these studies.

In [23], the authors also report the results for the fusion of 2D and the 2.5D information at feature and score levels. They conclude that taking the weighted sum of z-normalized scores performs better. Unfortunately, z-normalization parameters are computed and the weights are optimized using the test set. This assumes prior knowledge on the score distributions and mode reliabilities and hence it biases the results favourably.

In our experiments, we choose to skip normalization and we use same weights (0.6 for 2D and 0.4 for 2.5D) as proposed by the authors in [23].

For the sake of completeness, we additionally test other LBP types and classifiers mentioned previously in Section V on the Morpho database. This allows us to compare the previously employed algorithms on Morpho database with the new ones.



(a) (b) (c)

(d) (e) (f)

Fig. 7. ROC curves obtained using different LBP features and classifiers for Morpho database. (In (c) and (f) LBP features are extracted per-block.) (a) LBP features extracted per-image (2D). (b) LBP features extracted per-block (2D). (c) SVM results using linear and RBF kernels (2D). (d) LBP features extracted per-image (2.5D). (e) LBP features extracted per-block (2.5D). (f) SVM results using linear and RBF kernels (2.5D).

3) Anti-Spoofing Experiments on 3DMAD Database: In order to analyse the LBP-based anti-spoofing methods for 3DMAD, LOOCV method is adopted again. Mask attack detection performances are measured using each mask separately for 9 different sets of LBP features extracted from both 2D and 2.5D data.

Similar to face verification experiments, in each fold of LOOCV, the identity of the mask under analysis is assigned to the test set while the remaining 16 clients are divided in non-overlapping equal partitions for training and development.

For each one of the 17 folds, three different classifiers (χ^2, LDA, SVM with linear kernel) are trained using the training set and the classification scores are computed for both development and test sets. The decision threshold is determined as the operating point where FFR is equal to FLR in the development set and HTERs are calculated for both sets at this threshold.

Finally, for each LBP feature and for each classifier, the average values and the standard variations are computed for HTER rates of all folds.

4) Results: The obtained results for reproduction of prior work are given in Table II. Although the same preprocessed and cropped images are used in both experiments, the results show that our implementation fails to reproduce the exact or close results reported in [23]. We hypothesize that this disparity between the accuracies may be caused by the faulty LBP implementation[6] used by the authors. A similar occurrence was investigated in [14] and it was shown that the Matlab implementation of LBP responds unexpectedly in certain conditions, whereas they are handled correctly in the Bob toolbox that is used in our work. This may lead to completely different LBP codes in more than 5% of all computed codes.

TABLE II

REPORTED [23] AND OBTAINED RESULTS ON MORPHO DATABASE USING MICRO-TEXTURE ANALYSIS APPROACH IN [5] (BETTER PERFORMANCE IS IMPLIED BY HIGHER ACCURACY, HIGHER AUC AND LOWER EER.)

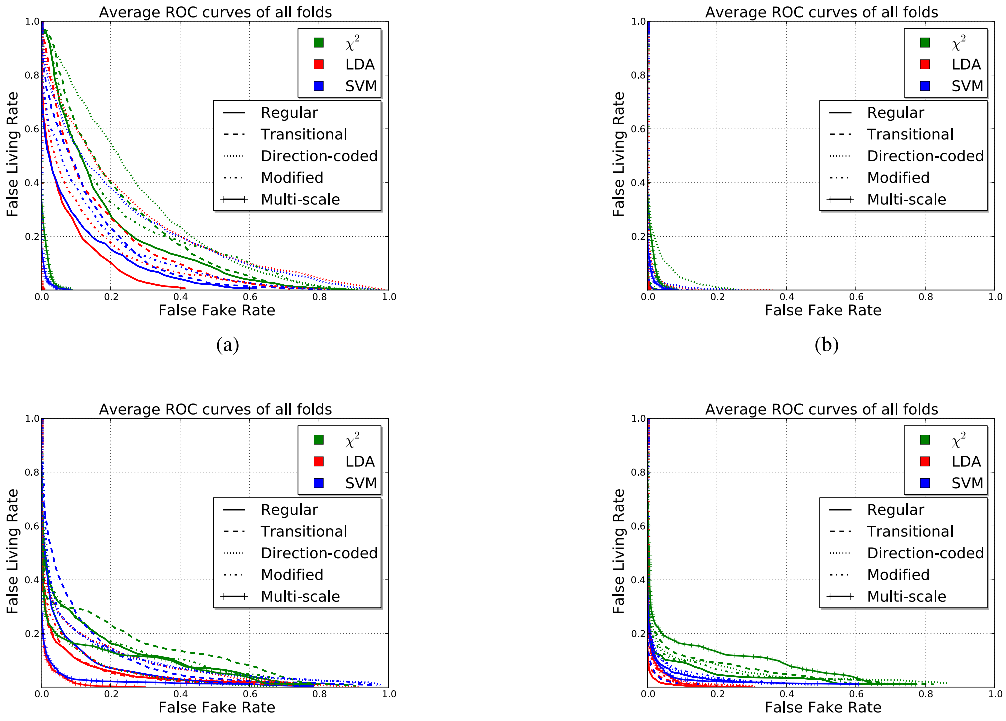
	Reported [23]			Obtained		
Data type	2D	2.5D	Fusion	2D	2.5D	Fusion
Accuracy	89.4%	82.4%	93.5%	97.0%	74.9%	94.5%
AUC	95.6%	91.5%	97.8%	99.4%	84.5%	98.9%
EER	-	-	-	5.0%	27.1%	7.0%

