Review of Face Presentation Attack Detection Competitions

Jukka Komulainen, Zinelabidine Boulkenafet and Zahid Akhtar

Abstract Face presentation attack detection has received increasing attention ever since the vulnerabilities to spoofing have been widely recognized. The state of the art in software-based face anti-spoofing has been assessed in three international competitions organized in conjunction with major biometrics conferences in 2011, 2013 and 2017, each introducing new challenges to the research community. In this chapter, we present the design and results of the three competitions. The particular focus is on the latest competition, where the aim was to evaluate the generalization abilities of the proposed algorithms under some real-world variations faced in mobile scenarios, including previously unseen acquisition conditions, presentation attack instruments and sensors. We also discuss the lessons learnt from the competitions and future challenges in the field in general.

1 Introduction

Spoofing (or presentation attacks as defined in the recent ISO/IEC 30107-3 standard [24]) poses serious security issue to biometric systems in general but face recognition systems in particular are easy to be deceived using images of the targeted person published in the web or captured from distance. Many works (e.g., [14, 30, 35]) have concluded that face biometric systems, even those presenting a

Jukka Komulainen
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
e-mail: jukka.komulainen@iki.fi

Zinelabidine Boulkenafet
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
e-mail: zinelabidine.boulkenafet@oulu.fi

Zahid Akhtar
INRS-EMT, University of Quebec, Canada
e-mail: zahid.akhtar.momin@emt.inrs.ca

high recognition performance, are vulnerable to attacks launched with different Presentation Attack Instruments (PAI), such as prints, displays and wearable 3D masks. The vulnerability to presentation attacks (PA) is one of the main reasons to the lack of public confidence in (face) biometrics. Also, face recognition based user verification is being increasingly deployed even in high-security level applications, such as mobile payment services. This has created a necessity for robust solutions to counter spoofing.

One possible solution is to include a specific Presentation Attack Detection (PAD) component into a biometric system. PAD (commonly referred to also as anti-spoofing, spoof detection or liveness detection) aims at automatically differentiating whether the presented biometric sample originates from a living legitimate subject or not. PAD schemes can be broadly categorized into two groups: hardware-based and software-based methods. Hardware-based methods introduce some custom sensor into the biometric system that is designed specifically for capturing specific intrinsic differences between a valid living biometric trait and others. Software-based techniques exploit either only the same data that is used for the actual biometric purposes or additional data captured with the standard acquisition device.

Ever since the vulnerabilities of face based biometric systems to PAs have been widely recognized, face PAD has received significant attention in the research community and remarkable progress has been made. Still, it is hard to tell what are the best or most promising practices for face PAD, because extensive objective evaluation and comparison of different approaches is challenging. While it is relatively cheap for an attacker to exploit a known vulnerability of a face authentication system (a ”golden fake”), such as a realistic 3D mask, manufacturing a huge amount of face artefacts and then simulating various types of attack scenarios (e.g. use-cases) for many subjects is extremely time-consuming and expensive. This is true especially in the case of hardware-based approaches because capturing new sensor-specific data is always required. Consequently, hardware-based techniques have been usually evaluated just to demonstrate a proof of concept, which makes direct comparison between different systems impossible.

Software-based countermeasures, on the other hand, can be assessed on common protocol benchmark datasets or, even better, if any new data is collected, it can be distributed to the research community. The early works in the field of software-based face PAD were utilizing mainly small proprietary databases for evaluating the proposed approaches but nowadays there exist several common public benchmark datasets, such as [9, 12, 16, 46, 49, 54]. The public databases have been indispensable tools for the researchers for developing and assessing the proposed approaches, which has had a huge impact on the amount of papers on data-driven countermeasures during the recent years. However, even if standard benchmarks are used, objective evaluation between different methods is not straightforward. First, the used benchmark datasets may vary across different works. Second, not all the datasets have unambiguously defined evaluation protocols, for example for training and tuning the methods, that provide the possibility for fair and unbiased comparison between different works.

Competitions play a key role in advancing the research on face PAD. It is important to organize collective evaluations regularly in order to assess, or ascertain, the current state of the art and gain insight on the robustness of different approaches using a common platform. Also, new more challenging public datasets are often collected and introduced within such collective efforts to the research community for future development and benchmarking use. The quality of PAIs keeps improving as technology (i.e., printers and displays) gets cheaper and better, which is another reason why benchmark datasets need to be updated regularly. Open contests are likely to inspire researchers and engineers beyond the field to participate, and their outside the box thinking may lead to new ideas on the problem of face PAD and novel countermeasures.

In the context of software-based face PAD, three international competitions [4, 10, 15] have been organized in conjunction with major biometric conferences in 2011, 2013 and 2017, each introducing new challenges to the research community. The first competition on countermeasures to 2D face spoofing attacks [10] provided an initial assessment of face PAD by introducing a precisely defined evaluation protocol and evaluating the performance of the proposed face PAD systems under print attacks. The second competition on countermeasures to 2D face spoofing attacks [15] utilized the same evaluation protocol but assessed the effectiveness of the submitted systems in detecting a variety of attacks, introducing display attacks (digital photos and video-replays) in addition to print attacks. While the first two contests considered only known operating conditions, the latest international competition on face PAD [4] aimed to compare the generalization capabilities of the proposed algorithms under some real-world variations faced in mobile scenarios, including unseen acquisition conditions, PAIs and input sensors.

This chapter introduces the state of the art in face PAD with particular focus on the three international competitions. The remainder of the chapter is organised as follows. First, we will give a brief overview on face PAD approaches proposed in the literature in Section 2. In Section 3, we will recapitulate the first two international competitions on face PAD, while Section 4 provides more comprehensive analysis on the latest competition focusing on generalized face PAD in mobile scenarios. In Section 5, we will discuss the lessons learnt from the competitions and future challenges in the field of face PAD in general. Finally, Section 6 summarizes the chapter, and presents conclusions drawn from the competitions discussed here.

2 Literature review on face PAD methods

There exists no universally accepted taxonomy for the different face PAD approaches. In this chapter, we categorize the methods into two very broad groups: hardware-based and software-based methods.

Hardware-based methods are probably the most robust ones for PAD because the dedicated sensors are able to directly capture or emphasize specific intrinsic differences between genuine and artificial faces in 3D structure [17, 42] and (multi-

spectral) reflectance [40, 42, 44, 55] properties. For instance, planar PAI detection becomes rather trivial if depth information is available [17], whereas near-infrared (NIR) or thermal cameras are efficient in display attack detection as most of the displays in consumer electronics emit only visible light. On the other hand, these kinds of unconventional sensors are usually expensive and not compact, thus not (yet) available in personal devices, which prevents their wide deployment.

