CHIF: Convoluted Histogram Image Features for Detecting Silicone Mask based Face Presentation Attack

Akshay Agarwal, Mayank Vatsa, and Richa Singh
IIIT-Delhi, India
{akshaya, mayank, rsingh}@iiitd.ac.in

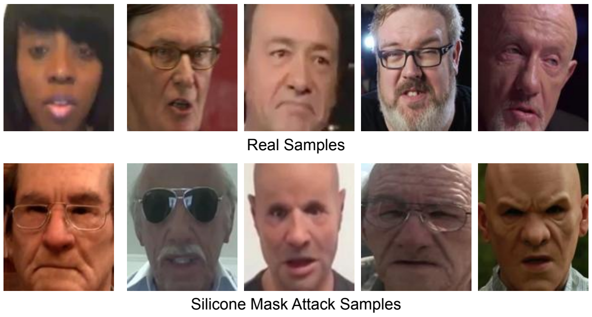
Abstract

Face recognition algorithms are generally vulnerable towards presentation attacks ranging from cost-effective ways such as print and replay to sophisticated mediums such as silicone masks. Carefully designed silicone masks have real-life face texture once wore and can exhibit facial motions; thereby making them challenging to detect. In the literature, while several algorithms have been developed for detecting print and replay based attacks, limited work has been done for detecting silicone mask-based attack. In this research, we propose a computationally efficient solution by utilizing the power of CNN filters, and texture encoding for silicone mask based presentation attacks. The proposed framework operates on the principle of binarizing the image region after convolving the region with the filters learned via CNN operations. On the challenging silicon mask face presentation attack database (SMAD), the proposed feature descriptor shows 3.8% *lower error rate than the state-of-the-art algorithms.*

1. Introduction

Face recognition is one of the widely used biometric modalities. However, attacking face recognition systems is also easy and researchers have shown that these systems are vulnerable against different kinds of presentation attacks [1, 4, 5, 7, 12, 18, 24]. For example, fooling a face recognition algorithm using print and replay attack is one of the easiest approaches. More sophisticated approaches include latex mask and silicone mask based presentation attack [5, 25]. Figure 1 illustrative some examples of silicone mask based presentation attacks.

The aim of presentation attacks is two folds: (i) *impersonation*: which is a targeted attack where the attacker wants to gain access to someone else's identity and (ii) *obfuscation*: which is an untargeted attack where the attacker wants to hide his/her own identity. Different kinds of presentation attack methodologies can be used to attack the face recognition systems. The attacking method such as printed photo and replay of a video comes under the category of 2D attacks and are the most popular in the literature. Due to the availability of multiple 2D attack databases, most of the existing detection algorithms are developed for print and replay attacks. However, these attacks generally suffer from edge artifacts, moiŕe pattern, and 3D shape. The above limitations of the attack mechanisms can be overcome by sophisticated silicone masks of the target identity for impersonation or generic identity for evasion. The algorithms which show higher detection rate on 2D attacks show higher error rates for sophisticated silicone mask-based attacks [11, 32, 35].



Real Samples

Silicone Mask Attack Samples

Figure 1. Real and silicone mask attack samples of SMAD database [25].

This research aims to develop a silicone mask based presentation attack detection algorithm, which is not only accurate but also computationally efficient. The proposed algorithm utilizes the filters of pre-trained Convolutional Neural Network (CNN) for computation of histogram-based feature descriptor. For an effective presentation attack detection algorithm, a challenging database captured in unconstrained settings is essential. For this purpose, we have used the silicone mask attack database (SMAD) [25] in which the images/videos vary in terms of illumination, pose, background, and facial accessories. The contributions of this research can be summarized as: **(i)** a new descriptor is proposed for textural feature extraction by utilizing pretrained CNN filters and **(ii)** comparison with state-of-the-art silicone mask face presentation attack detection algorithms is

978-1-7281-1522-1/19/$31.00 ©2019 IEEE

performed to show the efficacy of the proposed feature descriptor.