In our experiments, the anomaly affects the results adversely for mask attack detection in 2.5D. On the other hand, it produces better performance in 2D. Further analysis is needed to accurately identify the reasons behind this behaviour, however this investigation is not possible until the complete source code for [23] is available.

With our score-level fusion method, the EER is found to be 7.02% using the micro-texture analysis approach in [5]. The fusion in fact influenced the success rates negatively. With our implementation, 2D grayscale images outperform the 2.5D depth maps to the extent that simple fusion methods are inadequate.

The resulting ROC curves (FLR vs FFR) computed on the testing partition of Morpho database with all other LBP types and classifiers implemented are given in Fig. 7. Since the database is only divided into training and testing sets in the

[6] http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab



(a)

(b)

(c)

(d)

Fig. 8. ROC curves obtained using different LBP features and classifiers for 3DMAD database. The ROC curves are computed for the same threshold range for all masks and the average curves are calculated after merging development and test scores. For SVM, linear kernel is utilized. (a) LBP features extracted per-image (2D). (b) LBP features extracted per-block (2D). (c) LBP features extracted per-image (2.5D). (d) LBP features extracted per-block (2.5D).

previous studies, ROC curves are preferred to display the error rates at all possible operating points.

As for 3DMAD database, the results are given with two different plots. Firstly, in order to be able to compare the anti-spoofing performances with Morpho database, the ROC curves are presented in Fig. 8 for each LBP type and classifier. Since an operating point is not required to be calculated for ROC plots, development set is not need anymore and hence, these curves are generated by merging the development and test sets in each fold and finally taking their average. Secondly, using the development set to find the EER threshold, the average HTER rates for 2D and 2.5D images are computed and displayed in Fig. 9 as a bar graph where the variability of the HTER values are graphically represented by error bars.

5) Discussion: Several main points can be inferred from the results on the Morpho database as it can be seen in Fig. 7:

1) Mask detection performance in 2D images are much higher than the 2.5D images.
2) Block-based extraction of LBP features ameliorates the classification rates, especially for 2D images.
3) It is not possible to conspicuously point out one of the classifiers as the best among the three, but we can claim that SVM and LDA mostly draw ahead of χ^2.
4) Between extended LBP types, modified LBP delivers the best performance in all settings. Despite its shorter length (531), this descriptor is even better than the multi-scale LBP (833) [5], except when per-image feature extraction is utilized for 2.5D depth maps (Fig. 7-d).
5) For SVM classification, utilization of linear kernel yields to better results.

If we look at the findings from 3DMAD experiments given in Fig. 9 in parallel with Morpho database:

1) For block-based and multi-scale LBP approaches, performance of 2D images are again higher than the 2.5D images, on the other hand for LBP features extracted per-image, 2.5D images are observed to provide better results.
2) In accordance with previous observations, extraction of LBP features per-blocks is positively effectual for both 2D and 2.5D modes.
3) Again similarly to results with Morpho, SVM and LDA are better than χ^2, but this time LDA mostly yields to lower error rates than SVM.
4) For 2D images, all block-based and multi-scale LBP features yield to very good results except for dLBP. Particularly, regular block-based LBP shows an excellent performance with $0.12 \pm 0.47\%$ error, followed by modified block-based LBP ($0.50 \pm 0.95\%$) again using LDA. Similarly for 2.5D images, LDA gives lowest error rates for almost all LBP types, the best of them being again regular block-based ($3.91 \pm 6.04\%$) LBP, followed by multi-scale ($5.0 \pm 8.61\%$) LBP.