It is rather appealing to perform face PAD by further analyzing only the same data that is used for face recognition or additional data captured with the standard acquisition device (e.g., challenge-response approach). These kinds of software-based methods can be broadly divided into active (requiring user collaboration) and passive approaches. Additional user interaction can be very effectively used for face PAD because we humans tend to be interactive, whereas a photo or video-replay attack cannot respond to randomly specified action requirements. Furthermore, it is very difficult to perform liveness detection or facial 3D structure estimation by relying only on spontaneous facial motion. Challenge-response based methods aim at performing face PAD detection based on whether the required action (challenge), for example facial expression [25, 36], mouth movement [11, 25] or head rotation (3D structure) [20, 34, 48], was observed within a predefined time window (response). Also, active software-based methods are able to generalize well across different acquisition conditions and attack scenarios but at the cost of usability due to increased authentication time and system complexity.

Passive software-based methods are preferable for face PAD because they are faster and less intrusive than active countermeasures. Due to the increasing number of public benchmark databases, numerous passive software-based approaches have been proposed for face PAD. In general, passive methods are based on analyzing different facial properties, such as frequency content [28, 46], texture [2, 12, 17, 27, 32, 53] and quality [18, 21, 23], or motion cues, such as eye blinking [3, 38, 45, 47], facial expression changes [3, 25, 45, 47], mouth movements [3, 25, 45, 47], or even color variation due to blood circulation (pulse) [15, 29, 31], to discriminate face artifacts from genuine ones. Passive software-based methods have shown impressive results on the publicly available datasets but the preliminary cross-database tests, such as [19, 48], revealed that the performance is likely to degrade drastically when operating in unknown conditions.

Recently, the research focus on software-based face PAD has been gradually moving towards assessing and improving the generalization capabilities of the proposed and existing methods in a cross-database setup instead of operating solely on single databases. Among hand-crafted feature based approaches, colour texture analysis [5, 6, 7, 8], image distortion analysis [21, 23, 49], combination of texture and image quality analysis with interpupillary distance (IPD) based reject option [39], dynamic spectral domain analysis [41] and pulse detection [29] have been applied in the context of generalized face PAD but with only moderate success.

The initial studies using deep CNNs have resulted in excellent intra-test performance but the cross-database results have still been unsatisfactory [39, 52]. This is probably due to the fact that the current publicly available datasets may not provide enough data for training well-known deep neural network architectures from scratch

or even for fine-tuning pre-trained networks. As a result, the CNN models have been suffering from overfitting to specific data and learning database-specific information instead of generalized PAD related representations. In order to improve the generalization of CNNs with limited data, more compact feature representations or novel methods for cross-domain adaptation are needed. In [33], deep dictionary learning based formulation was proposed to mitigate the requirement of large amounts of training data with very promising intra-test results but the generalization capability was again unsatisfying. In any case, the potential of application-specific learning needs to be further explored when more comprehensive face PAD databases are available.

3 First and second competitions on countermeasures to 2D face spoofing attacks

In this section, we recapitulate the first [10] and second [15] competitions on countermeasures to 2D face spoofing attacks, which were held in conjunction with International Joint Conference on Biometrics (IJCB) in 2011 and International Conference on Biometrics (ICB) in 2013, respectively. Both competitions focused on assessing the stand-alone PAD performance of the proposed algorithms in restricted acquisition conditions, thus integration with actual face verification stage was not considered.

In 2011, the research on software-based face PAD was still in its infancy mainly due to lack of public datasets. Since there were no comparative studies on the effectiveness of different PAD methods under the same data and protocols, the goal of the first competition on countermeasures to 2D facial spoofing attacks [10] was to provide a common platform to compare software-based face PAD using a standardized testing protocol. The performance of different algorithms was evaluated under print attacks using a unique evaluation method. The used PRINT-ATTACK database [1] defines a precise protocol for fair and unbiased algorithm evaluation as it provides a fixed development set to calibrate the countermeasures, while the actual test data is used solely for reporting the final results.

While the first competition [10] provided an initial assessment of face PAD, the 2013 edition of the competition on countermeasures to 2D face spoofing attacks [15] aimed at consolidating the recent advances and trends in the state of the art by evaluating the effectiveness of the proposed algorithms in detecting a variety of attacks. The contest was carried out using the same protocol on the newly collected video REPLAY-ATTACK database [12], introducing display attacks (digital photos and video-replays) in addition to print attacks.

Both competitions were open to all academic and industrial institutions. A noticeable increase in the number of participants between the two competitions can be seen. Particularly, six different competitors from universities participated in the first contest, while eight different teams participated in the second competition. The

Table 1 Names and affiliations of the participating systems in the first competition on countermeasures to 2D facial spoofing attacks

Algorithm name	Affiliations
AMILAB	Ambient Intelligence Laboratory, Italy
CASIA	Chinese Academy of Sciences, China
IDIAP	Idiap Research Institute, Switzerland
SIANI	Universidad de Las Palmas de Gran Canaria, Spain
UNICAMP	University of Campinas, Brazil
UOULU	University of Oulu, Finland

Table 2 Names and affiliations of the participating systems in the second competition on countermeasures to 2D face spoofing attacks

Algorithm name	Affiliations
CASIA	Chinese Academy of Sciences, China
IGD	Fraunhofer Institute for Computer Graphics, Germany
MaskDown	Idiap Research Institute, Switzerland University of Oulu, Finland University of Campinas, Brazil
LNMIIT	LNM Institute of Information Technology, India
MUVIS	Tampere University of Technology, Finland
PRA Lab	University of Cagliari, Italy
ATVS	Universidad Autonoma de Madrid, Spain
UNICAMP	University of Campinas, Brazil

affiliation and corresponding algorithm name of the participating teams for the two competitions are summarized in Table 1 and Table 2.

In the following, we summarize the design and main results of the first and second competitions on countermeasures to 2D face spoofing attacks. The reader can refer to [10] and [15] for more detailed information on the competitions.

3.1 Datasets

The first face PAD competition [10] utilized PRINT-ATTACK [1] database consisting of 50 different subjects. The real access and attack videos were captured with a 320×240 pixels (QVGA) resolution camera of a MacBook laptop. The database includes 200 videos of real accesses and 200 videos of print attack attempts. The PAs were launched by presenting hard copies of high resolution photographs printed on A4 papers with a Triumph-Adler DCC 2520 color laser printer. The videos were recorded under controlled (uniform background) and adverse (non-uniform background with day-light illumination) conditions.

