A Competition on Generalized Software-based Face Presentation Attack Detection in Mobile Scenarios

Z. Boulkenafet[1], J. Komulainen[1], Z. Akhtar[2], A. Benlamoudi[3], D. Samai[3], SE. Bekhouche[4], A. Ouafi[4], F. Dornaika[5], A. Taleb-Ahmed[6], L. Qin[7], F. Peng[7], L.B. Zhang[7], M. Long[8], S. Bhilare[9], V. Kanhangad[9], A. Costa-Pazo[10], E. Vazquez-Fernandez[10], D. Pérez-Cabo[10], J. J. Moreira-Pérez[10], D. González-Jiménez[10], A. Mohammadi[11,12], S. Bhattacharjee[12], S. Marcel[12], S. Volkova[13], Y. Tang[14], N. Abe[15], L. Li[16], X. Feng[16], Z. Xia[16], X. Jiang[16], S. Liu[17], R. Shao[17], P. C. Yuen[17], W. R. Almeida[18], F. Andaló[18], R. Padilha[18], G. Bertocco[18], W. Dias[18], J. Wainer[18], R. Torres[18], A. Rocha[18], M. A. Angeloni[19], G. Folego[19], A. Godoy[19] and A. Hadid[1,16]

[1]University of Oulu (FI), [2]INRS-EMT, University of Quebec (CA), [3]University of Ouargla (DZ), [4]University of Biskra (DZ), [5]University of the Basque Country (ES), [6]University of Valenciennes (FR), [7]Hunan University (CN), [8]Changsha University of Science and Technology (CN), [9]Indian Institute of Technology Indore (IN), [10]GRADIANT (ES), [11]Ecole Polytechnique Federale de Lausanne (CH), [12]Idiap Research Institute (CH), [13]Vologda State University (RU), [14]Shenzhen University (CN), [15]FUJITSU LABORATORIES LTD LTD (JP), [16]Northwestern Polytechnical University (CN), [17]Hong Kong Baptist University (HK), [18]University of Campinas (BR), [19]CPqD (BR).

Abstract

In recent years, software-based face presentation attack detection (PAD) methods have seen a great progress. However, most existing schemes are not able to generalize well in more realistic conditions. The objective of this competition is to evaluate and compare the generalization performances of mobile face PAD techniques under some real-world variations, including unseen input sensors, presentation attack instruments (PAI) and illumination conditions, on a larger scale OULU-NPU dataset using its standard evaluation protocols and metrics. Thirteen teams from academic and industrial institutions across the world participated in this competition. This time typical liveness detection based on physiological signs of life was totally discarded. Instead, every submitted system relies practically on some sort of feature representation extracted from the face and/or background regions using hand-crafted, learned or hybrid descriptors. Interesting results and findings are presented and discussed in this paper.

1. Introduction

The vulnerabilities of face based biometric systems to presentation attacks have been widely recognized but still there is a lack of generalized software-based PAD methods performing robustly in practical (mobile) authentication scenarios [10, 26]. In recent years, many face PAD methods have been proposed and remarkable results have been reported on the existing benchmark datasets. For instance, in the first [6] and second [8] face PAD competitions, several methods have achieved perfect performances on the used databases. However, recent studies [4, 10, 22, 26] have revealed that most of these methods are not able to generalize well in more realistic scenarios, thus face PAD is still an unsolved problem in unconstrained operating conditions.

Focused large scale evaluations on the generalization of face PAD have not been conducted or organized after the issue was first pointed out by de Freitas Pereira *et al.* [10] in 2013. The aim of this competition is to compare and evaluate the generalization abilities of state-of-the-art PAD schemes under some real-world variations, including camera, attack, and illumination. Compared with the previous competitions, we observe that the number of participants has increased from six and eight in the first and second competitions, respectively, to 13 in this competition. Moreover, in the previous competitions, all the participated teams were from academic institutes and universities, while in this competition, we have registered the participation of three companies. This also highlights the importance of the topic for both academia and industry. The name and the affiliation of the participating teams are summarized in Table 1.

2. Database and evaluation protocols

The competition was carried out on the recent and publicly available[1] OULU-NPU face presentation attack database [5]. The dataset consists of 4950 real access and attack videos that were recorded using front facing cameras

[1]The dataset was not yet released at the time of the competition.

Table 1: Names and affiliations of the participating systems

Team	Affiliations
Baseline	University of Oulu, Finland
MBLPQ	University of Ouargla, Algeria
PML	University of Biskra, Algeria University of the Basque Country, Spain University of Valenciennes, France
Massy_HNU	Changsha University of Science and Technology Hunan University, China
MFT-FAS	Indian Institute of Technology Indore, India
GRADIANT	Galician Research and Development Center in Advanced Telecommunications, Spain
Idiap	Ecole Polytechnique Federale de Lausanne Idiap Research Institute, Switzerland
VSS	Vologda State University, Russia
SZUCVI	Shenzhen University, China.
MixedFasNet	FUJITSU laboratories LTD, Japan
NWPU	Northwestern Polytechnical University, China
HKBU	Hong Kong Baptist University, China
Recod	University of Campinas, Brazil
CPqD	CPqD, Brazil

