Appendix: Detection and Continual Learning of Novel Face Presentation Attacks

Mohammad Rostami, Leonidas Spinoulas, Mohamed Hussein, Joe Mathai, Wael Abd-Almageed
USC Information Sciences Institute
Los Angeles, CA, 90292 USA
{mrostami,lspinoulas,mehussein,jmathai,wamageed}@isi.edu

Abstract

In this Appendix, we provide details about the PADISI-Face and details about the experimental setup.

1. PADISI-Face Dataset

Previous efforts on generating face PAD datasets have been focused on a number of major attack types, including, disguises, printed photographs, 3D masks, and replays. As novel types of attacks emerge, existing datasets might be insufficient to guarantee designing suitable PAD algorithms because there has not been enough effort in generating datasets that cover a wide variety of attack categories. Table 1 summarizes a number of highly prevalent face PAD datasets in the literature. The table provides information about the types of presentation attack instruments (PAIs) present in each dataset. As it can be seen, these datasets are limited in terms of diversity of attack types they contain. PADISI-Face is a new dataset captured from 182 different participants to offer more diverse set of attack types. Due to granular labels on these attack types, PADISI-Face is a suitable dataset for testing models in continual learning settings or when there should be significant variations between the training and testing datasets.

The PADISI-Face Dataset is collected using a sensor array designed and built by our team, shown in Figure 1 [10]. The system is designed for more comprehensive future versions of PADISI-Face that will contain beyond the visible range information. The hardware comprises of six different cameras spanning visible (RGB), short-wave-infrared (SWIR) and long-wave infrared (Thermal) electromagnetic spectrum ranges. Additionally, there are two near-infrared (NIR) cameras for high quality stereo depth estimation. For acquisition of data in NIR and SWIR spectra, a synchronized illumination of different wavelength LEDs (shown in Figure 1) were used. The synchronized sequence of LED illuminations were designed to maximize the throughput of the camera suite while increasing the temporal coherence between frames. Figure 2 shows some examples of images

Table 1: Multi-spectral PAD Datasets

Dataset	Year	Participants	Attacks
Pavlidis Symosek [7]	2000	–	Facial disguises
3DMAD [5]	2013	17	3D mask attacks
I^2BVSD [4]	2013	75	Facial disguises
GUC-LiFFAD [8]	2015	80	2D print and replay
MS-Spoof [3]	2015	21	2D print
BRSU [11]	2016	50	3D masks
EMSPAD [9]	2017	50	2D print
MLFP [1]	2017	10	2D & 3D masks
CASIA-SURF [12]	2020	1000	2D print & cutouts
MAFPAD	2020	360	2D print, mannequins 3D masks, obfuscation makeup fake tattoo, eye area cover

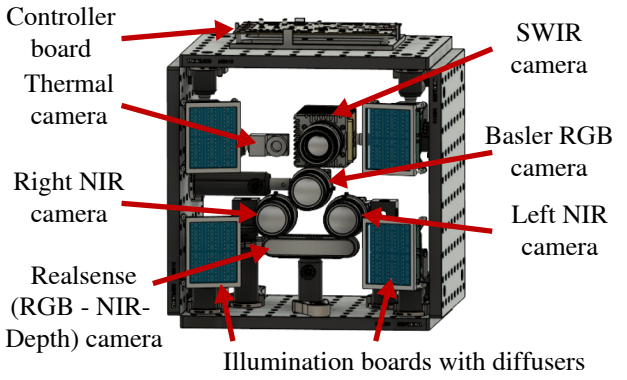

Figure 1: Face biometric sensor suite.

collected for bona-fide and several attack samples using the sensor array in various light ranges. For NIR and SWIR modalities, dark channel subtraction is performed to reduce the effect of ambient illumination. Data was collected from each participant over two rounds. In the first round, bona fide samples were collected. Participants presented a presentation attack instrument (PAI) in the second round. PADISI-Face will be available for the use of the research community.

