Generalized Presentation Attack Detection: a face anti-spoofing evaluation proposal

Artur Costa-Pazo[†§], David Jiménez-Cabello[†], Esteban Vazquez-Fernandez[†]
GRADIANT, Vigo (Spain)[†]
{acosta,djcabello,evazquez}@gradiant.org

José Luis Alba-Castro[§]
University of Vigo (Spain)[§]
jalba@gts.uvigo.es

Roberto J. López-Sastre[‡]
University of Alcalá (Spain)[‡]
robertoj.lopez@uah.es

Abstract

Over the past few years, Presentation Attack Detection (PAD) has become a fundamental part of facial recognition systems. Although much effort has been devoted to anti-spoofing research, generalization in real scenarios remains a challenge. In this paper we present a new open-source evaluation framework to study the generalization capacity of face PAD methods, coined here as face-GPAD. This framework facilitates the creation of new protocols focused on the generalization problem establishing fair procedures of evaluation and comparison between PAD solutions. We also introduce a large aggregated and categorized dataset to address the problem of incompatibility between publicly available datasets. Finally, we propose a benchmark adding two novel evaluation protocols: one for measuring the effect introduced by the variations in face resolution, and the second for evaluating the influence of adversarial operating conditions.

1. Introduction

Face recognition based systems have become very attractive solutions for a wide variety of applications. The presentation of the iPhone X in 2017 with its face verification system (the "FaceID") puts this technology in the spotlight as a strong candidate to substitute the fingerprint verification, not only to unlock the device but also as a mobile authentication mechanism.Furthermore, these systems are increasingly used in several applications as border controls, accesses to events or even for on-boarding processes. However, there remain two major challenges for the inclusion of these kind of systems in a larger number of applications: the Presentation Attacks (PAs) and their generalization capability.

The development of Presentation Attack Detection (PAD) solutions is a must, to guarantee the security of the users. Early adopters of face recognition (FR) systems have historically chosen *active face-PAD* approaches in order to preserve their security. This challenge-response strategy requires to mimic certain actions (e.g. eye blinking, smiling, head motion) increasing its robustness to attacks based on photographs or videos. However, latest advances in computer graphics [12, 15] threaten even this collaborative countermeasure. On the other hand, *passive face-PAD* is a more convenient and non-intrusive approach, but also entailing a far more challenging scenario. Relying on additional sensors such as 3D/IR/thermal cameras is an easy option, but it restricts the use case to a few specialized devices, dramatically incrementing costs. For the sake of accessibility and costs, we focus on the ubiquitous 2D-camera case, available in almost all mobile devices.

On the other hand, we find the generalization problem. After a first stage of enthusiasm among the research community, with PAD solutions that offer very low error rates in the available datasets, it comes a stage of disappointment when the same approaches are not able to work in more real environments. For instance, cross-evaluations between datasets quickly began to be included in the publications [9] in order to reveal a great problem: methods evaluated in a dataset other than training increase error rates by an order of magnitude. However, there is no common evaluation framework for conducting such cross-evaluation experiments in a consistent and comparable manner across publications. Therefore, addressing the generalization problem is critical.

We believe that generalization in face-PAD is in an early research stage and it is time to unify the criteria in order to perform fair evaluations. Thus, in this paper we propose to

978-1-7281-3640-0/19/$31.00

establish a new topic within the anti-spoofing research that deals with the generalization issues within face-PAD methods. We call this problem Generalized Presentation Attack Detection (GPAD).

We propose to analyze the status quo of the face anti-spoofing research with a proposal centered around the GPAD problem. The main contributions of this work include:

- We provide the largest aggregated dataset with a common categorization in two levels to represent four key aspects in anti-spoofing: attacks, lighting, capture devices and resolution.
- We release an open-source evaluation framework[1], introducing an unified benchmark for GPAD.
- We provide an evaluation of state-of-the-art methods in the proposed benchmark. We demonstrate the limitation of current dataset evaluation procedures (generalization, cross-domain evaluation, etc.), while showing the benefits of the proposed unified framework. All the experiments will be reproducible.
- Using the novel evaluation tool, we introduce two novel protocols for the GPAD problem.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the challenges in the field of generalization for anti-spoofing systems. The proposed framework and the description of the Aggregated-Datasets are presented in Section 3. In Section 4 we discuss in detail the proposed evaluation protocols and results. Conclusions are presented in Section 5.