The second competition on face PAD [15] was conducted using an extension of the PRINT-ATTACK database, named as REPLAY-ATTACK database [12]. The database consists of video recordings of real accesses and attack attempts corresponding to 50 clients. The videos were acquired using the built-in camera of a

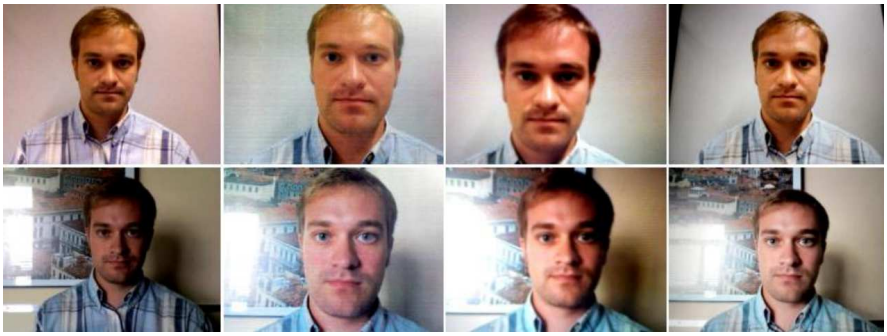


Fig. 1 Sample images from the PRINT-ATTACK [1] and REPLAY-ATTACK [12] databases. Top and bottom rows correspond to controlled and adverse conditions, respectively. From left to right columns: real accesses, print, mobile phone and tablet attacks.

MacBook Air 13 inch laptop under controlled and adverse conditions. Under the same conditions, high resolution pictures and videos were taken for each person using a Canon PowerShot SX150 IS camera and an iPhone 3GS camera, later to be used for generating the attacks. Three different attacks were considered: i) *print attacks* (i.e., high resolution pictures were printed on A4 paper and displayed to the camera); ii) *mobile attacks* (i.e., attacks were performed by displaying pictures and videos on the iPhone 3GS screen); iii) *high definition attacks* (i.e., the pictures and the videos were displayed on an iPad screen with 1024×768 pixels resolution). Moreover, attacks were launched with hand-held and fixed support modes for each PAI. Figure 1 shows sample images of real and fake faces from both PRINT-ATTACK and REPLAY-ATTACK databases.

3.2 Performance evaluation protocol and metrics

The databases used in both competition editions are divided into train, development and test sets with no overlap between them (in terms of subjects or samples). During the system development phase of the first competition, the participants were given access to the labelled videos of the training and the development sets that were used to train and calibrate the devised face PAD methods. In the evaluation phase, the performances of the developed systems were reported on anonymized and unlabelled test video files. In the course of the second competition, the participants had access to all subsets because the competition was conducted on the publicly available REPLAY-ATTACK database. The final test data consisted of anonymized videos of 100 successive frames cut from the original test set videos starting from a random time.

The first and second competitions considered a face PAD method to be prone to two types of errors: either a real access attempt is rejected (false rejection) or a PA is accepted (false acceptance). Both competitions employed Half Total Error

Table 3 Overview and performance (in %) of the algorithms proposed in the first face PAD competition (F stands for feature-level and S for score-level fusion)

Team	Features	Fusion	Development			Test		
			FAR	FRR	HTER	FAR	FRR	HTER
AMILAB	Texture, motion & liveness	S	0.00	0.00	0.00	0.00	1.25	0.63
CASIA	Texture, motion	S	1.67	1.67	1.67	0.00	0.00	0.00
IDIAP	Texture	-	0.00	0.00	0.00	0.00	0.00	0.00
SIANI	Motion	-	1.67	1.67	1.67	0.00	21.25	10.63
UNICAMP	Texture, motion & liveness	F	1.67	1.67	1.67	1.25	0.00	0.63
UOULU	Texture	-	0.00	0.00	0.00	0.00	0.00	0.00

Rate (HTER) as principal performance measure metric, which is the average of false rejection rate (FRR) and false acceptance rate (FAR) at a given threshold τ:

$$HTER(\tau) = \frac{FAR(\tau) + FRR(\tau)}{2} \tag{1}$$

For evaluating the proposed approaches, the participants were asked to provide two files containing a score value for each video in the development and test sets, respectively. The HTER is measured on the test set using the threshold τ corresponding to the equal error rate (EER) operating point on the development set.

3.3 Results and discussion

The algorithms proposed in the first competition on face PAD and the corresponding performances are summarized in Table 3. The participated teams used either single or multiple types of visual cues among motion, texture and liveness. Almost every system managed to obtain nearly perfect performance on both development and test sets of the PRINT-ATTACK database. The methods using facial texture analysis dominated because the photo attacks in the competition dataset suffered from obvious print quality defects. Particularly, two teams, IDIAP and UOULU, achieved zero percent error rates on both development and test sets relying solely on local binary pattern (LBP) [37] based texture analysis, while CASIA achieved perfect classification rates on the test set using combination of texture and motion analysis. Assuming that the attack videos usually are noisier than those of real videos, the texture analysis component in CASIA's system is based on estimating the difference in noise variance between the real and attack videos using first order Haar wavelet decomposition. Since the print attacks are launched with fixed and hand-held printouts with incorporated background (see Figure 1), the motion analysis component measures the amount of non-rigid facial motion and face-background motion correlation.

Table 4 gives an overview of the algorithms proposed within the second competition on face PAD and the corresponding performance figures for both development

Table 4 Overview and performance (in %) of the algorithms proposed in the second face PAD competition (F stands for feature-level and S for score-level fusion)

Team	Features	Fusion	Development			Test		
			FAR	FRR	HTER	FAR	FRR	HTER
CASIA	Texture & motion	F	0.00	0.00	0.00	0.00	0.00	0.00
IGD	Liveness	-	5.00	8.33	6.67	17.00	1.25	9.13
MaskDown	Texture & motion	S	1.00	0.00	0.50	0.00	5.00	2.50
LNMIIT	Texture & motion	F	0.00	0.00	0.00	0.00	0.00	0.00
MUVIS	Texture	F	0.00	0.00	0.00	0.00	2.50	1.25
PRA Lab	Texture	S	0.00	0.00	0.00	0.00	2.50	1.25
ATVS	Texture	-	1.67	0.00	0.83	2.75	21.25	12.00
Unicamp	Texture	-	13.00	6.67	9.83	12.50	18.75	15.62

and test sets. The participating teams developed face PAD methods based on texture, frequency, image quality, motion and liveness (pulse) features. Again, the use of texture was popular as seven out of eight teams adopted some sort of texture analysis in the proposed systems. More importantly, since the attack scenarios in the second competition were more diverse and challenging, a common approach was combining several complementary concepts together (i.e., information fusion at feature or score level). The category of the used features did not influence the choice of fusion strategy. The best-performing systems were based on feature-level fusion but it is more likely that the high level of robustness is largely based on the feature design rather than the used fusion approach.