of six different smartphones in the price range from €250 to €600 (see Figure 1). The real videos and attack materials were collected in three sessions with different illumination conditions (Session 1, Session 2 and Session 3). In order to simulate realistic mobile authentication scenarios, the video length was limited to five seconds and the subjects were asked to hold the mobile device like they were being authenticated but without deviating too much from their natural posture while normal device usage. The attack types considered in the OULU-NPU database are print and video-replay. These attacks were created using two printers (Printer 1 and Printer 2) and two display devices (Display 1 and Display 2). The videos of the real accesses and attacks, corresponding to the 55 subjects, are divided into three subject-disjoint subsets for training, development and testing with 20, 15 and 20 users, respectively.

During the system development phase of two months, the participants were given access to the labeled videos of the training and the development sets that were used to train and fine tune the devised face PAD methods. In addition to the provided training set, the participants were allowed to use external data to train their algorithms. In the evaluation phase of two weeks, the performances of the developed systems were reported on anonymized and unlabeled test video files. To assess the generalization of the developed face PAD methods, four protocols were used:

Protocol I: The first protocol is designed to evaluate the generalization of the face PAD methods under previously unseen environmental conditions, namely illumination and background scene. As the database is recorded in three sessions with different illumination condition and location, the train, development and evaluation sets are constructed using video recordings taken in different sessions.

Protocol II: The second protocol is designed to evaluate the effect of attacks created with different printers or displays on the performance of the face PAD methods as they may suffer from new kinds of artifacts. The effect of attack variation is assessed by introducing previously unseen print and video-replay attacks in the test set.

Protocol III: One of the critical issues in face PAD and image classification in general is sensor interoperability. To study the effect of the input camera variation, a Leave One Camera Out (LOCO) protocol is used. In each iteration, the real and the attack videos recorded with five smartphones are used to train and tune the algorithms, and the generalization of the models is assessed using the videos recorded with the remaining smartphone.

Protocol IV: In the most challenging protocol, all above three factors are considered simultaneously and generalization of face PAD methods are evaluated across previously unseen environmental conditions, attacks and input sensors.

Table 2 gives a detailed information about the video recordings used in the train, development and test sets of each test scenario. For every protocol, the participants were asked to provide separate score files for the development and test sets containing a single score for each video.

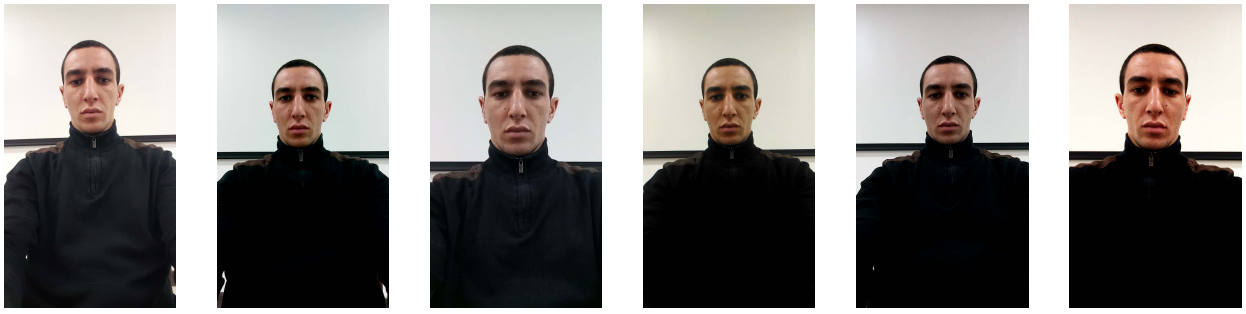
For the performance evaluation, we selected the recently standardized ISO/IEC 30107-3 metrics [15], Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER):

$$APCER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - Res_i) \tag{1}$$

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}} \tag{2}$$

where, N_{PAI}, is the number of the attack presentations for the given PAI, N_{BF} is the total number of the bona fide presentations. Res_i takes the value 1 if the ith presentation is classified as an attack presentation and 0 if classified as bona fide presentation. These two metrics correspond to the False Acceptance Rate (FAR) and False Rejection Rate (FRR) commonly used in the PAD related literature. However, APCER_{PAI} is computed separately for each PAI (e.g. print or display) and the overall PAD performance corresponds to the attack with the highest APCER, i.e. the "worst case scenario".

To summarize the overall system performance in a single value, the Average Classification Error Rate (ACER) is used, which is the average of the APCER and the BPCER at the decision threshold defined by the Equal Error Rate (EER) on the development set:



(a) Phone 1 (b) Phone 2 (c) Phone 3 (d) Phone 4 (e) Phone 5 (f) Phone 6

Figure 1: Sample images showing the image quality of the different camera devices.