2. Related Work

As [4] already points out, there is no unified taxonomy for the diverse set of attacks, nor for a unified criteria for the evaluation protocols or the metrics used to fairly assess the performance of face-PAD systems, specially when dealing with the generalization problem. We compliment the taxonomy initiated in [4] to fully categorize the current status of face-PAD approaches from different perspectives.

Current face-PAD methods can be classified regarding the following standpoints: i) from the hardware used for data acquisition as *rgb-only* [11, 21, 26] or *additional sensors* [2, 23] approaches; ii) from the required user interaction as *active* [14] or *passive* [16, 26] methods; iii) from the input data type as *single-frame* [26] or *video-based* [1, 22] approaches; iv) and, finally, depending on the feature extraction and classification strategy as *hand-crafted* [4, 26] or *deep learning* [13, 16]. Based on these classifications, we can depict that the most challenging scenario occurs when data is captured using rgb-only sensors using passive approaches that avoid any challenge-response interaction with the user (e.g. smiling or blinking) that increases drastically the usability of face-PAD systems. The most recent and successful methods [13, 17, 25] show that video-based solutions incorporate more meaningful information compared with those based on single frames, adopting the former as the main research direction. Finally, despite that many recent models attempt to detect face liveness, building upon representation based hand-crafted features [11, 19, 24], obtaining good results for intra-dataset protocols, the increasing interest on the topic has led to the appearance of new datasets [13, 17, 30], turning deep learning methods as a solid alternative adopted by almost every recent approach [13, 16, 17, 20, 30]. Considering the aforementioned challenging setting (rgb-only, passive, video-based and deep learning), these methods leverage recent advances on deep metric learning to extract highly discriminative features from image sequences, achieving state-of-the-art results for intra-dataset setups, especially when using auxiliary supervision based on depth reconstruction [13, 16]. Despite training data is not enough, this setting has been established as the right direction to discover the complex patterns and details that appear in any spoofing attempt.

2.1. Generalized Presentation Attack Detection

Current state-of-the-art solutions suffer a severe drop of performance during testing in realistic scenarios because they exhibit a sort of overfitting behavior maximizing the results for intra-dataset experimental setups. This fact has made that each approach proposes its own evaluation protocols and metrics, giving place to a missing unified criteria to evaluate generalization.

Generalization has been specifically addressed with no success from different perspectives: i) applying *domain adaptation* techniques [17]; ii) reformulating the problem as an *anomaly detection* [21] scenario; iii) learning *generalized deep features* [13, 16, 17]; or even iv) using generative models [13]. Besides, there is still a lack of unified benchmark and representative datasets, that might mitigate the constant improvement of Presentation Attack Instruments (PAIs) with more sophisticated strategies (e.g. 3D rendering) and newly unseen attack instruments (e.g. professional make-up). Regardless almost every proposal comes with its own reduced dataset [13, 17, 20, 30], there is *no agreement upon a PAD benchmark*, and generalization properties are not properly evaluated.

2.2. Datasets

During a brief inspection of the capture settings of video-based face-PAD datasets (see Table 1), one can easily observe that there is no unified criteria in the goals of each of them, leading to a manifest built-in bias. This specificity in the domain covered by most of the datasets can

[1] https://github.com/Gradiant/bob.paper.icb2019.gradgpad

be observed in different scenarios: i) some of them focus on a single type of attack (e.g masks - 3DMAD, HKBU, CSMAD), ii) others focus on the study of different image sources (depth/NIR/thermal) such as CASIA-SURF or CS-MAD, iii) others attempt to simulate a certain scenario like a mobile device setting, where the user hold the device (e.g. Replay-Mobile, OULU-NPU), or a webcam setting, where user is placed in front of fixed camera (e.g Replay-Attack, SiW), or even a stand-up scenario where users are recorded further from the camera (e.g UVAD), etc.

Beyond classical intra-dataset evaluation where state-of-the-art algorithms work really well, most of them also propose to assess generalization properties using a direct evaluation between different datasets (i.e. inter-dataset protocols). The most recent publications go a step further by proposing aggregated datasets such as [21] and [27]. The underlying idea is to generate a combined dataset that represents the anti-spoofing scenario in a richer way, increasing the diversity of the data, e.g. with a larger number of users, capture devices and PAIs.

It seems clear that the community has become aware of the importance of data representativeness in our topic and more and more publications are proposing new databases or different combinations of inter-dataset evaluations.

