CNN PATCH POOLING FOR DETECTING 3D MASK PRESENTATION ATTACKS IN NIR

Ketan Kotwal and *Sébastien Marcel*

Idiap Research Institute
Martigny, Switzerland

ABSTRACT

Presentation attacks using 3D masks pose a serious threat to face recognition systems. Automatic detection of these attacks is challenging due to hyper-realistic nature of masks. In this work, we consider presentations acquired in near infrared (NIR) imaging channel for detection of mask-based attacks. We propose a patch pooling mechanism to learn complex textural features from lower layers of a convolutional neural network (CNN). The proposed patch pooling layer can be used in conjunction with a pretrained face recognition CNN without fine-tuning or adaptation. The pretrained CNN, in fact, can also be trained from visual spectrum data. We demonstrate efficacy of the proposed method on mask attacks in NIR channel from WMCA and MLFP datasets. It achieves near perfect results on WMCA data, and outperforms existing benchmark on MLFP dataset by a large margin.

Index Terms— Biometrics, Face Presentation Attack Detection, 3D Mask Attacks, Patch Pooling Layer

1. INTRODUCTION

Face recognition (FR) systems have achieved excellent accuracies; however, their reliability and security in terms of detecting a presentation attack (PA), or *anti-spoofing* is still an important weakness. Presentation attacks can be classified as 2D or 3D depending on the nature of instrument used to construct an attack. The detection of all kinds of PAs is crucial for trustworthy functioning of the FR system. A majority of 2D attacks (print, digital display) can be recognized with a reasonable accuracy by RGB data alone, or by incorporating an additional data acquisition channel such as infrared, thermal, or depth. However, detection of 3D masks is a challenging task for RGB (visual spectrum) data, as well as, for data acquired from any aforementioned imaging channels. In this work, we address the problem of detection of 3D mask attacks—where the term 3D mask refers to a broad variety of masks in terms of quality and material- from a simple paper mask to a hyper-realistic, customized mask made from soft silicone.

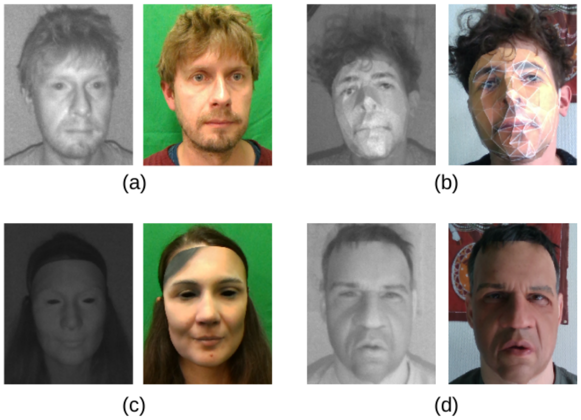


Fig. 1. Samples of *bona fide* and different mask attack presentations from WMCA [1] dataset: (a) *bona fide*, (b) paper mask, (c) custom mask, and (d) flexible silicone mask. In each case, the left image is captured in NIR, and right image is captured in RGB channel.

Fig. 1 shows samples of several types of masks and *bona fide* (BF) presentations captured in RGB channel. It can be seen that sophisticated masks are extremely good at mimicking visual appearance of a human face; and thus, PAD methods based on handcrafted features are less effective at detecting 3D masks. Along with improvising RGB-based PAD methods, a different imaging channel needs to be explored for devising a better PAD algorithm. We propose NIR-based PAD system to detect mask attacks as NIR devices are relatively cheaper, and easily available.

Deep learning-based methods, mainly using convolutional neural networks (CNNs), have demonstrated superior performance at face PAD [2]. Since textural features are important cues for detecting PAs, we propose a patch-based feature descriptor that encodes rich textural information from last convolutional layer of a CNN. Instead of processing the entire feature map as a whole, we define a novel *patch pooling* layer that facilitates learning complex texture features of an input without emphasizing information related to its shape. Whereas lack of training data is often a concern for CNN-based PAD methods; our proposed method can be deployed using a CNN, pretrained for FR tasks with visual spectrum data, without any explicit need for finetuning or domain adaptation. On training a suitable linear classifier, the proposed

Ketan Kotwal (ketan.kotwal@idiap.ch) is the corresponding author. This work has been supported by The European H2020-ICT project TeSLA (grant agreement no. 688520); and the Swiss Center for Biometrics Research and Testing.

patch-pooled features obtained excellent results with nearly 0% error rates on our test dataset. This dataset is a subset of challenging WMCA [1] dataset that consists of masks made from paper, rigid materials, and soft silicone.