Table 2: The detailed information about the video recordings in the train, development and test sets of each protocol

Protocol	Subset	Session	Phones	Users	Attacks created using	# real videos	# attack videos	# all videos
Protocol I	Train	Session 1,2	6 Phones	1-20	Printer 1,2; Display 1,2	240	960	1200
	Dev	Session 1,2	6 Phones	21-35	Printer 1,2; Display 1,2	180	720	900
	Test	Session 3	6 Phones	36-55	Printer 1,2; Display 1,2	120	480	600
Protocol II	Train	Session 1,2,3	6 Phones	1-20	Printer 1; Display 1	360	720	1080
	Dev	Session 1,2,3	6 Phones	21-35	Printer 1; Display 1	270	540	810
	Test	Session 1,2,3	6 Phones	36-55	Printer 2; Display 2	360	720	1080
Protocol III	Train	Session 1,2,3	5 Phones	1-20	Printer 1,2; Display 1,2	300	1200	1500
	Dev	Session 1,2,3	5 Phones	21-35	Printer 1,2; Display 1,2	225	900	1125
	Test	Session 1,2,3	1 Phone	36-55	Printer 1,2; Display 1,2	60	240	300
Protocol IV	Train	Session 1,2	5 Phones	1-20	Printer 1; Display 1	200	400	600
	Dev	Session 1,2	5 Phones	21-35	Printer 1; Display 1	150	300	450
	Test	Session 3	1 Phone	36-55	Printer 2; Display 2	20	40	60

$$ACER = \frac{\max\limits_{PAI=1...S}(APCER_{PAI}) + BPCER}{2} \quad (3)$$

where S is the number of the PAIs. In Protocols III and IV, these measures (i.e. APCER, BPCER and ACER) are computed separately for each mobile phone, and the average and standard deviation are taken over the folds to summarize the results. Since the attack potential of the PAIs may vary across the different folds, the overall APCER does not necessarily correspond to the highest mean APCER_{PAI}.

3. Brief description of the participating systems

Baseline In addition to the training and development data, the participants were given the source code of the baseline face PAD method that could be freely improved or used as it is in the final systems. The baseline face PAD method is based on the color texture technique [4], which has shown promising generalization abilities. The steps of the baseline method are as follows. First, the face is detected, cropped and normalized into 64×64 pixels. Then, the RGB face image is converted into HSV and YCbCr color spaces. The local binary pattern (LBP) texture features [20] are extracted from each channel of the color spaces. The resulting feature vectors are concatenated into an enhanced feature vector which is fed into a Softmax classifier. The final score for each video is computed by averaging the output scores of all frames.

MBLPQ After the face pre-processing step, cropped and normalized 128×128 face images are obtained. These RGB face images are then converted into YCbCr color space and divided into multiple blocks [3]. The local phase quantization (LPQ) [23] features are extracted from each block and then concatenated into a single feature vector. The LPQ features extracted from each channel are concatenated to form the overall face representation. Each video is represented with a single vector by averaging feature vectors extracted from the first ten frames. The score for each video is then computed using a Softmax classifier.

PML After sampling the video every four frames, the detected face is aligned, cropped and resized to 224×244 image and the resulting RGB image is converted to HSV color space. The face image is then transformed into a Pyramid Multi-Level (PML) representation [2] (six levels in our case). LPQ descriptor [23] is applied on each level and the resulting features are concatenated into a single feature

vector. Subsequently, the features extracted from the three channels are concatenated to form the overall face representation. Each video is represented with a single vector by averaging all PML vectors over the sampled frames. The aggregated feature vector is fed into a non-linear support vector machine (SVM) classifier to obtain the final score.

Massy_HNU First, the faces are detected, cropped and normalized into 64×64 images using the provided eye locations. Then, the RGB face images are converted into HSV and YCbCr color spaces and guided image filtering [13] is applied on the face images. After dividing the normalized facial images into 32×32 blocks with 16 pixels overlap, LBP features [20] are extracted from each channel of the color spaces. The LBP coding maps are calculated for each channel, and chromatic co-occurrence matrices [21] are extracted from the LBP coding maps as final features vectors. A Softmax classifier is used to compute the scores for 30 randomly selected frames. The scores are averaged to obtain the final score for a video.

MFT-FAS This face PAD method relies essentially on the texture information. First, the mean frame of the video is computed by averaging intensities of the corresponding pixels in the input video sequence. In order to reduce the computation time, the resulting mean video frame is resized to 480×270×3 pixels. Inspired by the baseline method [4], the resized image is then mapped to the YCbCr color space. Each channel of the resultant mean image in the YCbCr space is partitioned vertically into two parts resulting in a total of six sub-images. The binarized statistical image features (BSIF) [17] are then computed using two filter sets of size 5×5 and 13×13. This process generates four 256-dimensional feature vectors for every channel. Finally, the individual feature vectors are concatenated to obtain the final feature vector, which is fed to a Softmax classifier.