2.3. Challenges

Despite the effort of the research community, there is no standard methodology to evaluate the different approaches in a cross-dataset scenario, which motivates each publication to propose a new method with a different dataset. This makes it impossible to perform a more in-depth analysis of anti-spoofing algorithms, since by keeping the global results of a dataset, we are masking many variables. Thus, some questions quickly arise. Are our algorithms working badly for an specific type of attack, capture device or resolution?, or is it really just one of those hidden parameters that is causing that the performance drops drastically? It is obvious that there is a big challenge under the generalization problem and we need to unmask it in detail if we aim at giving a big step towards a general solution. We propose to emphasize the importance of data and to unify the procedure to evaluate face anti-spoofing systems in the most challenging scenario: the generalization problem.

3. The Evaluation Framework

In this paper, we present a novel evaluation framework to help the research community to address the problem of generalization in face-PAD. Our framework is publicly available as a Python package under the acronym GRAD-GPAD (Generalization Representation over Aggregated Datasets for Generalized Presentation Attack Detection). This framework provides a common taxonomy of existing datasets and allows to evaluate face-PAD algorithms

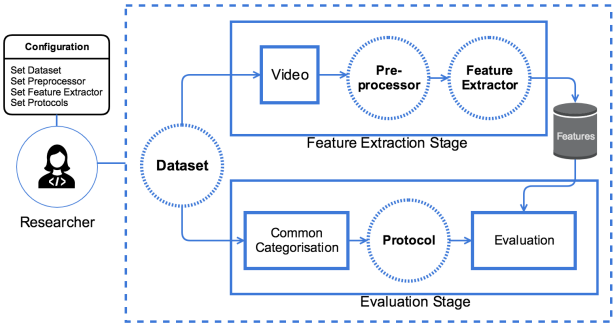

Figure 1: Overview of GRAD-GPAD evaluation framework

from additional points of view, revealing new interesting properties. The GRAD-GPAD framework is extendable and scalable by design. This will help researchers to create new protocols, to define categorizations over different datasets in order to make them compatible, and to enable the addition of new datasets.

The GRAD-GPAD framework has a top-down design and provides a very convenient level of abstraction. Figure 1 shows the proposed dataset evaluation pipeline, where circles with dotted lines represent the customizable and easy-to-extend elements. Each of these elements in the framework follows a specific interface to work properly. Thus, we can add, for instance, a new *Dataset* (e.g. using some videos recorded for an specific use case), or a new implementation of a *Feature Extractor* to evaluate an algorithm or even create an ad-hoc *Protocol*.

GRAD-GPAD presents two main stages: i) *feature extraction*, where features are computed for each input video applying a preprocessing step and following the interface of the Feature Extractor; and ii) *evaluation*, where a filtering step is applied, using the already extracted features and the common categorization of the datasets, to train and test over the selected features.

3.1. The Aggregated Dataset

We propose a novel paradigm of data collection where new datasets are aggregated in a larger dataset, following a global categorization to make the involved datasets compatible. We show that our framework can be used to replicate and extend the analysis of state-of-the-art protocols, as well as to obtain a fair analysis of the algorithms using novel protocols.

As we already pointed out in Section 2, current publicly available datasets for face anti-spoofing are very heterogeneous. The variability of the spoofing attacks and capture/display devices makes the problem actually unconstrained and dynamic. Moreover, existing datasets present different labeling, protocols and data splits. The GRAD-GPAD framework simplifies such a dynamic struc-