From Table 4, it can be seen that the two PAD techniques proposed by CASIA and LNMIIT achieved perfect discrimination between the real accesses and the spoofing attacks (i.e., 0.00% error rates on the development and test sets). Both of these top-performing algorithms employ a hybrid scheme combining the features of both texture and motion-based methods. Specifically, the used facial texture descriptions are based on LBP, while motion analysis components again measure the amount of non-rigid facial motion and face-background motion consistency as the new display attacks are inherently similar to the ”scenic” print attacks of the previous competition (see Figure 1). The results on the competition dataset suggested that face PAD methods relying on a single cue are not able to detect all types of attacks, and the generalizing capability of the hybrid approaches is higher but with high computational cost. On the other hand, MUVIS and PRA Lab managed to achieve excellent performance on the development and test sets using solely texture analysis. However, it is worth pointing out that both systems compute the texture features over whole video frame (i.e., including background region), thus the methods are severely overfitting to the scene context information that matches across the train, development and test data. All in all, the astonishing results also on the REPLAY-ATTACK dataset conclude that more challenging configurations are needed before the research on face PAD can reach the next level.

4 Competition on generalized face presentation attack detection in mobile scenarios

The vulnerabilities of face based biometric systems to PAs have been widely recognized but still we lack generalized software-based PAD methods performing robustly in practical (mobile) authentication scenarios. In recent years, many face PAD methods have been proposed and remarkable results have been reported on the existing benchmark datasets. For instance, as seen in Section 3, several methods achieved perfect error rates in the first [10] and second [15] face PAD competitions. More recent studies, such as [5, 8, 19, 49, 52], have revealed that the existing methods are not able to generalize well in more realistic scenarios, thus software-based face PAD is still an unsolved problem in unconstrained operating conditions.

Focused large scale evaluations on the generalization of face PAD had not been conducted or organized after the issue was first pointed out by de Freitas Pereira *et al.* [19] in 2013. To address this issue, we organized a competition on mobile face PAD [4] in conjunction with IJCB 2017 to assess the generalization abilities of state-of-the-art algorithms under some real-world variations, including unseen input sensors, PAIs, and illumination conditions. In the following, we will introduce the design and results of this competition in detail.

4.1 Participants

The competition was open to all academic and industrial institutions. The participants were required register for the competition and sign the end user license agreement (EULA) of the used OULU-NPU database [9] before obtaining the data for developing the PAD algorithms. Over 50 organizations registered for the competition and 13 teams submitted their systems in the end for evaluation. The affiliation and corresponding algorithm name of the participating teams are summarized in Table 5. Compared with the previous competitions, the number of participants increased significantly from six and eight in the first and second competitions, respectively. Moreover, in the previous competitions, all the participated teams were from academic institutes and universities, whereas in this competition, we had registered the participation of three companies as well, which highlights the importance of the topic for both academia and industry.

4.2 Dataset

The competition was carried out on the recently published[1] OULU-NPU face presentation attack database [9]. The dataset and evaluation protocols were designed

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

Table 5 Names and affiliations of the participating systems

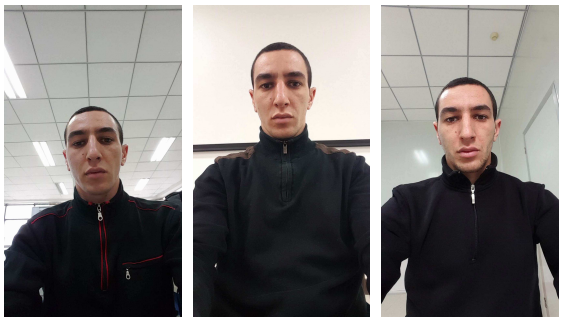
Algorithm name	Affiliations
Baseline	University of Oulu, Finland
MBLPQ	University of Ouargla, Algeria
PML	University of Biskra, Algeria University of the Basque Country, Spain University of Valenciennes, France
Massy_HNU	Changsha University of Science and Technology Hunan University, China
MFT-FAS	Indian 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

particularly for evaluating the generalization of face PAD methods in more realistic mobile authentication scenarios by considering three covariates: unknown environmental conditions (namely illumination and background scene), PAIs and acquisition devices, separately and at once.

The OULU-NPU database consists of 4950 short video sequences of real access and attack attempts corresponding to 55 subjects (15 female and 40 male). The real access attempts were recorded in three different sessions separated by a time interval of one week. During each session, a different illumination condition and background scene were considered (see Figure 2):

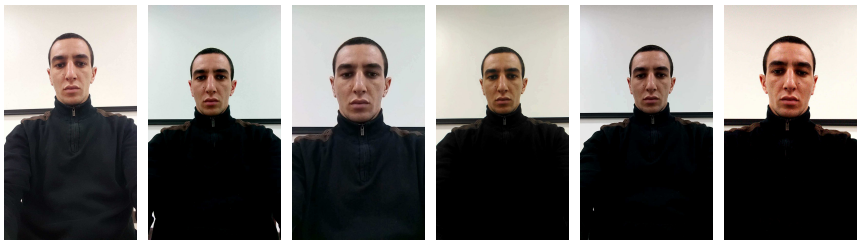
- *Session 1*: The recordings were taken in an open-plan office where the electric light was switched on, the windows blinds were open, and the windows were located behind the subjects.
- *Session 2*: The recordings were taken in a meeting room where the electric light was the only source of illumination.
- *Session 3*: The recordings were taken in a small office where the electronic light was switched on, the windows blinds were open, and the windows were located in front of the subjects.

During each session, the subjects recorded the videos of themselves using the front facing cameras of the mobile devices. In order to simulate realistic mobile authentication scenarios, the video length was limited to five seconds. Furthermore, the subjects were asked to use the device naturally while ensuring that the whole face is visible through the whole video sequence.



(a) Session 1 (b) Session 2 (c) Session 3

Fig. 2 Sample images of a real subject highlighting the illumination conditions across the three different scenarios.



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

Fig. 3 Sample images showing the image quality of the different camera devices.

Six smartphones with high-quality front-facing cameras in the price range from €250 to €600 were used for the data collection:

- Samsung Galaxy S6 edge with 5 MP frontal camera (Phone 1).
- HTC Desire EYE with 13 MP frontal camera (Phone 2).
- MEIZU X5 with 5 MP frontal camera (Phone 3).
- ASUS Zenfone Selfie with 13 MP frontal camera (Phone 4).
- Sony XPERIA C5 Ultra Dual with 13 MP frontal camera (Phone 5).
- OPPO N3 with 16 MP rotating camera (Phone 6).

The videos were recorded at Full HD resolution (i.e., 1920×1080) using the same camera software[2] installed on each device. Even though the nominal camera resolution of some mobile devices is the same, such as Phone 2, Phone 4 and Phone 5 (13 MP), significant differences can be observed in the quality of the resulting videos as demonstrated in Figure 3.