GRADIANT GRADIANT system fuses color [4], texture and motion information, exploiting both HSV and YCbCr color spaces. Computational restrictions have been taken into account during algorithm design, so that the final system can operate fully embedded into mobile devices. The system extracts dynamic information over a given video sequence and maps the temporal variations into a single image. This method is applied separately to all channels in both HSV and YCbCr color spaces, thus resulting in a pair of 3-channel images. For each image, ROI is cropped based on eye positions over the sequence and rescaled to 160×160 pixels. Each ROI is divided into 3×3 and 5×5 rectangular regions from which uniform LBP histogram features are extracted and concatenated into two 6018-length feature vectors. Recursive Feature Elimination [12] is applied to reduce feature dimensionality from 6018 to 1000. SVM-based supervised classification is performed for each feature vector, and fusion through the sum rule is applied to the resulting scores in order to obtain the decision result. Two versions were submitted: the first one (GRADIANT) was trained only with the train subset available in each protocol, while the second one (GRADIANT_extra) was trained also on external data, namely Replay-Mobile database [9].

Idiap This method is a score fusion of three other face PAD methods: Motion [1], Texture (LBP) [7] and Quality [11, 26]. The scores of these systems are first calibrated using logistic regression separately for each method. Then, the calibrated scores are used as three-dimensional features for the fusion system. The features are mean and standard deviation normalized so that their mean is zero and their standard deviation is one for each dimension. Then, a Gaussian mixture model (GMM) model with four mixtures is trained on the bona-fide samples only. The number of mixtures was chosen on the development set. All the calibration, normalization, and GMM training are done using the training set. Later, log-likelihood of a test sample belonging to the bona-fide GMM is reported as the final score.

VSS In the VSS face PAD method, first, the faces are detected, cropped and normalized into 128×128 gray-scale images, which are then fed into a CNN. The architecture of the CNN model consists of five convolution layers and two fully connected layers. Each convolution layer is combination of two independent convolution parts calculated from the input. The fully connected layers have 512 and two dimensions, respectively. The output of the second fully connected layer is used as an input to a Softmax classifier. Two sets of scores have been submitted. In the first set (VSS), only the provided train videos were used to train the model, while extra-training data was used to train the model in the second set (VSS_extra). The real faces, in this extra-data, were taken from the CASIA-WebFace database, while the fake faces were captured from magazines and movies displayed on monitor and TV using two mobile devices.

SZCVI After sampling the videos every six frames, the video frames were resized into 216×384 images and fed into a CNN model. The architecture of this model consists of five convolutional layers and one fully connected layer. The convolutional layers were inspired by the VGG model [24]. The scores of the sampled frames were averaged to obtain the final score for each video file.

MixedFASNet In the MixedFASNet method, the face images picked up every five frames are cropped and normalized to 64×64 images. These images are then converted to the HSV color space. To emphasize the differences between the real and fake face images, contrast limited adaptive histogram equalization [28] is applied. The feature extraction has some deep learning architectures trained with not only face images but also the specific background image patches. The number of layers is over 30 and several models are trained using different inputs, e.g. original and contrast-enhanced images. The extracted features are finally fed into MLP classifier and the final score for a video is computed

by averaging the scores of the sampled frames.

NWPU Since texture features have the potential to distinguish the real and fake faces, we built an end-to-end deep learning model, which auto-extracts the LBP features [20] from the convolutional layers. Compared with many previous deep learning methods, our method contains fewer parameters, thus does not require an enormous training dataset. First, the faces are detected, cropped and normalized into 64×64 RGB images, which are then fed into the network. The models are trained on the provided training set using back propagation algorithm and the error terms of the models are generated by the LBP features extracted from the convolutional layers. In testing stage, the obtained LBP features are fed into an SVM classifier and the final score for a video is computed by averaging the scores of individual video frames.

HKBU The basic idea of the proposed method is to extract the intrinsic properties of the presentation attack, which are located on the subtle differences of printed matter or screen. Three appearance-based features, namely image distortion analysis (IDA) features [26], multi-scale local binary patterns (msLBP) [19] and deep feature [16] are fused to give a robust representation. The steps of the method are as follows. First, the faces are located, aligned and then normalized into 256×256 RGB images. Then, the IDA and msLBP features are extracted from the preprocessed face images as distortion and texture components. To further undermine the fine-grained visual information, deep features are extracted through the lower convolutional layers (conv5) of a AlexNet model trained on the ImageNet object classification database. The lower convolution layers are used because they represent low-level subtle appearance features. Finally, multiple kernel learning [16] is employed to fuse the three types of features in a kernelised (RBF kernel) manner so that each component can optimally contribute to the final classification according to its discriminability. The final score is based solely on the first video frame.

Recod The SqueezeNet [14], which was originally trained with ImageNet, is the foundation of the Recod method. Since this CNN was trained to perform a different task, a transfer learning strategy is applied to fine-tune the pre-trained network to the binary problem of PAD. For this, two datasets were used: CASIA [27] and UVAD [22]. The faces were first detected and resized to 224×224 pixels. To select the best model during the fine tuning process, several checkpoints have been evaluated and the one with highest accuracy on the development sets is chosen, considering several frames from each video. For each protocol of the competition, the previously selected network was further fine tuned on the corresponding training sets. During the second fine tuning process, the best ten checkpoints were selected, based on their average error rate. These checkpoints are stored and used to generate the scores for the competition. From each video of the test set, roughly every seventh frame is selected and forwarded through the ten CNNs. The resulting scores are averaged to obtain the final score. To further improve the performance, the obtained score is fused with the aggregated score of the provided baseline method.