Dataset	Year	Num Identities	Number samples real/ attack	Spoof attacks	Capture Devices	Modal types	Display Devices	Pose Range	Different Expression	Additional Lightning
CASIA-FASD [31]	2012	50	150/450	Print, Replay	low-quality webcam, medium-quality webcam, Sony NEX-5	RGB	iPad	Frontal	No	No
REPLAY-ATTACK [6]	2012	50	200/1000	Print, 2 Replay	macbook webcam	RGB	iPhone 3GS, iPad	Frontal	No	Yes
3DMAD [10]	2013	17	170/85	Mask (rigid)	Microsoft Kinect	RGB/Depth	-	Frontal	No	No
MSU-MFSD [26]	2015	35	110/330	Print, 2 Replay	macbook air webcam, nexus 5	RGB	iPad Air, iPhone 5S	Frontal	No	No
UVAD [22]	2015	404	808/16268	7 Replay	cybershot_dsc-hx1, canon_powershot_sx1, nikon_coolpix_p100, kodak_z981, olympus_sp_800UZ, panasonic_fz35_digital	RGB	7 different unknown monitors	Frontal	No	No
REPLAY-MOBILE [7]	2016	40	390/640	Print, Replay	iPad Mini 2, LG G4	RGB	Philips 227ELH	Frontal	No	Yes
HKBU [18] (v1)	2016	8	70/40	Mask (rigid)	Logitech C9200 webcam	RGB	-	Frontal	No	No
OULU-NPU [5]	2017	55	1980/3960	2 Print, 2 Replay	Samsung Galaxy S6 Edge, htc_desire_eye, meizu_x5, asus_xenfone, sony_xperia_c5, oppo_n3	RGB	Dell1905FP, MacBook Retina	Frontal	No	Yes
SMAD [20]	2017	-	65/65	Mask (silicone)	-	RGB	-	-	-	-
ROSE-YOUTU [17]	2018	20	3350	2 Print, 2 Replay, 2 Mask (Paper)	hasee, huawei, ipad_4, iphone_5s, zte	RGB	Lenovo LCD display, Mac LCD display	Frontal	Yes	No
SIW [13]	2018	165	1320/330	2 Print, 4 Replay	canon_eos_t6, logitech_c920_webcam	RGB	iPad Pro, iPhone7 Galaxy S8, Asus MB168B	[-90^o, 90^o]	Yes	Yes
CS-MAD [2]	2018	14	88/220	Print,Mask (silicone)	Intel Realsense SR300 (full HD), Nikon Coolpix P520 (still images)	RGB/Depth IR/Thermal	-	Frontal	No	Yes
CASIA-SURF [30]	2019	1000	3500/17500	Print, Mask (paper, cut)	Intel RealSense	RGB/Depth IR	-	[-30^o, 30] (attacks)	No	No

Table 1: List of existing databases for anti-spoofing based on videos and their main characteristics.

ture thanks to its scalable nature and only requires the data to be split into three subsets in order to evaluate face-PAD algorithms: a training set to find the classification model, the development set to estimate the threshold that provides Equal Error Rate (EER) performance, and a test set that is used to report the final results using fair metrics for the generalization problem that has been recently standardized in the *ISO/IEC 30107-3* [2]: i.e. HTER (*Half Total Error Rate*), ACER (*Average Classification Error Rate*), APCER (*Attack Presentation Classification Error Rate*) and BPCER (*Bona fide Presentation Classification Error Rate*). Some datasets (e.g. Replay-Attack [6], 3DMAD [10], Replay-Mobile [7], HKBU [18], OULU-NPU [5]) are already divided that way, others, however, need to be restructured. Based on the work of Pinto et al. in [22], we have split CASIA-FASD [31], Rose-Youtu [17] and SiW [13], keeping the test subset unmodified and splitting the original training set in a training subset comprising 80% of the users and a development subset comprising the remaining 20%. Furthermore, CS-MAD [2] does not contain explicit subsets, so we randomly partitioned the data into the mentioned subsets (40% in Train, 30% in Dev and 30% in Test) from the users' identities. Finally MSU-MFSD [26] is originally divided in two folds, nevertheless we re-divided it based on this Python package [3]. At the moment of writing this paper, ten out of the thirteen datasets shown in Table 1 have been integrated in our framework. Work is currently underway to add both SMAD [10] and UVAD [22], as well as the recently introduced CASIA-SURF [29].

3.2. Categorization

Although some of the datasets share a common notation, there is no common taxonomy, so we can categorize the different types of accesses represented in current datasets. Each of these datasets has useful categories and labels that help researchers for a better understanding of their algorithms. However, the amount of data is usually not enough to train the algorithms properly, so they cannot be effectively used. In this paper, we propose an updated inter-dataset categorization based on four general categories: *common PAI*, *common capture device*, *common lighting* and *common face resolution*. Table 2 and the next paragraphs present, in more detail, each of the proposed categories, as well as their subclasses (type and subtype).

Category	Types	Sub-type	Criteria
Common PAI	print	low	dpi $\leq$ 600
		medium	600 $<$ dpi $\leq$ 1000
		high	dpi $>$ 1000
	replay	low	res $\leq$ 480 pix
		medium	480 $<$ res $<$ 1080 pix
		high	res $\geq$ 1080 pix
	mask	paper	paper masks
		rigid	non-flexible plaster-like
		silicone	silicone masks
Common Capture Devices	webcam	low	SD res
		high	HD res
	mobile/tablet	low	SD res
		high	HD res
	digital camera	low	SD res
		high	HD res
Common Face Resolution	small face	-	IOD $\leq$ 120 pix
	medium face	-	120 $<$ IOD $\leq$ 240 pix
	large face	-	IOD $>$ 240 pix
Common Lighting	controlled	-	-
	adverse	-	-
	no info	-	-

Table 2: Inter-Dataset common categorization.