When the results are analysed closely, the deviations are observed to be larger for test sets, compared to development sets. This points to a serious generalization problem for most of the tested counter measures, especially when the 2.5D depth maps are utilized. For 2D images, regular

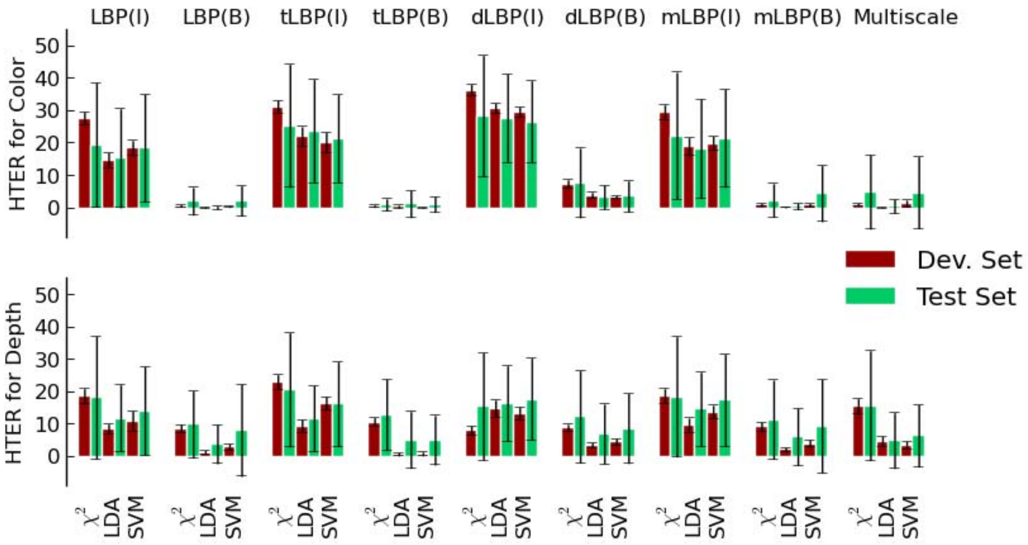


Fig. 9. The HTERs of different LBP types applied per-image (I) and per-block (B) for 2D and 2.5D images in 3DMAD are presented with error bars, where the uncertainties are indicated by standard deviations.

and modified block-based LBP with LDA behave as wanted, having almost no deviation, but unfortunately, this is not the case for any of the methods using 2.5D images. This also shows us that reporting best possible performance or AUC on the test set can be highly misleading. Generalization should be handled and analysed as seriously as differentiability.

VII. Conclusion

Spoofing attacks continue to be a security threat for biometric recognition systems and face is among the most vulnerable traits due to its high accessibility. Majority of previous studies in face spoofing focus on preventing 2D attacks performed by displaying printed photos or replaying recorded videos on mobile devices. However, utilization of 3D masks for face spoofing attacks has become easier and cheaper with the advancements in 3D reconstruction and printing technologies.

In our paper, we aim to contribute to the current state of the art in the research domain of 3D mask attacks. For this purpose, we extend our previous work [26] in three directions; firstly by assessing spoofing performances on 2.5D and 3D systems, secondly by analysing each mask separately with LOOCV and lastly experimenting on another 3D mask spoofing database which has been used in some of the previous studies but is not publicly available, in addition to 3DMAD. The parallel evaluations of LBP based anti-spoofing methods on these two databases allow us to associate previously published results on the Morpho database with our current work and with possible future studies on 3DMAD.

The face verification experiments conducted on 2D, 2.5D and 3D baseline systems reveal their vulnerability to spoofing attacks using facial masks. Additionally, they help us to see the different nature of the two types of masks used in two databases. The results show that the masks in 3DMAD for which the facial shapes are reconstructed from 2D images are not as capable as the ones in Morpho database for which he facial shapes are obtained via a 3D scanner.