Specific contributions of this paper are as follows:

- We propose a PAD method to detect a variety of mask attacks using a single NIR channel. Very few works have addressed this problem, while most existing works rely on multiple imaging channels, and their fusion.
- Through patch-pooled feature descriptors, we demonstrate novel mechanism to encode complex texture cues from CNN for face PAD.
- We show that a CNN pretrained for FR tasks is an efficient feature extractor for detecting masks without explicit transfer learning or fine-tuning. The imaging channels used to train the CNN need not be same as that of presentations for PAD.
- We demonstrate efficacy of the proposed PAD method on two publicly available datasets consisting a variety of masks—where it outperforms state-of-the-art results by a large margin.

2. RELATED WORK

Initial research in detecting 3D PAs was confined to rigid masks provided by 3DMAD [3], and HKBU-MARs [4] datasets. Most of the face PAD methods were based on handcrafted features [5–7]. Recently, the research focus has been detection of customized and high quality masks—since such masks have become relatively affordable.

Manjani *et al.* [8] introduced a face PAD dataset, named SMAD, consisting of silicone mask-based attacks captured in visual spectrum. They also proposed a deep dictionary learning method for mask PAD. Agarwal *et al.* [9] analyzed a variety of texture-based PAD approaches for detection of latex masks captured in visual, NIR, and thermal channels. Their results show that a combination of Redundant Discrete Wavelet Transform (RDWT) and Haralick features provides better separation of features (BF v/s PA) for NIR and thermal data.

Liu and Kumar investigated several CNN configurations to detect mask attacks captured in RGB and NIR channels [10]. Their results indicate superiority of NIR-based PAD methods over processing of presentations acquired in visual spectrum. In [1], a multi-channel CNN (MC-CNN) has been proposed to detect a variety of 2D and 3D PAs from WMCA dataset—that are captured in RGB, NIR, thermal, and depth channels. The initial convolutional layers of MC-CNN are adapted for each imaging channel, while the higher layers are shared across all imaging channels. Kotwal *et al.* demonstrated that embeddings (output of pre-final fully connected layer) of FR CNN can be directly utilized to identify PAs constructed using custom silicone masks [11].

3. PROPOSED METHOD

Masks, with various materials, exhibit different textural patterns than that of a natural human skin. Although PAD methods using handcrafted texture features have provided good results for a variety of PAs, their performance at detecting masks is relatively poor. Most of these methods process low-level texture features obtained directly from the input presentations. Therefore, learning a set of complex texture features might be helpful at improving accuracy of mask detection.

Deep CNNs have proved to be an excellent choice for extracting complex features for a wide range of applications. A typical CNN consists of a set of convolutional (`conv`) layers, followed by one or more fully connected (`FC`) layers. The conv layers learn characteristics of local regions in the input image; wherein the shape related properties are captured through the FC layers [12]. The output of FC layer of a CNN encodes spatial information; and thus, emphasizes shape-related (global) features of input rather than local ones. Texture, being a local or region-based feature, can be well-learnt through conv layers with minimal inference of location information. A better representation of information from final (or intermittent) conv layers of CNN, with lesser influence of spatial details, can be more discriminative toward detecting mask attacks. To obtain this representation, we propose a novel patch-pooling (`PP`) layer to be incorporated after the final conv layer of a CNN. In [11], it has been demonstrated that FR CNN is a good choice for extracting or learning features for PAD. Therefore, we also consider FR CNN as a backbone or base network for this work.

Patch Pooling (`PP`) Layer:
Few researchers have addressed the problem of computing a texture descriptor from conv layers of a deep CNN [12–14]. These works, however, have been developed for data from visual spectra. To the best of our knowledge, despite increasing popularity of NIR-based face PAD, no work on defining texture descriptor suitable for PAD using NIR data has been conducted.