1. *Common PAI*. So far, every existing dataset have in common that the PAI used in each attack is precisely labeled. In this work, we also propose an additional subcategorization in three subtypes based on both quality and material criteria. Print attacks are categorized as: 1) low quality, if the attacks were performed with an image printed on a printer with less than $600dpi$ (dots per inch), 2) high, if the printer exceeds $1000dpi$, and 3) medium for the remaining range. The resolution of the screen is used to sub-categorize the Replay attacks as follows: 1) high, if the attacks are performed replaying a video or a photo on a screen with a resolution higher than $1080pix$,

[2] https://www.iso.org/standard/67381.html

[3] https://gitlab.idiap.ch/bob/bob.db.msu_mfsd_mod

2) low, if it is performed on screens with lower resolution than $480pix$, and 3) medium, as in the previous case, for the remaining range. In the case of mask-type attacks, the categorization is more subjective, so it has been divided on the basis of the material used in each of the masks as: 1) paper, 2) non-flexible plaster-like and 3) silicone.

2. *Common capture devices*. Production ready systems that are currently working in the real world have been developed for specific hardware, being able to adjust the parameters to a specific scenario. However, all state-of-the-art systems suffer a dramatic performance loss when they are evaluated in inter-dataset protocols with no reasoning about the parameters involved. GRAD-GPAD allows, in much more detail, an analysis of the influence of these parameters, e.g. the capture devices. Three different types have been added: 1) webcams, 2) mobile/tablet and 3) digital cameras. Additionally, they can be divided by the resolution of their sensors as: high-definition (HD) or standard-definition (SD).
3. *Common lighting*. Features extracted from color and texture information (crucial in most of current face-PAD methods) are highly influenced by the varying lighting conditions. However, their influence in face-PAD generalization has not been studied properly. In the proposed categorization, we have collected all the information related to lighting. We have found 3 novel divisions: 1) controlled, 2) adverse and 3) no-info. Videos with no information about the lighting have been categorized as no info, waiting for a future categorization.
4. *Common Face Resolution*: Texture-based methods suffer from different face resolutions. Thus, we propose a inner division in three types: 1) small, 2) medium and 3) large faces. The applied criterion has been extracted from the *ISO/IEC 29794-5:2010* [4] recommendations, applying thresholds to the Interocular Distance (IOD). Eyes landmarks have been extracted with [28] for all the frames available on the aggregated dataset, and the categorization has been made for each video, based on an average of the inter eyes distances of the frames.

3.3. Protocols

The GRAD-GPAD framework allows, using configuration files, a simple filtering of data to create our own protocols. We can determine for each of the subsets of the aggregated dataset (Train, Dev and Test) which dataset and subsets are used in the experiment, and filter depending on the common categories selected.

Furthermore, the framework also offers some protocols that have been already proposed in the literature to benchmark the anti-spoofing algorithms. The default protocols are the following:

[4] https://www.iso.org/standard/50912.html

1. **Grandtest**: a protocol that evaluates face-PAD algorithms without any filter. All sets include all the previous categories.
2. **Cross-Dataset**: this is the most common protocol used to assess generalization of face-PAD algorithms. It is based on training on one or several datasets and testing in others.
3. **One-PAI**: a protocol that evaluates face-PAD algorithms filtering only by PAI. This protocol is useful to evaluate systems that are focused on the detection of only one kind of PAI.
4. **Unseen Attacks (Cross-PAI)**: a protocol that evaluates the performance under unseen PAI. In the training and dev stage a PAI is excluded, which is then used for testing.
5. **Unseen Capture Devices**: protocols that evaluate how a face-PAD algorithm works under capture devices that were excluded on training and dev stages.

Despite these protocols already appeared in the literature, the proposed categorization and the aggregated dataset allow a better representation of the anti-spoofing scenario (more examples of different devices, attacks, conditions, etc.). That way, these protocols can extract finer information from the methods under evaluation. Additionally, we propose in this work two novel protocols that evaluate parameters of great relevance in the generalization to real environments.

1. **Cross-FaceResolution**. Building on top of some papers [4] that mention the influence of the distance between camera and subject, we propose a protocol to focus the analysis on the role of resolution of the region of the face in the task of detecting fake attempts. Two variants are possible for this protocol. The first one, *Cross-FaceResolution-LF-Test*, uses small and medium faces (determined by the resolution) for training, while using high resolution faces (LF - Large Faces) for testing. The second, *Cross-FaceResolution-SF-Test*, represents the opposed setup: it uses large and medium faces samples on training and small faces (SF) for testing.
2. **Cross-Conditions**. The underlying idea is to assess the performance of the anti-spoofing algorithms under adversarial conditions. Two variants are proposed: *Cross-Conditions-Test-Adverse*, where we evaluate the performance of the system when training on optimal conditions (high quality capture devices, low and medium quality PAIs, paper masks and both controlled and no info lighting conditions) and testing on adverse conditions (low quality capture devices, high quality PAIs, silicon and non-flexible plaster-like mask, and adverse lighting conditions); and *Cross-Conditions-Test-Optimal*, where we do the opposite.