Table 3: Categorization of the proposed systems based on hand-crafted, learned and hybrid features

Category	Teams
Hand-crafted features	Baseline, MBLPQ, PML, Massy_HNU, MFT-FAS, GRADIANT, Idiap
Learned features	VSS, SZCVI, MixedFASNet
Hybrid features	NWPU, HKBU, Recod, CPqD

CPqD The CPqD method is based on the Inception-v3 Convolutional Neural Network (CNN) model [25]. This model is trained for object classification on the ImageNet database. To adapt it for face PAD problem, the pre-trained model is modified by replacing the last layer with a one-way fully connected layer and a sigmoid activation function. The face regions are cropped based on the eye locations and resized into 224×224 RGB images. The modified model is then fine-tuned on the training face images (sampled at 3 fps) using binary cross-entropy loss function and Adam optimizer [18]. Since the provided database is unbalanced, class weights inversely proportional to class frequencies are adopted. To avoid overfitting, training is limited to ten epochs, and data augmentation is employed. The model with lowest EER on the development set among all epochs is selected. A single score for a video is obtained by averaging scores obtained on sampled frames. For each video, the final score is given by the average between the scores of the described method and the provided baseline method.

4. Results and analysis

In this competition, typical "liveness detection" was not adopted as none of the submitted systems is explicitly aiming at detecting physiological signs of life, like eye blinking, facial expression changes and mouth movements. Instead, every proposed face PAD algorithm relies on one or more types of feature representations extracted from the face and/or the background regions. The used descriptors can be categorized into three groups (see Table 3): hand-crafted, learned and hybrid (fusion of hand-crafted and learned). The performances of the submitted systems under the four test protocols are reported in Tables 4, 5, 6 and 7.

It appears that the analysis of mere grayscale or even RGB images does not result in particularly good generalization. In the case of hand-crafted features, every algorithm is based on the recently proposed color texture analysis [4] in which RGB images are converted into HSV and/or YCbCr color spaces prior feature extraction. The only well-generalizing feature learning based method, MixedFASNet,

Table 4: The performance of the proposed methods under different illumination and location conditions (Protocol I)

Methods	Dev	Test				
	EER(%)	Display	Print	Overall		
		APCER(%)	APCER(%)	APCER(%)	BPCER(%)	ACER(%)
GRADIANT_extra	0.7	7.1	3.8	7.1	5.8	**6.5**
CPqD	0.6	1.3	2.9	2.9	10.8	6.9
GRADIANT	1.1	**0.0**	1.3	1.3	12.5	6.9
Recod	2.2	3.3	0.8	3.3	13.3	8.3
MixedFASNet	1.3	**0.0**	**0.0**	**0.0**	17.5	8.8
PML	0.6	7.5	11.3	11.3	9.2	10.2
Baseline	4.4	5.0	1.3	5.0	20.8	12.9
Massy_HNU	1.1	5.4	3.3	5.4	20.8	13.1
HKBU	4.3	9.6	7.1	9.6	18.3	14.0
NWPU	**0.0**	8.8	7.5	8.8	21.7	15.2
MFT-FAS	2.2	0.4	3.3	3.3	28.3	15.8
MBLPQ	2.2	31.7	44.2	44.2	**3.3**	23.8
Idiap	5.6	9.6	13.3	13.3	40.0	26.7
VSS	12.2	20.0	12.1	20.0	41.7	30.8
SZUCVI	16.7	11.3	**0.0**	11.3	65.0	38.1
VSS_extra	24.0	9.6	11.3	11.3	73.3	42.3

Table 5: The performance of the proposed methods under novel attacks (Protocol II)

Methods	Dev	Test				
	EER(%)	Display	Print	Overall		
		APCER(%)	APCER(%)	APCER(%)	BPCER(%)	ACER(%)
GRADIANT	0.9	**1.7**	3.1	**3.1**	**1.9**	**2.5**
GRADIANT_extra	0.7	6.9	**1.1**	6.9	2.5	4.7
MixedFASNet	1.3	6.4	9.7	9.7	2.5	6.1
SZUCVI	4.4	3.9	3.3	3.9	9.4	6.7
MFT-FAS	2.2	10.0	11.1	11.1	2.8	6.9
PML	0.9	11.4	9.4	11.4	3.9	7.6
CPqD	2.2	9.2	14.7	14.7	3.6	9.2
HKBU	4.6	13.9	12.5	13.9	5.6	9.7
Recod	3.7	13.3	15.8	15.8	4.2	10.0
MBLPQ	1.9	5.6	19.7	19.7	6.1	12.9
Baseline	4.1	15.6	22.5	22.5	6.7	14.6
Massy_HNU	1.3	16.1	26.1	26.1	3.9	15.0
Idiap	8.7	21.7	7.5	21.7	11.1	16.4
NWPU	**0.0**	12.5	5.8	12.5	26.7	19.6
VSS	14.8	25.3	13.9	25.3	23.9	24.6
VSS_extra	23.3	36.1	33.9	36.1	33.1	34.6