Inspired from texture descriptors in [13, 14], we develop a simple, and computationally inexpensive patch pooling layer for NIR-based mask detection. Let the last conv layer of a given FR CNN generate N feature maps of $(X \times Y)$ dimensions each. Each feature map represents densely pooled local features of the input presentation. We divide each feature map into spatially non-overlapping patches of $(x \times y)$ dimensions, such that, $X = m_1 x$, and $Y = m_2 y$, for $m_1, m_2 \in \mathbb{Z}^+$. With this procedure, we obtain tessellated feature maps of $(N \times x \times y)$ dimensions for $m_1 m_2$ patches. Each of these maps is then linearized to produce a patch-level descriptor, $\mathbf{p}_k$, such that, $\mathbf{p}_k \in \mathbb{R}^{Nxy},\ k = 1, 2, \cdots, (m_1 m_2)$. At this stage, we obtain such $m_1 m_2$ patch-level descriptors, each representing a (linearized version of) complex features learnt by conv layers of CNN over a small patch of input presen-

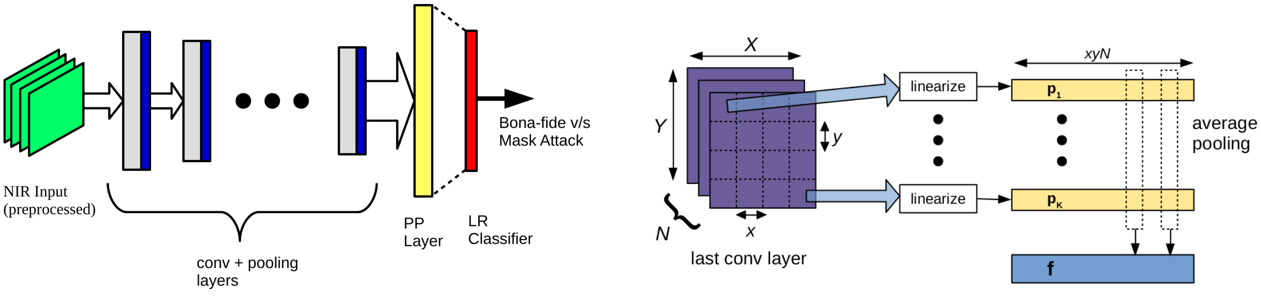


Fig. 2. Framework of the proposed PAD method for mask detection (left), and patch pooling (PP) layer (right).

tation. We compute the final descriptor, $\mathbf{f}$, through *average pooling* of $m_1 m_2$ vectors such that,

$$\mathbf{f} \equiv f^i = \sum_{k=1}^{m_1 m_2} p_k^i; \quad i = 0, 1, \cdots, Nxy - 1. \tag{1}$$

The index i refers to the i-th element of a descriptor; and $\mathbf{p}_k$ is the descriptor for k-th patch. We refer to the overall process of obtaining a feature descriptor, $\mathbf{f}$, through pooling of patch-level features of last conv layer of a CNN as patch pooling layer. The schematics of overall PAD framework and PP layer are provided in Fig. 2.

It may be noted that our PP layer does not produce output that is strictly independent of spatial information. The local geometry of a patch is implicitly encoded during its linearization. A small region of image, however, is essential to learn texture through spatial neighborhood. Since, area of a patch is significantly lesser than the entire input presentation, it results in nominal emphasis on overall shape information. Additionally, we are average-pooling such $m_1 m_2$ patch-level features ($\mathbf{p}$) to obtain the final descriptor, $\mathbf{f}$; thereby further mitigating the effect of location information. The output of PP layer is, thus, *mean-local representation* of textural features learnt by conv layers of FR CNN.

Classifier: We formulate mask-based PAD as a binary classification problem. We train a linear classifier from the outputs of PP layer to compute final score.

4. EXPERIMENTS

We demonstrate efficacy of the proposed PAD method over two publicly available datasets[1].

4.1. Datasets & Protocols

Very few publicly available PAD datasets consist of 3D mask attacks in NIR channel. We conduct our experiments on two such datasets- (*a*) Wide Multi Channel Presentation Attack (WMCA) [1]; and (*b*) Multispectral Latex Mask based Video Face Presentation Attack (MLFP) [9].

We consider the subset of WMCA dataset consisting of BF and 3D mask attacks acquired in NIR channel. The PAs include masks made from paper, latex, and silicone. Some samples of these presentations in RGB and NIR ($860\,nm$) channels are shown in Fig. 1. This subset of WMCA consists of 240 BF presentations, and 487 PA mask-based PAs. We use a *grandtest* protocol derived from the original protocol of creators of dataset. It consists of three fixed partitions (training, development, evaluation) that are disjoint (non-overlapping) in terms of subjects. Our second test dataset comprises of NIR presentations from MLFP dataset. The PAs in this dataset are constructed using paper and latex masks. It consists of 40 BF and 400 attack presentations in NIR channel captured using Kinect (Windows V2). We design our cross validation (*CV*) protocol using the subject-based partitioning devised by the creators of dataset [9, Sec. 4.1]. For brevity of space, we do not provide details of any protocol. The exact details of both protocols can be obtained from our code.