To enable face detection in all captured frames, we use a standard calibration process using checkerboards [13]. For the checkerboard to be visible in all wavelength regimes,

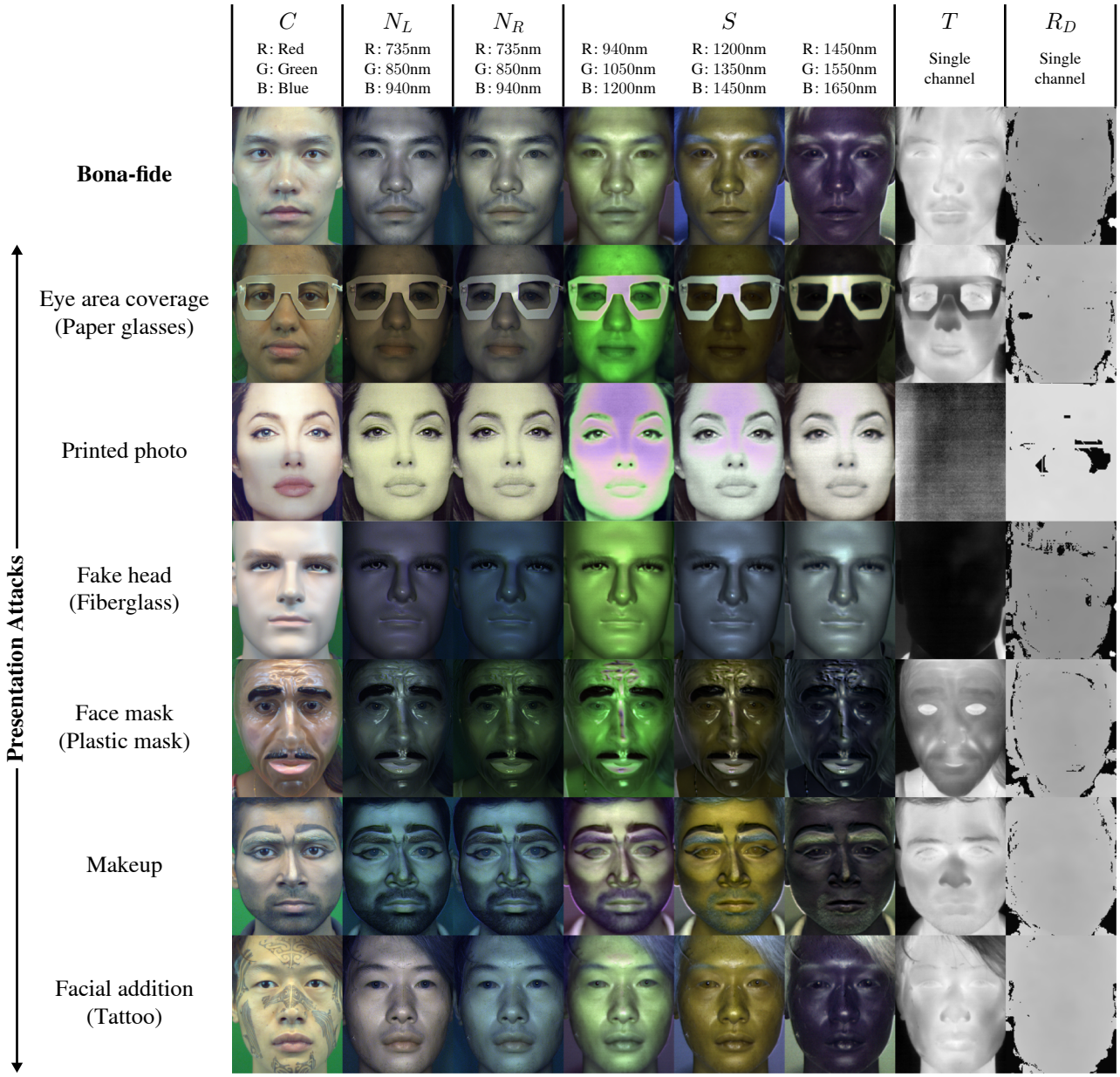

Figure 2: RGB visualization of samples collected for all types of PAIs are summarized in this figure. For each image, the corresponding dark channel has been subtracted and each RGB channel has been min-max normalized for visualization purposes. Note, not all images have the same resolution but are resized for fine arrangement.

a manual approach is used when a sequence of frames is captured offline while the checkerboard is being lit with a bright halogen light. This makes the checkerboard pattern visible and detectable by all cameras which allows the standard calibration estimation process to be followed. The face can then be easily detected in the RGB space [2] and the calculated transformation for each camera can be applied to detect the face in the remaining camera frames.