2. Related Work

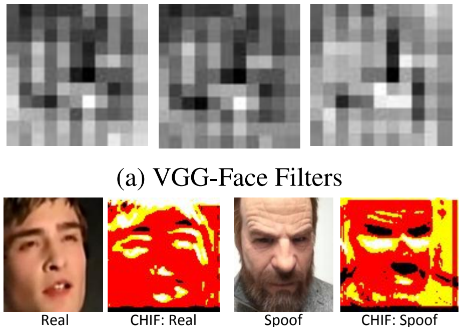
In the literature, texture-based algorithms are popular for presentation attack detection because of the artifacts such as blurring and Moiŕe patterns. For example, Määttä et al. [23] have proposed multi-scale local binary pattern (LBP) to differentiate real face from fake face. Chingovska et al. [9] have used variants of LBP and other classifiers. Agarwal et al. [3] have used the combination of wavelet decomposition and Haralick texture features for 2D and 3D hard resin attacks. To handle the variation in illumination and attack mediums, Boulkenafet et al. [8] have extracted different texture features from chrominance and luminance components of different image domains such as HSV and YCbCr. To handle the limitations of simple print and photo-based attack, motion features are also an important cue. Pan et al. [29] have proposed eye-lid motion detection using a conditional random field. Li et al. [36] have used the pulse motion for the detection of 3D masks. Bharadwaj et al. [6] and Siddiqui et al. [32] have performed the fusion of texture and motion features. Erdogmus and Marcel [10] have used multi-scale LBP for 3D mask attack detection. rPPG signals have gained attention for detecting 3D mask attacks [19, 20]. Recently, Liu et al. [21] have presented the progress towards detection of 3D mask attacks. Agarwal et al. [2] have proposed feature tampering for fooling the face presentation attack detection algorithms.

With the success of CNN models for object recognition and face recognition, several researchers have explored them for face PAD task. Yang et al. [38] have proposed an end-to-end framework for presentation attack detection. Manjani et al. [25] have proposed deep dictionary learning for silicone mask detection. Xu et al. [37] and Li et al. [17] have used long-short-term-memory and 3D CNN for presentation attack detection. Mehta et al. [26] have proposed the CNN architecture for detecting 2D/3D physical and digital presentation attacks.

In this research, we have proposed a new texture-based feature descriptor by first filtering the input images using non-linearly learned filters. The proposed descriptor is inspired from popular histogram feature descriptors such as LBP, its variants [27, 28], and binarized histogram image features (BSIF) [16]. While existing texture based approaches, generally, use handcrafted filters, the proposed feature descriptor utilizes the pre-trained CNN filters, which are computed using various non-linear operations.

3. Proposed CHIF Algorithm

The proposed feature engineering algorithm combines the concepts from histogram-based image descriptors such as LBP/BSIF and the filter learning capability of CNN. The proposed algorithm first convolves the image patches using CNN learned filters followed by the comparison of a center pixel with surrounding pixels in that patch. The responses are then binarized and a binary string is obtained by concatenating each binarized output in the patch. The binarized string is then converted into the decimal value, and finally, the histogram feature vector is calculated to represent the image. The vector represents the local intensity pattern of an image convolved with non-linearly learned filters. Hence, the proposed feature descriptor is termed as *'Convoluted Histogram Image Features (CHIF)'*. We assert that the proposed filter response captures the image texture better than existing texture-based approaches.



(a) VGG-Face Filters

(b) Original and CHIF Images

Figure 2. Sample VGG-Face learned filters of size 11×11 at layer 1 and CHIF visualization of real and spoof faces.

Filter Learning: The filters to convolve the image patches are obtained by learning a CNN architecture. The traditional CNN consists of multiple layers stacked together, and initially images are convolved with randomly initialized filters. The convolution operation is defined as:

$$f = W^T \times I + b \tag{1}$$

where, I is the image, W is the filter, and b is the bias term.

The convolved filter outputs are then passed through the sub-sampling layer with max or average pooling. ReLU non-linearity is applied on the sub-sampled filtered outputs (f'). The ReLU outputs are computed as: $f'(x) = max(0, x)$, where max is the maximum operator and x is the value in f'. The multiple convolutional, pooling, and ReLU layers are stacked together with a combination of flattening layer at the end for error calculation. The randomly initialized filters, i.e., W, are then optimized using a gradient descent algorithm.