is using HSV images as input, whereas the networks operating on gray-scale or RGB images are not generalizing very well. On the other hand, it is worth mentioning that VSS and SZCVI architectures consist only of five convolutional layers, whereas the MixedFASNet is much deeper. The best performing hybrid methods, Recod and CPqD, are fusing the scores of their deep learning based method and the provided baseline in order to increase the generalization capabilities. Since only the scores of hybrid systems were provided, the robustness of the proposed fine-tuned CNN models operating on RGB images remains unclear. Among the methods solely based on RGB image analysis, HKBU fusing IDA, LBP and deep features is the only one that generalizes fairly well across the four protocols.

Table 6: The performance of the proposed methods under input camera variations (Protocol III)

Methods	Dev	Test				
	EER(%)	Display	Print	Overall		
		APCER(%)	APCER(%)	APCER(%)	BPCER(%)	ACER(%)
GRADIANT	0.9±0.4	1.0±1.7	2.6±3.9	2.6±3.9	**5.0±5.3**	**3.8±2.4**
GRADIANT_extra	0.7±0.2	1.4±1.9	**1.4±2.6**	**2.4±2.8**	5.6±4.3	4.0±1.9
MixedFASNet	1.4±0.5	1.7±3.3	5.3±6.7	5.3±6.7	7.8±5.5	6.5±4.6
CPqD	0.9±0.4	4.4±3.4	5.0±6.1	6.8±5.6	8.1±6.4	7.4±3.3
Recod	2.9±0.7	4.2±3.8	8.6±14.3	10.1±13.9	8.9±9.3	9.5±6.7
MFT-FAS	0.8±0.4	**0.8±0.9**	10.8±18.1	10.8±18.1	9.4±12.8	10.1±9.9
Baseline	3.9±0.7	9.3±4.3	11.8±10.8	14.2±9.2	8.6±5.9	11.4±4.6
HKBU	3.8±0.3	7.9±5.8	9.9±12.3	12.8±11.0	11.4±9.0	12.1±6.5
SZUCVI	7.0±1.6	10.0±8.3	7.5±9.5	12.1±10.6	16.1±8.0	14.1±4.4
PML	1.1±0.3	8.2±12.5	15.3±22.1	15.7±21.8	15.8±15.4	15.8±15.1
Massy_HNU	1.9±0.6	5.8±5.4	19.0±26.7	19.3±26.5	14.2±13.9	16.7±10.9
MBLPQ	2.3±0.6	5.8±5.8	12.9±4.1	12.9±4.1	21.9±22.4	17.4±10.3
NWPU	**0.0±0.0**	1.9±0.7	1.9±3.3	3.2±2.6	33.9±10.3	18.5±4.4
Idiap	7.9±1.9	8.3±3.0	9.3±10.0	12.9±8.2	26.9±24.4	19.9±11.8
VSS	14.6±0.8	21.4±7.7	13.8±7.0	21.4±7.7	25.3±9.6	23.3±2.3
VSS_extra	25.9±1.7	25.0±11.4	32.2±27.9	40.3±22.2	35.3±27.4	37.8±6.8

Table 7: The performance of the proposed methods under environmental, attack and camera device variations (Protocol IV)

Methods	Dev	Test				
	EER(%)	Display	Print	Overall		
		APCER(%)	APCER(%)	APCER(%)	BPCER(%)	ACER(%)
GRADIANT	1.1±0.3	**0.0±0.0**	5.0±4.5	5.0±4.5	15.0±7.1	**10.0±5.0**
GRADIANT_extra	1.1±0.3	27.5±24.2	5.8±4.9	27.5±24.2	**3.3±4.1**	15.4±11.8
Massy_HNU	1.0±0.4	20.0±17.6	26.7±37.5	35.8±35.3	8.3±4.1	22.1±17.6
CPqD	2.2±1.7	16.7±16.0	24.2±39.4	32.5±37.5	11.7±12.1	22.1±20.8
Recod	3.7±0.7	20.0±19.5	23.3±40.0	35.0±37.5	10.0±4.5	22.5±18.2
MFT-FAS	1.6±0.7	**0.0±0.0**	12.5±12.9	12.5±12.9	33.3±23.6	22.9±8.3
MixedFASNet	2.8±1.1	10.0±7.7	4.2±4.9	10.0±7.7	35.8±26.7	22.9±15.2
Baseline	4.7±0.6	19.2±17.4	22.5±38.3	29.2±37.5	23.3±13.3	26.3±16.9
HKBU	5.0±0.7	16.7±24.8	21.7±36.7	33.3±37.9	27.5±20.4	30.4±20.8
VSS	11.8±0.8	21.7±8.2	9.2±5.8	21.7±8.2	44.2±11.1	32.9±5.8
MBLPQ	3.6±0.7	35.0±25.5	45.0±25.9	49.2±27.8	24.2±27.8	36.7±4.7
NWPU	**0.0±0.0**	30.8±7.4	6.7±11.7	30.8±7.4	44.2±23.3	37.5±9.4
PML	0.8±0.3	59.2±24.2	38.3±41.7	61.7±26.4	13.3±13.7	37.5±14.1
SZUCVI	9.1±1.6	**0.0±0.0**	**0.8±2.0**	**0.8±2.0**	80.8±28.5	40.8±13.5
Idiap	6.8±0.8	26.7±35.2	13.3±8.2	33.3±30.4	54.2±12.0	43.8±20.4
VSS_extra	21.1±2.7	13.3±17.2	15.8±21.3	25.8±20.8	70.0±22.8	47.9±12.1