4.2. Experimental Setup

Our proposed face PAD method, based on PP layer, derives features from conv layers of FR CNN. We utilize the 9-layer LightCNN [15], one of the state-of-the-art FR CNNs, as a backbone or feature extractor (for conv layers). We consider the outputs of final conv layer of LightCNN-9 (MFM5 as per [15]) as input to the PP layer to compute the proposed feature descriptor. This descriptor is then used to train or test the logistic regression (LR)-based classifier.

We use the equal error rate (EER) on the dev set to compute the score threshold—which refers to approximately equal number of incorrect classifications in both classes. We evaluate PAD experiments using the following performance metrics: (*a*) **APCER** (Attack presentation classification error rate) defined as the proportion of incorrectly classified PAs; (*b*) **BPCER** (*Bona fide* presentation classification error rate) defined as the proportion of incorrectly classified BF presentations; and (*c*) **ACER** (Average classification error rate) calculated as the average of APCER and BPCER.

[1] Python code for all experiments described in this paper: `https://gitlab.idiap.ch/bob/bob.paper.nir_patch_pooling`

For all experiments, we prepared the input presentation to match the specifications of our FR CNN (LightCNN-9). Each frame of NIR data has been clipped to 8-bit range, where clipping limits were computed using the median absolute deviation (MAD). We have used the Multi-Task Cascaded Convolutional Network (MTCNN) [16] to detect facial region. For very few samples, we were not able to detect the faces either due to highly non-face like appearance of sample (such as poor mask), or due to limitation in our face detection mechanism.

4.3. Experimental Results

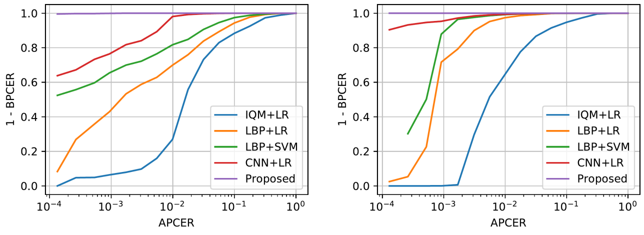
WMCA dataset using grandtest protocol: In this main experiment, we evaluated performance of the proposed PAD method on the WMCA dataset using grandtest protocol. The train set is used to compute the feature descriptor using (a part of) LightCNN and PP layer. These features are then used to train LR classifier. The score thresholds are computed on the dev set. Table 1 provides performance evaluation of the proposed method along with some common and recent baseline methods.

For baseline, we consider following commonly used PAD approaches: (*a*) **IQM+LR** method: based on image quality measures (IQM) as described in [17], and classified using an LR classifier. (*b*) **LBP+LR** method: We compute uniform $\text{LBP}_{8,1}^{u2}$ codes on input presentations, and their histograms are classified using an LR classifier. (*c*) **LBP+SVM** method: Here, the LBP histograms (as described in previous method) are classified using a support vector machine (SVM) with a radial basis function kernel. (*d*) **CNN+LR** method: This method considers CNN embeddings as features [11]. These are classified using an LR classifier.

For the proposed PAD method, using patch-based pooling, the ACER on dev set was dropped to 0.1%. On eval set, only a single BF frame was misclassified as attack, and all mask presentations were correctly identified. Therefore, for 11390 frames in eval set, we obtained ACER of 0.008%. With near-perfect classification, the proposed method clearly outperformed the baselines on every set. The CNN+LR method from [11] is of particular interest since it also uses FR CNN to generate feature descriptors. However, it considers the output of prefinal FC layer as the features, while the proposed method replaces FC layer(s) with a novel PP layer that pools the features from patches of input presentations. The improvement in results indicates superiority of patch-level features over FC-level features toward extracting textural cues; and thereby, detecting mask attacks on FR systems.

Table 1. Performance evaluation of the proposed method and baselines on the WMCA dataset for grandtest protocol. All measure rates are in %. The numbers in parenthesis indicate the number of incorrectly classified samples for total samples in the given class.