CHIF: To compute the binary string let P be the image patch, and W be the non-linear filter obtained from a trained CNN. The filter is multiplied with the patch P as follows:

$$f_P = \sum_{x,y} W(x,y)P(x,y) = W \times P \tag{2}$$

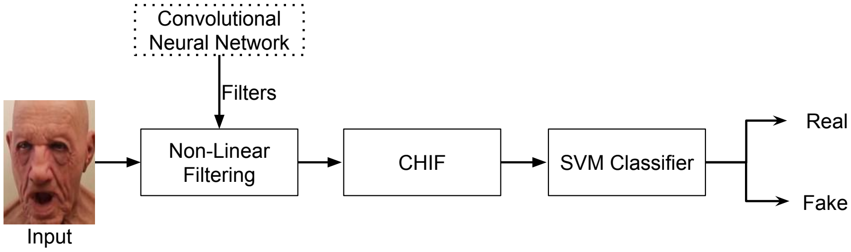


Figure 3. Pipeline of the proposed PAD algorithm with CHIF feature engineering block.

Next, f_P is binarized based on its sign i.e., if $f_P > 0$, then the binary value 1 is assigned, else 0 is assigned.

$$s(x,y) = \begin{Bmatrix} 1 & f_P > 0 \\ 0 & f_P <= 0 \end{Bmatrix} \tag{3}$$

As discussed earlier, each layer of CNN can have multiple filters. The filters can be combined at third dimension to form one big matrix of size $n \times a \times b$, where n is the number of filters of size $a \times b$. This filter matrix is then convolved with image patch and each response is binarized. The decimal value from the binary string is calculated using the following equation:

$$CHIF_8 = \sum_{i=1}^{8} s(x,y) \times 2^i \tag{4}$$

In this research, we have experimented with the first and second layer filters (i.e., rich in edge information [39]) from pre-trained[1] VGG16, VGG19 [33], GoogLeNet [34], ResNet50, ResNet101 [14], and VGG-Face [30] for computing the CHIF image descriptor. Figure 2 shows some of the learned filters of VGG-Face which served as the basis of the proposed algorithm and the associated CHIF image visualization. The PAD framework using CHIF is shown in Figure 3. Here, we have utilized the first layer filters of VGG-Face for CHIF features and linear support vector machine (SVM) classifier for binary (real vs spoof) classification. The parameters of linear SVM are optimized using grid search over training set.

4. Experimental Settings and Results

In this section, the PAD database used to perform the experiments is described along with its associated protocol, followed by the algorithms used for comparison. Later in this section, the PAD results using the proposed and existing algorithms are discussed.

Databases and Protocol: For PAD experiments, SMAD database prepared by Manjani et al. [25] is used. The database consists of 130 real and silicone mask attack videos. In total, the database consists of a total of $27,897$ frames. For the experiments, the standard protocol as defined in the paper(i.e., 5 folds cross-validation) is used and the performance of presentation attack detection (PAD) is reported using equal error rate (EER).

The database contains both videos and frames. Therefore, the performance of the PAD algorithm is reported with respect to both frame-based (where the results of the classification of an individual frame are reported) and video-based (where the entire video is classified as real or spoof).

Comparative Algorithms: The comparison of the proposed CHIF algorithm is performed with state-of-the-art algorithms developed for silicone mask attack detection, including deep dictionary proposed by Manjani et al. [25]. The existing PAD algorithms used for comparison are described below:

- Deep Dictionary: Manjani et al. [25] have proposed two variants of deep dictionary, namely over complete and under complete dictionary. The basis and coefficients of 3 layers over the complete dictionary are learned in a greedy layer by layer fashion. After learning the dictionary, the linear SVM classifier is trained for two-class classification.