In general, the submitted systems are processing each video frame (of a video sequence) independently then the final score for a given video is obtained by averaging the resulting scores of individual frames. None of the deep learning or hybrid methods were exploiting temporal variations but in the case of hand-crafted features two different temporal aggregation approaches were proposed for encoding the dynamic information within a video sequence, e.g. motion. MBLPQ and PML averaged the feature vectors over the sampled frames, whereas GRADIANT and MFT-FAS map the temporal variations into a single image prior feature extraction. The approach by GRADIANT turned out to be particularly successful as the achieved performance was simply the best and most consistent across all the four

protocols.

In this competition, the simple color texture based face descriptions were very powerful compared to deep learning based methods, of which the impressive results by GRADIANT are a good example. On the other hand, the current (public) datasets may not probably provide enough data for training CNNs from scratch or even fine-tuning the pre-trained models to their full potential. NWPU extracted LBP features from convolutional layers in order to reduce the number of trainable parameters, thus relieving the need for enormous training sets. Unfortunately, the method was not able to generalize well on the evaluation set.

Few teams used additional public and/or proprietary datasets for training and tuning their algorithms. VSS team augmented the subset of real subjects with CASIA-WebFace and collected own attack samples. The usefulness of these external datasets remains unclear because the grayscale image analysis based face PAD method was not very efficient. Recod used publicly available datasets for fine tuning the pre-trained network but the resulting generalization was comparable to similar method, CPqD, not using any extra-data. GRADIANT submitted two systems with and without external training data. Improved BPCER was obtained in unseen acquisition conditions but APCER is much better in general when using only the provided OULU-NPU training data.

Since unseen attack scenarios will be definitely experienced in operation, the problem of PAD could be easily ideally solved using one-class classifiers for modeling the variations of the only known class, i.e. bona-fide. Idiap method is based on the idea of anomaly detection but it lacked generalization mainly because the individual grayscale image analysis based methods were performing poorly[2]. Thus, one-class modeling would be worth investigating when combined with more robust feature representations.

Few general observations can be concluded based on the results of protocols I, II and III assessing the generalization of the PAD method across unseen conditions, i.e. acquisition conditions, attack types and input sensors, separately:

Protocol I: In general, a significant increase in BPCER can be noticed compared to APCER when the PAD systems are operating in new acquisition conditions. The reason behind this may be in the data collection principles of the OULU-NPU dataset. Legitimate users have to be verified in various conditions, while attackers aim probably at high-quality attack presentation in order to increase the chance of successfully fooling a face biometric system. From the usability point of view, the bona-fide samples were collected in three sessions with different illumination. In contrast, the bona-fide data matching to each session was used to create face artifacts but the attacks themselves were always launched with short standoff and captured in the same laboratory setup. Thus, the intrinsic properties of the attacks do not vary too much across the different sessions.

Protocol II: In most cases, previously unseen attack leads into dramatic increase in APCER, which is expected as only one PAI of each print and video-replay attacks is provided for training and tuning purposes.

Protocol III: It is also interesting to notice that the standard deviation of APCER across different input sensors is much larger in the case of print attacks compared to video-replay attacks, which suggests that the nature of print attacks seems to vary more although both attack types can be detected equally well on average.

Based on the results of the protocol IV, it is much harder to make general conclusions because all the factors are combined and different approaches seem to be more robust to different covariates. The last protocol reveals, however, that none of the methods is able to achieve either reasonable trade-off between usability and security. For instance, in the case of GRADIANT, either the APCER or BPCER of the two systems is too high for practical applications. Nevertheless, the overall performance of GRADIANT, MixedFASNET, CPqD and Recod is very impressive considering the conditions of the competition and the OULU-NPU dataset.

5. Conclusion

The deployment of face recognition applications in mobile authentication has created a necessity for robust face PAD solutions. Despite the recent progress, the existing mobile face PAD methods have shown lack of generalization in real-world operating conditions. This was the first large-scale competition to evaluate the generalization capability of state-of-the-art countermeasures for face presentation attacks on a publicly available database.