PAD Method	dev set	eval set		
	ACER	APCER	BPCER	ACER
IQM + LR	10.9	4.8 (372/7691)	9.7 (360/3699)	7.3
LBP + LR	7.6	1.4 (104/7691)	2.1 (79/3699)	1.7
LBP + SVM	5.4	0.7 (52/7691)	1.0 (38/3699)	0.9
CNN + LR	1.4	0.3 (26/7691)	1.4 (52/3699)	0.9
Proposed	0.1	0.0 (0/7691)	0.0 (1/3699)	0.0



(a) dev set

(b) eval set

Fig. 3. ROC of PAD methods on WMCA dataset using *grandtest* protocol.

Fig. 3 illustrates the receiver operating characteristics (ROC) curves for dev and eval sets of the WMCA dataset. For both sets, a near-zero value of BPCER can be observed for the entire range of APCER—which is an ideal condition for any PAD system.

MLFP dataset using subject-based protocol: We conducted 3-fold CV experiments using subject-based partitioning of MLFP dataset. The score thresholds were chosen *a posteriori* on the testing partition for given trial. The ACER values for each trial are provided in Table 2. The proposed PAD method resulted in ACER of 1.9% averaged across trials. This is nearly 20× improvement over the baseline results from [9]. It should also be noted that the classification accuracy of the worst trial is above 97%.

5. CONCLUSION

We have proposed a CNN-based face PAD method to detect 3D mask attacks in NIR channel. This method employs a patch-pooling mechanism to learn textural cues from final conv layer of CNN. We have also demonstrated that a CNN, pretrained for FR using visual spectrum data, can be directly used to compute the patch-pooled feature descriptor. The proposed PAD method has been tested on two publicly available datasets that consists of masks made of paper, latex, and silicone. Excellent results, on both datasets, indicate that the patch pooling mechanism is well-suited for discriminating mask-based PAs in NIR channel.

Table 2. Performance evaluation of the proposed method on the MLFP dataset for cross validation trials. All values are ACER (%).

Trial	CV1	CV2	CV3	Average
Baseline (frame-based)	-	-	-	44.4
Baseline (video-based)	-	-	-	42.0
Proposed	2.9	1.5	1.4	1.9

6. REFERENCES

[1] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2020.

[2] A. Liu et al., "Multi-Modal Face Anti-Spoofing Attack Detection Challenge at CVPR2019," in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, June 2019.

[3] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with Kinect," in *Proceedings of International Conference on Biometrics: Theory, Applications and Systems*, Sep. 2013, pp. 1–6.

[4] S. Liu, B. Yang, P. Yuen, and G. Zhao, "A 3D Mask Face Anti-Spoofing Database with Real World Variations," in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, June 2016, pp. 1551–1557.

[5] T. Siddiqui, S. Bharadwaj, T. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha, "Face anti-spoofing with multifeature videolet aggregation," in *Proceedings of International Conference on Pattern Recognition*, Dec 2016, pp. 1035–1040.

[6] R. Raghavendra and C. Busch, "Robust 2D/3D face mask presentation attack detection scheme by exploring multiple features and comparison score level fusion," in *Proceedings of International Conference on Information Fusion*, July 2014, pp. 1–7.

[7] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral SWIR imaging," in *Proceedings of International Conference on Biometrics*, June 2016, pp. 1–8.

[8] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, July 2017.

[9] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face Presentation Attack with Latex Masks in Multispectral Videos," in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, July 2017, pp. 275–283.

[10] J. Liu and A. Kumar, "Detecting Presentation Attacks from 3D Face Masks Under Multispectral Imaging," in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops*, June 2018, pp. 47–52.

[11] K. Kotwal, S. Bhattacharjee, and S. Marcel, "Multispectral deep embeddings as a countermeasure to custom silicone mask presentation attacks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 4, pp. 238–251, Oct 2019.

[12] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of Conference on Computer Vision and Pattern Recognition*, June 2015, pp. 3828–3836.

[13] V. Andrearczyk and P. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016.

[14] K. Kotwal, Z. Mostaani, and S. Marcel, "Detection of Age-Induced Makeup Attacks on Face Recognition Systems Using Multi-Layer Deep Features," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 15–25, Jan 2020.

[15] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov 2018.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[17] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The Replay-Mobile Face Presentation-Attack Database," in *Proceedings of International Conference of the Biometrics Special Interest Group*, Sep. 2016, pp. 1–7.