Following the aforementioned approach, face landmarks are detected on the visible spectrum using [2]. A bounding box is then constructed from the landmarks to have a tight crop of the face. The bounding box of the visible spectrum is then projected to the corresponding co-ordinate system of the other cameras to extract approximately aligned faces on different modalities. Each channel is then scaled to the range $[0, 1]$ by dividing the bit depth of the camera and then resized to 160×160 pixels. In our experiments, we use the visible range information as input to our algorithm.

Table 2: Effect of knowing labels for the challenging data points using the PADISI-Face dataset in the single PA/task scenario. Results for NACL are taken from the bottom left of Fig. 5.

Task	APCER (%)		BPCER (%)		ACER (%)	
No.	NACL	DL	NACL	DL	NACL	DL
1	46.1	27.8	10.7	4.8	28.4	16.3
2	38.2	24.8	10.1	1.0	24.1	12.9
3	41.4	27.0	14.1	1.5	27.8	14.2
4	29.1	19.9	15.0	1.1	22.0	10.5
5	15.1	12.5	18.1	4.3	16.6	8.4
6	19.4	16.0	18.0	3.7	18.7	9.8
7	17.2	15.1	17.0	2.6	17.1	8.9
8	18.2	13.8	17.4	2.8	17.8	8.3

A future direction includes considering beyond the visible information to perform PAD.

2. Reducing False-Positive Predictions

Our primary focus has been on reducing the false-negative predictions. In practical settings, we can assume labels for detected novel data points can be accessible with a delay by the end of each task, e.g., using manual annotation. To model this possibility, we performed an experiment using a selection scheme that assumes manual annotation is possible, i.e., label pollution is reduced to zero. Since updating the model occurs at discrete periods at the end of each task, we have assumed the delay for labeling is less than the time needed to update the model. Hence by the time a task finishes, we assume the labels for the novel samples are accessible before updating the model. Table 2 presents results for the PADISI-Face dataset in the single PA/task scenario, where we have compared Delayed Labels (DL) with NACL. We observe that this sampling scheme leads to reduced false-negative predictions. Additionally, we observe BPCER performance also improves.

3. Experimental Setup Details

We provide details that we used to perform experiments.

3.1. Network structure

In our experiments, the network is consisted of a pre-trained fixed backbone encoder, followed by fully connected layers to reach to the label space.

Backbone Model

An important limitation of CNN models when trained on small datasets, such as biometric datasets, is that they tend to select features which are not generalizable due to overfitting. For this purpose, we opted for employing MoCo-v1 as a fixed backbone network [6] to improve generalizability of the extracted features. This network is trained on ImageNet using a contrastive loss that attempts to find similarities and dissimilarities among synthesized variants of training data samples in an unsupervised way. It is subclass of self-supervised learning at which a deep neural network is trained to solve pseudo-tasks. As a result, the network learns to extract discriminative features at its early layers to solve the pseudo-tasks. Thus, when we use MoCo-v1 as our backbone for PAD in a continual learning setting, no input or label information of the future PA types have been used to train the feature extraction model. This property ensures that no information about the training dataset has been used in training the model. This allows to claim that the new attacks are indeed unseen. We note that extracting features using this pre-trained network leads to an separability of different types of attacks, as shown in the t-SNE visualizations of Figure 3, for PADISI-Face dataset as an example which enables our model to identify data points that belong to new attack classes. This observation demonstrates we can use the backbone model as a good feature extractor to identify OTDS.

Learnable Layers

We use the same end-to-end network structure for a fair comparison among the methods. The MoCo backbone is followed by three fully connected layers with 64, 32, and 2 (turn into 3 nodes when the model is trained to identify OTDS) nodes each. MoCo's encoder is in essence a ResNet50 architecture with 128 output nodes, used here as discriminative feature vectors to improve classification. In all experiments, the weights of the backbone network are frozen and learnable parameters θ in our formulated would refer to the last fully connected layers. We use ReLU non-linearity in the first two layers and softmax non-linearity in the final layer. We have selected the layer with 32 nodes to represent the embedding space $\mathcal{Z}$ on which the CL approach is performed, as described in the paper. At each task, the network is trained with 2 output nodes and performance on the testing split is measured during training. When the task is learned, the network output is extended to include a third output. After identifying the OTDS, the network again is trained with 2 outputs. This process is continued until all tasks are learned. To reduce redundancy of inference at stochastic gradient descent step, we compute the input features initially and perform optimization just on the learnable layers. By doing so, we reduce learning time but have the understanding that in practice, inference also needs to be performed end-to-end.