During each of the three sessions, a high-resolution photo and a video of each user was captured using the back camera of the Phone 1 capable of taking 16 MP still images and Full HD videos. These high resolution photos and videos were then

[2] http://opencamera.sourceforge.net/

(a) Print 1 (b) Print 2 (c) Display 1 (d) Display 2

Fig. 4 Samples of print and display attacks taken with the front camera of Sony XPERIA C5 Ultra Dual.

used to create the PAs. The attack types considered in this database are print and video-replay attacks:

- *Print attacks*: The high resolution photos were printed on A3 glossy paper using two different printers: a Canon imagePRESS C6011 (Printer 1) and a Canon PIXMA iX6550 (Printer 2).
- *Video-replay attacks*: The high-resolution videos were replayed on two different display devices: a 19" Dell UltraSharp 1905FP display with 1280×1024 resolution (Display 1) and an early 2015 Macbook 13" laptop with Retina display of 2560×1600 resolution (Display 2).

The print and video-replay attacks were then recorded using the front-facing cameras of the six mobile phones. While capturing the print attacks, the facial prints were held by the operator and captured with stationary capturing devices in order to maximize the image quality but still introduce some noticeable motion in the print attacks. In contrast, when recording the video-replay attacks both of the capturing devices and PAIs were stationary. Furthermore, we paid special attention that the background scene of the attacks matched that of the real accesses during each session and that the attack videos did not include the bezels of the screens or borders of the prints. Figure 4 shows samples of the attacks captured using the Phone 5.

4.3 Performance evaluation protocol and metrics

During the system development phase of two months, the participants were given access to the labelled videos of the training and the development sets that were used to train and 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

Table 6 The detailed information about the video recordings in the train, development and test sets of each protocol (P stands for print and D for display attack)

Protocol	Subset	Session	Phones	Subjects	Attacks	Real / attack videos
Protocol I	Train	1,2	6	1-20	P 1,2; D 1,2	240 / 960
	Dev	1,2	6	21-35	P 1,2; D 1,2	180 / 720
	Test	3	6	36-55	P 1,2; D 1,2	120 / 480
Protocol II	Train	1,2,3	6	1-20	P 1; D 1	360 / 720
	Dev	1,2,3	6	21-35	P 1; D 1	270 / 540
	Test	1,2,3	6	36-55	P 2; D 2	360 / 720
Protocol III	Train	1,2,3	5	1-20	P 1,2; D 1,2	300 / 1200
	Dev	1,2,3	5	21-35	P 1,2; D 1,2	225 / 900
	Test	1,2,3	1	36-55	P 1,2; D 1,2	60 / 240
Protocol IV	Train	1,2	5	1-20	P 1; D 1	200 / 400
	Dev	1,2	5	21-35	P 1; D 1	150 / 300
	Test	3	1	36-55	P 2; D 2	20 / 40

reported on anonymized and unlabelled test video files. To assess the generalization of the developed face PAD methods, four protocols have been used:

Protocol I: This 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: This 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 sensors.

Table 6 gives 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 [24], Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER):

$$APCER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - Res_i) \tag{2}$$

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}} \tag{3}$$

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:

$$ACER = \frac{\max_{PAI=1\ldots S}(APCER_{PAI}) + BPCER}{2} \tag{4}$$

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}.

4.4 Baseline

In addition to the training and development data, the participants were given the source code[3] of the baseline face PAD method that could be freely improved or used as it is in the final systems. The colour texture based method [5] was as the baseline because it has shown promising generalization abilities. In this method, the texture features are extracted from the colour images instead of the gray-scale representation that has been more commonly used in face PAD, for example in [17, 27, 32, 53]. The key idea behind colour texture based face PAD is that an image of an artificial face is actually an image of a face which passes through two different camera systems and a printing system or a display device, thus it can be referred to in fact as a recaptured image. As a consequence, the observed artificial face image is likely to suffer from different kinds of quality issues, such as printing defects, video artifacts, PAI dependent (local) colour variations and limited colour reproduction (gamut), that can be captured by analyzing the texture content of both luminance and chrominance channels.

The steps of the baseline method are the following. First, the face is detected, cropped and normalized into 64×64 pixels. Then, the RGB face image is converted

[3] The source code for baseline can be downloaded along with the OULU-NPU database.

Table 7 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

into HSV and YCbCr colour spaces. The local binary pattern (LBP) texture features [37] are extracted from each channel of the colour 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 ten random frames.

4.5 Results and discussion

In this competition, typical "liveness detection" was not adopted as none of the submitted systems is explicitly aiming at detecting physiological signs of life, such as 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 7): 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 8, 9, 10 and 11.

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 colour texture analysis [5] in which RGB images are converted into HSV and/or YCbCr colour spaces prior feature extraction. The only well-generalizing feature learning based method, MixedFASNet, uses HSV images as input, whereas the networks operating on gray-scale or RGB images do not generalize well. On the other hand, it is worth mentioning that VSS and SZCVI architectures consist only of five convolutional layers, whereas the MixedFASNet, consisting of over 30 layers, is much deeper. The best performing hybrid methods, Recod and CPqD, fuse 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.

In general, the submitted systems process each (selected) 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 exploited temporal variations but in the case of hand-crafted features two

Table 8 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 9 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

different temporal aggregation approaches were proposed for encoding the dynamic information within a video sequence, for example 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 [4]. The

Table 10 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 11 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

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 colour texture based face descriptions were very powerful compared to deep learning based methods, of which the impressive results

achieved 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 avoiding the need for enormous training sets. Unfortunately, the method did not generalize well on the evaluation set.

Few teams used additional public and/or proprietary datasets for training and tuning their algorithms. The VSS team augmented the subset of real subjects with CASIA-WebFace and collected their own attack samples. The usefulness of these external datasets remains unclear because their grayscale image analysis based face PAD method did not perform well. Recod used publicly available datasets for fine tuning the pre-trained network but the resulting generalization was comparable to similar method, CPqD, that did not use any extra-data. GRADIANT submitted two systems with and without external training data. Improved BPCER was obtained in unseen acquisition conditions but APCER was 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[4]. Thus, one-class modeling would be worth investigating when combined with more robust feature representations.

Several general observations can be made 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 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. The bona-fide samples were collected in three sessions with different illumination. In contrast, the bona-fide data corresponding 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 sensors is much larger in the case of print attacks compared to video-

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

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 a 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 challenging conditions of the competition and the OULU-NPU dataset.

5 Discussion

All the three competitions on face PAD were very successful in consolidating and benchmarking the current state of the art. In the following, we provide general observations and further discussion on the lessons learnt and potential future challenges.

5.1 General observations

It can be noticed that the used datasets and evaluation protocols, and also the recent advances in the state of the art reflect the face PAD scheme trends seen in the different contests. The algorithms proposed in the first and second competitions on countermeasures to 2D face spoofing attacks exploited the evident visual cues that we humans can observe in the videos of the PRINT-ATTACK and REPLAY-ATTACK databases, including localized facial movements, global motion, face-background motion correlation, print quality defects and other degradations in facial texture quality. While simple texture analysis was sufficient for capturing the evident printing artefacts in the PRINT-ATTACK database, fusion of multiple visual cues was needed for achieving robust performance under variety of attacks of the REPLAY-ATTACK database. The perfect error rates of the best-performing PAD schemes in homogeneous development and test conditions indicated that more challenging configurations were needed for future benchmarks.