4. Experiments

In this section we present the results of the experiments for the proposed benchmark. The GRAD-GPAD framework

is used to extract the features, using two well-known face-PAD methods. We then train the models and filter the categories in order to evaluate several protocols for the proposed aggregated dataset.

4.1. Face-PAD methods

Two recent and popular approaches have been selected to guide the explanation of the proposed protocols, which will also serve as a baseline for the aggregated dataset: 1) the Color-based face-PAD proposed in [3], and 2) the Quality-Based method proposed in [21] based on a concatenation of quality measures (IQM) introduced in [11] and [26]. The code for these algorithms is publicly available in the GRAD-PAD framework based on the reproducible material[5, 6] shared by the authors.

Since we want to compare systems using the same conditions, we fix the video preprocessing and the classification stage for both algorithms. Following the recommendations given in [8], the experiments have been designed trying to simulate a real scenario and considering usability as a key aspect of the systems.

For face detection we use the method proposed in [28]. Faces are cropped and rescaled to 64 x 64 pixels. This cropped image is the input for the method under evaluation. In the case of the Quality-Based algorithm, we obtain for every image a 139-length feature vector from the concatenation of the quality measurements proposed in [11, 26]. On the other hand, the Color-Based face-PAD computes for each image a vector of a much larger dimensionality (19998-length feature vector) by concatenating texture features (LBP-based) extracted in two color spaces (YCrCb and HSV). Then, for each access video, a feature vector is obtained as the average of the extracted features in each frame. Finally, an SVM classifier with an RBF kernel ($\gamma = 1/numfeatures$) is trained for both systems.

4.2. Classical Evaluation

First of all, the face-PAD methods are evaluated with two protocols well known to the community (*Grandtest* and *Cross-Dataset*). The aim of the experiments is two-fold. First, to test the selected algorithms on the aggregated dataset and to provide a baseline for further comparisons. Second, to take advantage of the capabilities of the GRAD-GPAD framework to perform an in-depth analysis of the two approaches. Initially, we evaluate the algorithms with the *Grandtest* protocol.

Now, from the results in Table 3 we can easily observe that the Color-based approach performs better in general. Besides, the large difference, in both approaches, between HTER and ACER, suggests the influence of some PAIs.

[5] https://github.com/zboulkenafet

[6] https://gitlab.idiap.ch/bob/bob.pad.face/

	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
Quality-Based	17.03	25.25	34.09	16.41
Color-Based	6.33	10.22	13.86	6.58

Table 3: Results for *Grandtest* protocol

In order to understand the source of the error regarding the PAI, we have used the One-PAI protocol. As this protocol has only one type of attack for testing, the reported metric is HTER (equivalent to ACER for the single case). The Quality-based face-PAD obtains better results for replay (HTER = 10.22%) and mask attacks (HTER = 10.08%), compared with print attacks (HTER = 14.53%). The color-based approach, on the contrary, has a slightly more stable behavior with a better performance for mask attacks (HTER = 2.90%), while print and replay attacks get HTER values of 4.41% and 3.92% respectively.

Once we have clarified the strengths and weaknesses of the two approaches in the aggregated dataset, it is time to check their generalization capabilities. Table 4 represents the results for leave-one-out *Cross-Dataset* protocol, where the named dataset is used for testing the generalization of the models and the nine remaining datasets are used for training and tuning the algorithms.