Also this competition was a huge success in consolidating and benchmarking the state-of-the-art approaches in face PAD. The number of participants, from both academic and industrial institutions, is growing with respect to previous competitions. As a matter of fact, 13 entries were submitted from ten academic institutes and three companies. All submitted systems relied on one or more features of three different kinds: hand-crafted, learned and hybrid, i.e. fusion of both. All in all, though deep learning-based methods achieve impressive results, hand-crafted features coupled with appropriate color spaces can generalize remarkably well not only against previously unseen input cameras but also across environmental conditions and attack types.

A possible future study would be combining match scores with both PAD and quality measures to improve the resilience of face verification systems. The OULU-NPU database could be expanded to increase variability in user demographics and mobile devices, and introduce uncontrolled outdoor conditions and new attacks, e.g. 3D masks.

[2]Idiap submitted also the scores of the individual sub-systems.

Acknowledgement

The financial support of the Academy of Finland, Infotech Oulu, the Nokia Foundation and the Finnish Foundation for Technology Promotion is fully acknowledged.

References

[1] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: A public database and a baseline. In *International Joint Conference On Biometrics (IJCB)*, 2011. 4

[2] S. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, and A. Hadid. Pyramid multi-level features for facial demographic estimation. *Expert Systems with Applications*, 80:297 – 310, 2017. 3

[3] A. Benlamoudi, D. Samai, A. Ouafi, S. E. Bekhouche, A. Taleb-Ahmed, and A. Hadid. Face spoofing detection using multi-level local phase quantization (ML-LPQ). In *International Conference on Automatic control, Telecommunications and Signals (ICATS)*, 2015. 3

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *IEEE International Conference on Image Processing (ICIP)*, 2015. 1, 3, 4, 5

[5] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017. 1

[6] M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, F. Roli, J. Yan, D. Yi, Z. Lei, Z. Zhang, S. Li, W. Schwartz, A. Rocha, H. Pedrini, J. Lorenzo-Navarro, M. Castrillon-Santana, J. Määttä, A. Hadid, and M. Pietikäinen. Competition on counter measures to 2-D facial spoofing attacks. In *International Joint Conference on Biometrics (IJCB)*, 2011. 1

[7] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference on Biometrics Special Interests Group (BIOSIG)*, 2012. 4

[8] I. Chingovska, J. Yang, Z. Lei, D. Yi, S. Li, O. Kahm, C. Glaser, N. Damer, A. Kuijper, A. Nouak, J. Komulainen, T. Pereira, S. Gupta, S. Khandelwal, S. Bansal, A. Rai, T. Krishna, D. Goyal, M.-A. Waris, H. Zhang, I. Ahmad, S. Kiranyaz, M. Gabbouj, R. Tronci, M. Pili, N. Sirena, F. Roli, J. Galbally, J. Ficrrcz, A. Pinto, H. Pedrini, W. Schwartz, A. Rocha, A. Anjos, and S. Marcel. The 2nd competition on counter measures to 2D face spoofing attacks. In *International Conference on Biometrics (ICB)*, 2013. 1

[9] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *International Conference on Biometrics Special Interests Group (BIOSIG)*, 2016. 4

[10] T. de Freitas Pereira, A. Anjos, J. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *International Conference on Biometrics (ICB)*, 2013. 1

[11] J. Galbally, S. Marcel, and J. Fierrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing*, 23(2):710–724, 2014. 4

[12] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002. 4

[13] K. He, J. Sun, and X. Tang. Guided image filtering. In *European Conference on Computer Vision (ECCV)*, 2010. 4

[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*, 2016. 5

[15] ISO/IEC JTC 1/SC 37 Biometrics. Information technology – Biometric presentation attack detection – Part 1: Framework. International Organization for Standardization, 2016. 2

[16] A. Jain, S. V. Vishwanathan, and M. Varma. Spf-gmkl: generalized multiple kernel learning with a million kernels. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 750–758, 2012. 5

[17] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *IEEE International Conference on Pattern Recognition (ICPR)*, 2012. 4

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[19] J. Määttä, A. Hadid, and M. Pietikainen. Face spoofing detection from single images using micro-texture analysis. In *International joint conference on Biometrics (IJCB)*, 2011. 5

[20] T. Ojala, M. Pietikäinen, and T. Mäenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 3, 4, 5

[21] C. Palm. Color texture classification by integrative co-occurrence matrices. *Pattern Recognition*, 37(5):965 – 976, 2004. 4

[22] A. Pinto, W. R. Schwartz, H. Pedrini, and A. Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE Transactions on Information Forensics and Security*, 10(5):1025–1038, 2015. 1, 5

[23] E. Rahtu, J. Heikkilä, V. Ojansivu, and T. Ahonen. Local phase quantization for blur-insensitive image analysis. *Image and Vision Computing*, 30(8):501 – 512, 2012. 3

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 13(13):1–13, 2015. 4

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 5

[26] D. Wen, H. Han, and A. Jain. Face Spoof Detection with Image Distortion Analysis. *IEEE Transactions on Information Forensic and Security*, 10(4):746–761, 2015. 1, 4, 5

[27] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Li. A face antispoofing database with diverse attacks. In *ICB*, 2012. 5

[28] K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pages 474–485. 1994. 4