3.2. Implementation Parameters

We use the cross entropy loss as the discrimination loss. We used Keras for implementation of the algorithm and the Adam optimizer to perform stochastic gradient descent. The learning rate is set to be 2×10^{-4} with a decay rate of 10^{-4}. We use a batch size of 100. At each batch, we select 100 points randomly and make sure the batch is balanced. To learn each task, we randomly initialize all the trainable weights (fully connected layers) and perform opti-

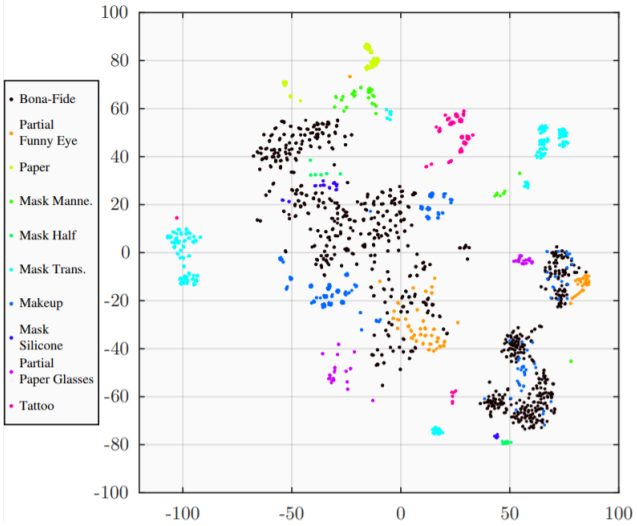

Figure 3: t-SNE visualization of features obtained using the pre-trained model of [6] (MoCo-v1), for PADISI-Face dataset. One frame, per capture, is used for this visualization.

mization using 10000 batches. At each training epoch, we computed the loss function on the training data split and the performance metrics on the testing split. We ran our code on a cluster node equipped with 4 Nvidia Tesla P100-SXM2 GPU's. Our code is provided as part of the supplementary material.

References

[1] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore. Face Presentation Attack with Latex Masks in Multispectral Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 275–283, 2017. 1

[2] A. Bulat and G. Tzimiropoulos. Super-FAN: Integrated Facial Landmark Localization and Super-Resolution of Real-World Low Resolution Faces in Arbitrary Poses with GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 109–117, 2018. 2

[3] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face Recognition Systems Under Spoofing Attacks. In Thirimachos Bourlai, editor, *Face Recognition Across the Imaging Spectrum*, pages 165–194. Springer International Publishing, Cham, 2016. 1

[4] Tejas Indulal Dhamecha, Richa Singh, Mayank Vatsa, and Ajay Kumar. Recognizing Disguised Faces: Human and Machine Evaluation. *PLOS ONE*, 9(7):1–16, 07 2014. 1

[5] N. Erdogmus and S. Marcel. Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–6, 2013. 1

[6] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 3, 4

[7] I. Pavlidis and P. Symosek. The imaging issue in an automatic face/disguise detection system. In *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No.PR00640)*, pages 15–24, 2000. 1

[8] R. Raghavendra, K. B. Raja, and C. Busch. Presentation Attack Detection for Face Recognition Using Light Field Camera. *IEEE Transactions on Image Processing*, 24(3):1060–1075, 2015. 1

[9] R. Raghavendra, K. B. Raja, S. Venkatesh, F. A. Cheikh, and C. Busch. On the vulnerability of extended Multispectral face recognition systems towards presentation attacks. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–8, 2017. 1

[10] Leonidas Spinoulas, Mohamed E. Hussein, David Geissbühler, Joe Mathai, Oswin G. Almeida, Guillaume Clivaz, Sébastien Marcel, and Wael Abd Almageed. Multispectral biometrics system framework: Application to presentation attack detection. *IEEE Sensors Journal*, pages 1–1, 2021. 1

[11] Holger Steiner, Sebastian Sporrer, Andreas Kolb, and Norbert Jung. Design of an Active Multispectral SWIR Camera System for Skin Detection and Face Verification. *Journal of Sensors*, 2016:16, 2016. 1

[12] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li. CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 1

[13] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, Nov 2000. 1