Furthermore, the success rates of LBP based features in 3D mask attack detection are assessed via exhaustive tests using three different classifiers. The results for both 2D and 2.5D images indicate an advantage in the block-based approach. Among different LBP types tested, modified LBP is observed to deliver best results for Morpho database, despite its shorter length compared to multi-scale LBP which was proposed in previous publications. On the other hand in 3DMAD, regular block-based LBP shows the best performance for both 2D and 2.5D data. As for classification, LDA and SVM are found to be better than χ^2, while LDA is proved to be best in case of 3DMAD database.

A possible extension to this work is to search for more generalizable algorithms to detect the mask attacks, in order to avoid large variations in error rates. The obtained score distributions for 3DMAD are observed to vary between development and test sets, resulting in suboptimal decision thresholds and hence, increased error rates.

Another point that needs to be deliberated is the utilization of mask attack samples for training the anti-spoofing systems. Ideally, a countermeasure algorithm against spoofing should be able to decide whether the face image captured by the sensor belongs to a real face or not, regardless of the attack type. Because it is not realistic for a biometric system to employ a different anti-spoofing module for each attack type. In all of the previous works and in this study, the classifiers are trained using both real and attack samples.

Acknowledgment

The authors would like to thank the FP7 European TABULA RASA project for the financial support and to the partners of the consortium for their collaboration; particularly to Morpho for giving access to their database and to Eurecom for cooperating in reproduction of their results.

References

[1] K. Nixon, V. Aimale, and R. Rowe, "Spoof detection schemes," in *Handbook of Biometrics*, A. Jain, P. Flynn, and A. Ross, Eds. New York, NY, USA: Springer-Verlag, 2008, pp. 403–423.

[2] B. Biggio, Z. Akhtar, G. Fumera, G. L. Marcialis, and F. Roli, "Security evaluation of biometric authentication systems under real spoofing attacks," *IET Biometrics*, vol. 1, no. 1, pp. 11–24, Mar. 2012.

[3] J. Bai, T.-T. Ng, X. Gao, and Y.-Q. Shi, "Is physics-based liveness detection truly possible with a single image?" in *Proc. IEEE ISCAS*, May/Jun. 2010, pp. 3425–3428.

[4] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," *Proc. SPIE*, vol. 5404, pp. 296–303, Aug. 2004.

[5] J. Maatta, A. Hadid, and M. Pietikainen, "Face spoofing detection from single images using micro-texture analysis," in *Proc. IEEE IJCB*, Oct. 2011, pp. 1–7.

[6] K. Kollreider, H. Fronthaler, and J. Bigun, "Evaluating liveness by face images and the structure tensor," in *Proc. IEEE Workshop Autom. Identificat. Adv. Technol.*, Oct. 2005, pp. 75–80.

[7] M. De Marsico, M. Nappi, D. Riccio, and J. Dugelay, "Moving face spoofing detection via 3D projective invariants," in *Proc. 5th IAPR ICB*, Mar./Apr. 2012, pp. 73–78.

[8] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *Proc. IEEE ICCV*, Oct. 2007, pp. 1–8.

[9] G. Chetty and M. Wagner, "Multi-level liveness verification for face-voice biometric authentication," in *Proc. Biometrics Symp., Special Session Res. Biometric Consortium Conf.*, Sep./Aug. 2006, pp. 1–6.

[10] K. Kollreider, H. Fronthaler, and J. Bigun, "Verifying liveness by multiple experts in face biometrics," in *Proc. IEEE CVPRW*, Jun. 2008, pp. 1–6.

[11] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 436–441.

[12] M. M. Chakka *et al.*, "Competition on counter measures to 2-D facial spoofing attacks," in *Proc. IJCB*, Oct. 2011, pp. 1–6.

[13] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *Proc. IJCB*, Oct. 2011, pp. 1–7.

[14] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group*, Sep. 2012, pp. 1–7.

[15] F. Tsalakanidou, C. Dimitriadis, and S. Malassiotis, "A secure and privacy friendly 2D+3D face authentication system robust under pose and illumation variation," in *Proc. Int. WIAMIS*, Jun. 2007, p. 40.

[16] N. Erdogmus and S. Marcel, "Spoofing 2D face recognition systems with 3D masks," in *Proc. Int. Conf. Biometrics Special Interest Group*, 2013.

[17] F. J. Prokoski, "Disguise detection and identification using infrared imagery," *Proc. SPIE*, vol. 0339, pp. 27–31, Jun. 1983.