In the competition on generalized face PAD, typical liveness detection and motion analysis were hardly used. In general, the proposed solutions relied on one or more types of feature representations extracted from the face and/or background regions using hand-crafted and/or learned descriptors, which is not surprising considering the recent trends in (face) PAD. Colour texture analysis had shown promising generalization capabilities in preliminary studies [5, 6, 7]. This explains why most teams proposed new facial colour texture representations or used the provided baseline as a complementary PAD method. Although it was nice to see a diverse set of deep learning based systems and further improved versions of the provided baseline

method, it was bit disappointing that entirely novel generalized face PAD solutions were not proposed. While the best-performing approaches were able to generalize remarkably well under the individual unknown conditions, no major breakthrough in generalized face PAD was achieved as the none of the methods was able to achieve satisfying performance under the most challenging test protocol, Protocol IV.

5.2 Lessons learnt

The competitions have given valuable lessons on designing databases and test protocols, and competitions in general. In the second competition on countermeasures to 2D face spoofing attacks, two teams managed to achieve perfect discrimination on the REPLAY-ATTACK database, and consequently PRINT-ATTACK database, by computing texture features over the whole video frame. The two background conditions in the REPLAY-ATTACK dataset are the same across the training, development and test sets and the corresponding scene is incorporated in the attack presentations (see Figure 1). Thus, also the differences in background scene texture between the real access and attack videos match between the development and test data, while only the facial texture is unknown due to previously unseen subjects. It is also worth mentioning that the original video encoder of the REPLAY-ATTACK dataset was not used for creating the randomly sampled test videos. The resulting video encoding artefacts and noise patterns did not match between the development and test phases, which might explain the increase in FRR for the methods relying largely on static and dynamic texture analysis.

In the third competition, focusing on generalization in face PAD, the time between the release of test data and submission of results was two weeks. The labelled test set of OULU-NPU database was not yet publicly available during the competition. However, we humans are apt in differentiating attack videos from real ones and the test subset of the OULU-NPU database contains still only 1800 videos. Therefore, it was feasible to label the anonymized and unlabelled test data by hand for ”data peeking”, that is calibrating, or even training, the systems on the test data. This kind of cheating could be prevented by hiding some ”anchor” videos from the development set (with randomized file names) in the evaluation data and releasing the augmented test set once the development set scores have been submitted (fixed), as done in the BTAS 2016 Speaker Anti-spoofing Competition [26]. The scores of the anchor videos could be used for checking whether the scores for the development and test sets have been generated by the same system.

An even more serious concern with the third competition is that the data provided for system development contained all variations in attacks, input sensors and acquisition conditions that the generalization was tested for. While only a specific subset defined in the test protocols (see Table 6) was supposed to be used for training, no measures were taken to prevent cheating by training and calibrating a single system on all data (containing also the unknown scenarios) and using it for reporting the scores for the according development and test sets of the individual protocols. In this

case, only the test subjects would be unknown to the system. Since only the integrity of the participants was trusted, the overall conclusions should be handled with care. However, it is worth pointing out that none of the submitted algorithms managed to achieve satisfying PAD performance on the OULU-NPU dataset even though cheating was possible. Although promising generalization was achieved across the different protocols, the best results are far from perfect, unlike in the previous competitions.

The best solution to prevent ”data peeking” or cheating in general would be to keep the test data, including unknown scenarios, inaccessible during algorithm development phase and to conduct independent (third-party) evaluations, in which the organizers run the provided executables or source codes of the submitted systems on the competition data. The results of the iris liveness detection competitions (LivDet-Iris) [50, 51] have shown already that the resulting performances can be far from satisfactory even for the winning methods, thus probably reflecting better the true generalization capabilities. It is worth highlighting that any later comparison to this kind of competition results should be treated with caution because it is impossible to reproduce the ”blind” evaluation conditions any more and, consequently, to achieve a fair comparison.

Over 50 organizations registered for the competition on generalized face PAD but only 13 teams made a final submission. Among the remaining 37 registered, there were also many companies. In general, the industrial participants should be encouraged to make an ”anonymous” submission, for example if the results might be unsatisfactory or details of the used algorithm cannot be revealed, as the results can still provide extremely useful additional information on the performance of the state of the art. For instance, in the LivDet-Iris 2017 competition [50], the best-performing algorithm was submitted anonymously.

5.3 Future challenges

The test cases in the OULU-NPU database measuring the generalization across the different covariates are still very limited. The video sequences have been captured with six different mobile devices but the attacks consists of only two different print attacks and display attacks and the acquisition conditions are quite controlled and restricted to three indoor office locations. Also, the variability in user demographics could be increased. The results of the third competition suggest that among the three tested covariates previously unseen acquisition conditions cause the most significant degradation in performance due to increase in BPCER, whereas unknown attacks have huge impact in APCER, especially in the case of print attacks. This observation is consistent with cross-dataset experiments conducted in other studies (e.g., [8, 16, 49]). While there is still plenty of room for improvement in the results obtained on the OULU-NPU dataset, more comprehensive datasets for investigating face presentation attack detection ”in the wild” will be eventually needed.

In general, the evaluation of biometric systems under presentations attacks can be conducted either at algorithm or system level [22]. In the first case, the robustness of the PAD modules is evaluated independently of the performance of the rest of the system, for instance, the face recognition stage. System level evaluation considers the performance of the biometric system as a whole. The advantage of system based evaluations is that it provides better insight into the overall robustness of the whole system to spoofing attacks, and how a proposed PAD technique affects the overall system accuracy (in terms of FRR). All three competitions have considered only stand-alone face PAD. Therefore, a possible future study would be combining match scores with both PAD and quality measures to improve the resilience of face verification systems [13, 43]. So far, the competitions have assessed the proposed PAD algorithms based on single liveness score values that have been assigned to each video after processing all or some of its frames. It would be also useful to measure the complexity, speed and latency of the participating systems, for example by computing the error rates over time.

Due to the recent advances in technology and vulnerabilities to spoofing, manufacturers, such as Microsoft, Apple and Samsung, have introduced new sensors (e.g., active NIR and depth cameras) for face verification purposes on personal devices. The dedicated imaging solutions are better capable of capturing the intrinsic differences between bona-fide samples and face artefacts than conventional cameras (hardware-based PAD). Since the new sensors are emerging in consumer devices, algorithm-based evaluations on sensor-specific data would be valuable addition in upcoming competitions. Alternatively, system-based evaluations of complete biometric systems with novel sensors and PAD modules could be assessed on the spot, as conducted already in LivDet-Iris 2013 [51], for instance. Naturally, this kind of arrangement requires careful competition design and execution, let alone significant efforts compared to algorithm level evaluation.