Test	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
CASIA-FASD	41.57	48.98	81.11	16.85
REPLAY-ATTACK	27.61	34.06	33.96	34.17
3DMAD	29.00	29.00	0.00	58.00
MSU-MFSD	31.11	46.66	46.66	46.66
REPLAY-MOBILE	26.89	28.19	34.37	22.02
HKBU	45.00	45.00	90.0	0.00
OULU-NPU	34.68	41.11	75.27	6.94
ROSE-YOUTU	37.88	45.81	42.40	49.22
SIW	31.97	48.40	53.07	43.74
CSMAD	40.51	40.51	10.20	70.83

(a) Results using the Quality-Based face-PAD

Test	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
CASIA-FASD	15.45	16.75	17.78	15.73
REPLAY-ATTACK	25.11	33.35	31.25	35.44
3DMAD	0.00	0.00	0.00	0.00
MSU-MFSD	17.78	35.00	56.66	13.33
REPLAY-MOBILE	18.30	22.99	23.96	22.02
HKBU	0.00	0.00	0.00	0.00
OULU-NPU	34.27	37.78	72.22	3.33
ROSE-YOUTU	27.42	34.78	25.25	44.32
SIW	9.90	22.06	30.43	13.69
CSMAD	40.05	40.05	55.10	25.00

(b) Results using the Color-Based face-PAD

Table 4: Results for *Cross-Dataset* protocol

To extract realistic conclusions from the results of the cross-dataset protocol, we have to take into account the global parameters of each of the datasets (see Table 1). From the experiments we observe that the Color-based method generalizes consistently better in all the datasets considered in the cross-dataset protocol. It even gets remarkable results in 3DMAD and HKBU databases. The common element between these two databases is that all attacks are performed using rigid masks. It seems that the Color-based method (texture features over different color

spaces) achieves fair performance for this type of attacks. However, none of the methods generalizes properly in scenarios with different PAIs, as in all of them the ACER is above 22%. In this classical protocol there are many factors that are involved in an uncontrolled way and, without additional information, it is very complicated to draw fair conclusions. It is therefore necessary to carry out evaluations focusing on fewer parameters.

4.3. Proposed Evaluation: the new protocols

The proposed benchmark has three protocols that allow us to draw conclusions from another perspective: the *Cross-Devices* protocol evaluates performance on unseen capture devices; the *Cross-FaceResolution* protocol measures the effects of variations in face resolution; and the *Cross-Conditions* protocol evaluates the influence of adversarial operating conditions.

Table 5 shows the results of Quality-Based and Color-Based approaches using the Cross-Device protocol, leaving out one type of capture device (digital camera, webcam or mobile/tablet as named in the table) for training and development stages and using it on testing. Such results indicate that the most challenging scenario for the two systems evaluated is *Cross-Device-Webcam-Test*. This result seems to be closely related to the quality of the capture device, as webcams have the worst image quality. We can conclude that there is a drop in the performance of both face-PAD systems as the image quality degrades. In addition, it can be appreciated how, with this type of protocols, the Quality-based method suffers much more than the Color-based one.

	Cross-Device Protocol	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
Quality-Based	DigitalCamera-Test	24.85	52.27	86.67	17.89
	Webcam-Test	28.55	53.57	29.52	47.62
	MobileTablet-Test	21.11	25.76	29.33	22.19
Color-Based	DigitalCamera-Test	7.42	16.26	26.76	5.75
	Webcam-Test	12.16	31.90	48.98	14.83
	MobileTablet-Test	9.08	12.30	15.07	9.54

Table 5: Results for *Cross-Device* protocol

To compliment the study of capture devices, it is worth to carry out another interesting experiment, taking into account the image resolution. Besides, if we also had the information of the resolution of the detected face available, we could extract very interesting properties to figure out the optimal distance for realistic scenarios. Results using the *Cross-FaceResolution* protocol are shown in Table 6. We can observe that the Color-based method on *Cross-FaceResoultion-LF-Test* is the only experiment within this protocol that seems to achieve good generalization capabilities. In view of these results, it seems that there is a better capacity for generalization when we move towards higher resolution scenarios, having been trained with lower resolution, than in the opposite case (for the Color-based face-PAD).

	Cross-FaceResolution Protocol	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
Quality-Based	LF-Test	24.48	51.86	86.21	17.52
	SF-Test	29.98	48.79	50.00	47.58
Color-Based	LF-Test	8.33	15.81	27.50	4.12
	SF-Test	25.47	29.62	12.20	47.04

Table 6: Results for *Cross-FaceResolution* protocol

If we analyze the results reported in both Table 5 and Table 6, we can see that the most favorable case for the color-based method occurs when we train using webcams and digital cameras where the resolution of the detected faces are medium and low, and testing on mobile or tablets with a large face resolution. The results for the described combination are an ACER of 51.77% for the Quality-based method, while it is 39.07% in the case of Color-based. A smaller number of samples for training (more filtering) and a high dimensionality of the features can be the cause of this drop in performance.

Finally, the results of the *Cross-Condition* protocol are presented in Table 7. It seems that none of the baselines are able to generalize in this protocol. Results indicate that the working point shift is very large, reaching APCER values close to or above 90%. This novel protocol is therefore postulated as one of the greatest challenges for the aggregated dataset presented.