[18] I. Pavlidis and P. Symosek, "The imaging issue in an automatic face/disguise detection system," in *Proc. Workshop Comput. Vis. Beyond Vis. Spectr., Methods Appl.*, 2000, pp. 15–24.

[19] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *J. Opt. Soc. Amer. A*, vol. 26, no. 4, pp. 760–766, 2009.

[20] J. Yan, Z. Zhang, Z. Lei, D. Yi, and S. Z. Li, "Face liveness detection by exploring multiple scenic clues," in *Proc. 12th ICARCV*, Dec. 2012, pp. 188–193.

[21] N. Kose and J.-L. Dugelay, "Countermeasure for the protection of face recognition systems against mask attacks," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6.

[22] N. Kose and J.-L. Dugelay, "On the vulnerability of face recognition systems to spoofing mask attacks," in *Proc. IEEE ICASSP*, May 2013, pp. 2357–2361.

[23] N. Kose and J.-L. Dugelay, "Shape and texture based countermeasure to protect face recognition systems against mask attacks," in *Proc. IEEE Conf. CVPRW*, Jun. 2013, pp. 111–116.

[24] N. Kose and J.-L. Dugelay, "Reflectance analysis based countermeasure technique to detect face mask attacks," in *Proc. Int. Conf. DSP*, Jul. 2013, pp. 1–6.

[25] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. IAPR ICB*, Mar./Apr. 2012, pp. 26–31.

[26] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect," in *Proc. Biometrics, Theory, Appl. Syst.*, Sep./Oct. 2013, pp. 1–6.

[27] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Inter-session variability modelling and joint factor analysis for face authentication," in *Proc. IJCB*, Oct. 2011, pp. 1–8.

[28] N. Erdogmus and J.-L. Dugelay, "On discriminative properties of TPS warping parameters for 3D face recognition," in *Proc. ICIEV*, May 2012, pp. 225–230.

[29] B. B. Amor, M. Ardabilian, and L. Chen, "New experiments on ICP-based 3D face recognition and authentication," in *Proc. 18th Int. Conf. ICPR*, vol. 3. 2006, pp. 1195–1199.

[30] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition," *Comput. Vis. Image Understand.*, vol. 101, no. 1, pp. 1–15, 2006.

[31] G. Medioni and R. Waupotitsch, "Face modeling and recognition in 3-D," in *Proc. IEEE Int. Workshop AMFG*, Oct. 2003, pp. 232–233.

[32] C. McCool, J. Sanchez-Riera, and S. Marcel, "Feature distribution modelling techniques for 3D face verification," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1324–1330, 2010.

[33] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992.

[34] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, 1996.

[35] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Int. Conf. Biometrics Special Interest Group*, Sep. 2012, pp. 1–7.

[36] J. Trefnỳ and J. Matas, "Extended set of local binary patterns for rapid object detection," in *Proc. Comput. Vis. Winter Workshop*, 2010.

[37] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, p. 5.

Nesli Erdogmus received the B.S. and M.S. degrees from the Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey, in 2005 and 2008, respectively, and the Ph.D. degree from the Department of Multimedia Communications, EURECOM, Biot, France, in 2012, focusing on utilization of 3D data in face recognition. She is currently a Postdoctoral Researcher with the Biometric Person Recognition Research Team, Idiap Research Institute, Martigny, Switzerland, with a particular interest in spoofing attacks to face recognition systems.

Sébastien Marcel received the Ph.D. degree in signal processing from the Centre National d'Études des Télécommunications, Université de Rennes I, Rennes, France, in 2000, the research center of France Telecom (now Orange Labs). He is a Senior Research Scientist with the Idiap Research Institute, Martigny, Switzerland, where he leads the Biometrics Group and conducts research on multimodal biometrics, including face recognition, speaker recognition, vascular recognition, and spoofing and anti-spoofing. In 2010, he was a Visiting Professor with the University of Cagliari, Cagliari, Italy, where he taught a series of lectures in face recognition. In 2013, he was a Lecturer with Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, where he taught Fundamentals in Statistical Pattern Recognition. He was the main organizer of a number of special scientific events or competitive evaluations all involving biometrics, and serves as an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He is also the Principal Investigator of international research projects, including MOBIO, TABULA RASA, and BEAT. Finally, he leads the development of the Bob, the signal processing and machine learning toolbox.