6 Conclusions

Competitions play a vital role in consolidating the recent trends and assessing the state of the art in face PAD. This chapter introduced the design and results of the three international competitions on software-based face PAD. These contests have been important milestones in advancing the research on face PAD to the next level as each competition has offered new challenges to the research community and resulted in novel countermeasures and new insight. The number of participants has grown in each successive competition, which indicates the increasing interest and importance of the research problem. The first and second competitions had six and eight participants from academic institutes, while the latest contest had 13 entries including three companies.

The first two competitions provided initial assessments of the state of the art by introducing a precisely defined test protocol and evaluating the performance of the systems under print and display attacks in homogeneous conditions. The best-

performing teams achieved perfect results in the first two competitions, because the test data did not introduce conditions (e.g., sensors, illumination or attacks) not seen during the algorithm development phase. Despite significant progress in the field, existing face PAD methods have shown lack of generalization in real-world operating conditions. Therefore, the latest contest considered a more unconstrained setup than in previous competitions, and aimed at measuring the generalization capabilities of the proposed algorithms under some real-world variations faced in mobile scenarios, including unknown acquisition conditions, PAIs and sensors. While the best results were promising, no major breakthrough in generalized face PAD was achieved even though the use of external training data was allowed.

Although none of the systems proposed in the latest competition managed to achieve satisfying PAD performance on the recent OULU-NPU database, more comprehensive datasets on presentation attack detection are still needed, especially considering the needs of data-hungry deep learning algorithms. So far, the competitions have focused only on stand-alone PAD, thus joint-operation with face verification would be worth investigating in future. Since new imaging solutions, such as NIR and depth cameras, are already emerging in consumer devices, it would be important to include these kinds of sensors in the upcoming benchmark datasets and competitions.

Acknowledgements The financial support from the Finnish Foundation for Technology Promotion and Infotech Oulu Doctoral Program is acknowledged.

Index

References

1. Anjos, A., Marcel, S.: Counter-measures to photo attacks in face recognition: a public database and a baseline. In: Proceedings of IAPR IEEE International Joint Conference on Biometrics (IJCB) (2011)
2. Bai, J., Ng, T.T., Gao, X., Shi, Y.Q.: Is physics-based liveness detection truly possible with a single image? In: IEEE International Symposium on Circuits and Systems (ISCAS), pp. 3425–3428 (2010)
3. Bharadwaj, S., Dhamecha, T.I., Vatsa, M., Richa, S.: Computationally efficient face spoofing detection with motion magnification. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2013)
4. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Qin, L., Peng, F., Zhang, L., Long, M., Bhilare, S., Kanhangad, V., Costa-Pazo, A., Vazquez-Fernandez, E., Perez-Cabo, D., Moreira-Perez, J.J., Gonzalez-Jimenez, D., Mohammadi, A., Bhattacharjee, S., Marcel, S., Volkova, S., Tang, Y., Abe, N., Li, L., Feng, X., Xia, Z., Jiang, X., Liu, S., Shao, R., Yuen, P.C., Almeida, W.R., Andalo, F., Padilha, R., Bertocco, G., Dias, W., Wainer, J., Torres, R., Rocha, A., Angeloni, M.A., Folego, G., Godoy, A., Hadid, A.: A competition on generalized software-based face presentation attack detection in mobile scenarios. In: IEEE International Joint Conference on Biometrics (IJCB) (2017)
5. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: IEEE International Conference on Image Processing (ICIP) (2015)
6. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters **24**(2), 141–145 (2016)
7. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security **11**(8), 1818–1830 (2016)
8. Boulkenafet, Z., Komulainen, J., Hadid, A.: On the generalization of color texture-based face anti-spoofing. Image and Vision Computing (2018)
9. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: IEEE International Conference on Automatic Face and Gesture Recognition (2017)
10. Chakka, M., Anjos, A., Marcel, S., Tronci, R., Muntoni, D., Fadda, G., Pili, M., Sirena, N., Murgia, G., Ristori, M., Roli, F., Yan, J., Yi, D., Lei, Z., Zhang, Z., Li, S., Schwartz, W., Rocha, A., Pedrini, H., Lorenzo-Navarro, J., Castrillon-Santana, M., Määttä, J., Hadid, A., Pietikäinen, M.: Competition on counter measures to 2-D facial spoofing attacks. In: International Joint Conference on Biometrics (IJCB) (2011)
11. Chetty, G., Wagner, M.: Liveness verification in audio-video speaker authentication. In: Australian International Conference on Speech Science and Technology, pp. 358–363 (2004)
12. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–7 (2012)
13. Chingovska, I., Anjos, A., Marcel, S.: Anti-spoofing in action: Joint operation with a verification system. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 98–104 (2013)
14. Chingovska, I., Erdogmus, N., Anjos, A., Marcel, S.: Face recognition systems under spoofing attacks. In: T. Bourlai (ed.) Face Recognition Across the Imaging Spectrum, pp. 165–194. Springer International Publishing (2016)
15. Chingovska, I., Yang, J., Lei, Z., Yi, D., Li, S.Z., Kähm, O., Glaser, C., Damer, N., Kuijper, A., Nouak, A., Komulainen, J., Pereira, T., Gupta, S., Khandelwal, S., Bansal, S., Rai, A., Krishna, T., Goyal, D., Waris, M.A., Zhang, H., Ahmad, I., Kiranyaz, S., Gabbouj, M., Tronci, R., Pili, M., Sirena, N., Roli, F., Galbally, J., Fierrez, J., Pinto, A., Pedrini, H., Schwartz, W.S., Rocha, A., Anjos, A., Marcel, S.: The 2nd competition on counter measures to 2D face spoofing attacks. In: International Conference on Biometrics (ICB) (2013)

16. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The REPLAY-MOBILE face presentation-attack database. In: International Conference on Biometrics Special Interests Group (BIOSIG) (2016)
17. Erdogmus, N., Marcel, S.: Spoofing attacks to 2D face recognition systems with 3D masks. In: IEEE International Conference of the Biometrics Special Interest Group (2013)
18. Feng, L., Po, L.M., Li, Y., Xu, X., Yuan, F., Cheung, T.C.H., Cheung, K.W.: Integration of image quality and motion cues for face anti-spoofing: A neural network approach. Journal of Visual Communication and Image Representation **38**, 451–460 (2016)
19. de Freitas Pereira, T., Anjos, A., De Martino, J., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: International Conference on Biometrics (ICB) (2013)
20. Frischholz, R.W., Werner, A.: Avoiding replay-attacks in a face recognition system using head-pose estimation. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (2003)
21. Galbally, J., Marcel, S.: Face anti-spoofing based on general image quality assessment. In: IAPR/IEEE International Conference on Pattern Recognition (ICPR), pp. 1173–1178 (2014)
22. Galbally, J., Marcel, S., Fiérrez, J.: Biometric antispoofing methods: A survey in face recognition. IEEE Access **2**, 1530–1552 (2014)
23. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. IEEE Transactions on Image Processing **23**(2), 710–724 (2014)
24. ISO/IEC JTC 1/SC 37 Biometrics: Information technology – Biometric presentation attack detection – Part 1: Framework. Tech. rep., International Organization for Standardization (2016)
25. Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-time face detection and motion analysis with application in liveness assessment. IEEE Transactions on Information Forensics and Security **2**(3), 548–558 (2007)
26. Korshunov, P., Marcel, S., Muckenhirn, H., Gonalves, A.R., Mello, A.G.S., Violato, R.P.V., Simoes, F.O., Neto, M.U., de Assis Angeloni, M., Stuchi, J.A., Dinkel, H., Chen, N., Qian, Y., Paul, D., Saha, G., Sahidullah, M.: Overview of BTAS 2016 speaker anti-spoofing competition. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS) (2016)
27. Kose, N., Dugelay, J.L.: Countermeasure for the protection of face recognition systems against mask attacks. In: International Conference on Automatic Face and Gesture Recognition (FG) (2013)
28. Li, J., Wang, Y., Tan, T., Jain, A.K.: Live face detection based on the analysis of fourier spectra. In: Biometric Technology for Human Identification, pp. 296–303 (2004)
29. Li, X., Komulainen, J., Zhao, G., Yuen, P.C., Pietikäinen, M.: Generalized face anti-spoofing by detecting pulse from face videos. In: International Conference on Pattern Recognition (ICPR) (2016)
30. Li, Y., Li, Y., Xu, K., Yan, Q., Deng, R.: Empirical study of face authentication systems under OSNFD attacks. IEEE Transactions on Dependable and Secure Computing (2016)
31. Liu, S., Yuen, P.C., Zhang, S., Zhao, G.: 3D mask face anti-spoofing with remote photoplethysmography. In: European Conference on Computer Vision (ECCV), pp. 85–100. Springer International Publishing (2016)
32. Määttä, J., Hadid, A., Pietikäinen, M.: Face Spoofing Detection From Single Images Using Micro-Texture Analysis. In: Proceedings of International Joint Conference on Biometrics (IJCB) (2011). DOI 10.1109/IJCB.2011.6117510
33. Manjani, I., Tariyal, S., Vatsa, M., Singh, R., Majumdar, A.: Detecting silicone mask based presentation attack via deep dictionary learning. IEEE Transactions on Information Forensics and Security (2017)
34. De Marsico, M., Nappi, M., Riccio, D., Dugelay, J.L.: Moving face spoofing detection via 3D projective invariants. In: IAPR International Conference on Biometrics (ICB) (2012)
35. Mohammadi, A., Bhattacharjee, S., Marcel, S.: Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. IET Biometrics **7**(1), 15–26 (2018)

36. Ng, E.S., Chia, A.Y.S.: Face verification using temporal affective cues. In: International Conference on Pattern Recognition (ICPR), pp. 1249–1252 (2012)
37. Ojala, T., Pietikäinen, M., Mäenpaa, T.: 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)
38. Pan, G., Wu, Z., Sun, L.: Liveness detection for face recognition. In: Recent Advances in Face Recognition, pp. 109–124. In-Teh (2008)
39. Patel, K., Han, H., Jain, A.K.: Cross-database face antispoofing with robust feature representation. In: Chinese Conference on Biometric Recognition (CCBR), pp. 611–619 (2016)
40. Pavlidis, I., Symosek, P.: The imaging issue in an automatic face/disguise detection system. In: IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (CVBVS), pp. 15–24 (2000)
41. Pinto, A., Pedrini, H., Robson Schwartz, W., Rocha, A.: Face spoofing detection through visual codebooks of spectral temporal cubes. IEEE Transactions on Image Processing **24**(12), 4726–4740 (2015)
42. Raghavendra, R., Raja, K.B., Busch, C.: Presentation attack detection for face recognition using light field camera. IEEE Transactions on Image Processing **24**(3), 1060–1075 (2015)
43. Rattani, A., Poh, N., Ross, A.: A bayesian approach for modeling sensor influence on quality, liveness and match score values in fingerprint verification. In: IEEE International Workshop on Information Forensics and Security (WIFS), pp. 37–42 (2013)
44. Rudd, E.M., Gnther, M., Boult, T.E.: PARAPH: Presentation attack rejection by analyzing polarization hypotheses. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 171–178 (2016)
45. Siddiqui, T., Bharadwaj, S., Dhamecha, T., Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Face anti-spoofing with multifeature videolet aggregation. In: International Conference on Pattern Recognition (ICPR) (2016)
46. Tan, X., Li, Y., Liu, J., Jiang, L.: Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: European Conference on Computer vision (ECCV), pp. 504–517 (2010)
47. Tirunagari, S., Poh, N., Windridge, D., Iorliam, A., Suki, N., Ho, A.T.S.: Detection of face spoofing using visual dynamics. IEEE Transactions on Information Forensics and Security **10**(4), 762–777 (2015)
48. Wang, T., Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection using 3D structure recovered from a single camera. In: International Conference on Biometrics (ICB) (2013)
49. Wen, D., Han, H., Jain, A.: Face spoof detection with image distortion analysis. Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
50. Yambay, D., Becker, B., Kohli, N., Yadav, D., Czajka, A., Bowyer, K.W., Schuckers, S., Singh, R., Vatsa, M., Noore, A., Gragnaniello, D., Sansone, C., Verdoliva, L., He, L., Ru, Y., Li, H., Liu, N., Sun, Z., Tan, T.: LivDet-Iris 2017 - Iris Liveness Detection Competition 2017. In: IEEE International Joint Conference on Biometrics (IJCB) (2017)
51. Yambay, D., Doyle, J.S., Bowyer, K.W., Czajka, A., Schuckers, S.: LivDet-iris 2013 - Iris Liveness Detection Competition 2013. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8 (2014)
52. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. CoRR **abs/1408.5601** (2014). URL `http://arxiv.org/abs/1408.5601`
53. Yang, J., Lei, Z., Liao, S., Li, S.Z.: Face liveness detection with component dependent descriptor. In: International Conference on Biometrics (ICB) (2013)
54. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: International Conference on Biometrics (ICB), pp. 26–31 (2012)
55. Zhang, Z., Yi, D., Lei, Z., Li, S.Z.: Face liveness detection by learning multispectral reflectance distributions. In: International Conference on Face and Gesture, pp. 436–441 (2011)