	Cross-Conditions Protocol	HTER (%)	ACER (%)	APCER (%)	BPCER (%)
Quality-Based	Test-Adverse	36.62	40.48	72.50	8.46
	Test-Optimal	45.50	66.06	96.67	35.46
Color-Based	Test-Adverse	41.13	45.43	86.25	4.61
	Test-Optimal	34.37	55.12	93.33	16.91

Table 7: Results for *Cross-Conditions* protocol

Through the experiments carried out, we have been able to analyze two methods that achieve state-of-the-art results when evaluated in individual databases and observe that, however, they do not manage to generalize adequately in conditions closer to real scenarios. But the information provided by the framework goes further, allowing us to systematically analyze the performance of such systems in the face of other variations such as changes in capture devices, image resolution or extreme mismatches between display conditions. This allows for a fairer and more consistent evaluation and comparison of face-PAD systems.

5. Conclusions

In this work we have proposed a framework, GRAD-GPAD, for systematic evaluation of the generalization properties of face-PAD methods. The GRAD-GPAD framework allows the aggregation of heterogeneous datasets, the inclusion of new feature-extraction algorithms and new evaluation protocols. We have studied the generalization capabilities of two well-known face-PAD methods using a large ag-

gregated dataset comprising ten publicly available datasets, and several protocols, including two new ones proposed in this work. The GRAD-GPAD framework allows to learn some deficiencies of these algorithms that could eventually drive to more robust and generalized feature representations for face-PAD. In short, this paper highlights the importance of data and the necessity of fair evaluation methodologies to improve the generalization of existing face-PAD methods.

Acknowledgments We thank our colleagues of the Biometrics Team at Gradiant for their valuable contributions.

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*, 2011. 2

[2] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing Deep Face Recognition With Custom Silicone Masks. In *Biometrics: Theory, Applications, and Systems*, 2018. 2, 4

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security (TIFS)*, 2016. 6

[4] Z. Boulkenafet, J. Komulainen, and A. Hadid. On the generalization of color texture-based face anti-spoofing. *Image and Vision Computing*, 2018. 2, 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. 2017. 4

[6] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012. 4

[7] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel. The replay-mobile face presentation-attack database. In *BioSIG*, 2016. 4

[8] A. Costa-Pazo, E. Vazquez-Fernandez, J. L. Alba-Castro, and D. González-Jiménez. Challenges of face presentation attack detection in real scenarios, 2019. 6

[9] T. de Freitas Pereira, A. Anjos, J. M. D. Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB 2013*, 2013. 1

[10] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. 2013. 4

[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*, 2014. 2, 6

[12] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. *International Conference on Computer Vision*, 2017. 1

[13] A. Jourabloo*, Y. Liu*, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proc. European Conference on Computer Vision*, Munich, Germany, 2018. 2, 4

[14] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE TIFS*, 2007. 2

[15] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *ICCV*, 2017. 1

[16] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE TIFS*, 2018. 2

[17] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE TIFS*, 2018. 2, 4

[18] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV*. Springer International Publishing, 2016. 4

[19] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB 2011*. IEEE, 2011. 2

[20] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE TIFS*, 2017. 2, 4

[21] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In *ICB*, 2018. 2, 3, 6

[22] A. Pinto, W. R. Schwartz, H. Pedrini, and A. d. R. Rocha. Using visual rhythms for detecting video-based facial spoof attacks. *IEEE TIFS*, 2015. 2, 4

[23] A. Sepas-Moghaddam, F. Pereira, and P. L. Correia. Light field-based face presentation attack detection: Reviewing, benchmarking and one step further. *IEEE TIFS*, 2018. 2

[24] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. Ho. Detection of face spoofing using visual dynamics. *IEEE TIFS*, 2015. 2

[25] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, and Z. Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv preprint arXiv:1811.05118*, 2018. 2

[26] D. Wen, H. Han, and A. Jain. Face Spoof Detection with Image Distortion Analysis. *IEEE TIFS*, 2015. 2, 4, 6

[27] F. Xiong and W. Abdalmageed. Unknown presentation attack detection with face rgb images. In *BTAS*, 2018. 3

[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, 2016. 5, 6

[29] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li. Casia-surf: A dataset and benchmark for large-scale multi-modal face anti-spoofing. *arXiv preprint arXiv:1812.00408*, 2018. 4

[30] S. Zhang, X. Wang, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z.Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *In Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4

[31] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *ICB 2012*, 2012. 4