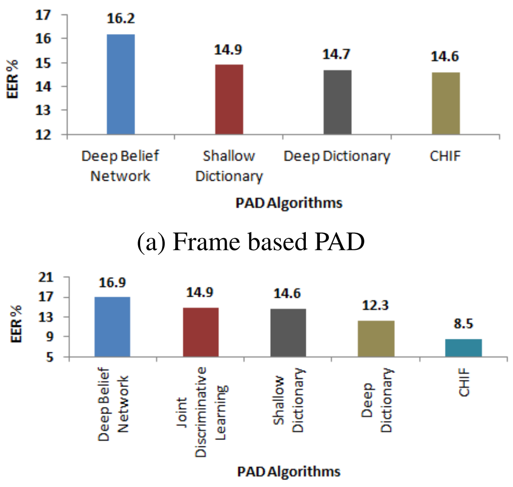
- Deep Belief Network (DBN): It is a generative algorithm which consists of multiple hidden latent variables. Hinton et al. [15] have proposed the greedy layer-wise unsupervised learning of DBN. The results of DBN are taken from Manjani et al. [25].

- Joint Discriminative Learning [31]: In this algorithm, each frame of a video is first passed through a pretrained CNN, and then the subtle motion is calculated using optical flow. Later, the discriminative features are jointly learned along the channel and spatial dimension.

4.1. PAD Results

Comparison of the proposed CHIF and existing PAD algorithms for the frame-based protocol are reported in Figure 4 (a). In the frame-based protocol, the DBN model yields an EER of 16.2%, which is 1.6% higher than the proposed

[1] http://www.vlfeat.org/matconvnet/pretrained/#imagenet-ilsvrc-classification



(a) Frame based PAD

(b) Video based PAD

Figure 4. Comparison of CHIF with SOTA algorithms such as deep dictionary [25], deep belief network (DBN), and and joint disriminative learning [31] developed for frame and video based silicone mask attack detection.

Table 1. Frame based PAD performance in terms of EER ± std (%) (lower the better) of the proposed CHIF features. The comparative results with several histogram feature descriptors are also reported.

Algorithm	Features	SVM
Existing	VGG16	18.74 ±2.47
	LBP	17.84 ±3.40
	BSIF:3 × 3	17.09 ±2.53
	BSIF:5 × 5	17.62 ±4.54
	BSIF:7 × 7	16.06 ±4.42
	BSIF:9 × 9	15.23 ±3.54
	BSIF:11 × 11	16.53 ±4.33
Proposed	CHIF	**14.61 ±3.71**

CHIF based approach. When the single layer (i.e., shallow) and multiple layers (i.e., deep) dictionary is trained for PAD, the EER is at-least 0.1% higher than the proposed PAD. For video-based PAD (Figure 4 (b)), EER of CHIF (i.e., 8.5%) is 3.8% and 6.4% lower than deep dictionary (EER = 12.3%) [25] and recently proposed joint discriminative learning (EER = 14.9%) [31], respectively. The proposed CHIF texture descriptor outperforms existing PAD algorithms which are based on learning deep features either using dictionary learning or CNN features+motion estimation. The PAD performance of other CNN filters such as GoogLeNet and ResNet is at-least 2%–3% lower than VGG-Face filters based CHIF features.

To further justify the advantage of the proposed non-linear filters based texture features, we have performed the comparison with linear filters based texture algorithms. The frame-based PAD results are reported in Table 1. LBP based approach yields 3.23% higher EER than CHIF. The BSIF features are extracted using the filters of multiple sizes ranging from 3 × 3 to 11 × 11. BSIF is a linear filter learning-based algorithm. The EER of best performing BSIF features (i.e., with 9 × 9 filters) is 0.62% higher than CHIF. It shows that the non-linear filters have an advantage in extracting discriminative texture features. For comparison, the final layer features from VGG16 are also extracted and are given to SVM for two-class classification. While the standard deviation of VGG16 based algorithm is the lowest, the average EER is the highest among all the algorithms used. Finally, Figure 5 shows a t-SNE [22] plot in which we compare frames of a video with real faces and frames of a video with silicone mask faces. This plot clearly highlights that CHIF is able to encode the presentation attack variations caused due to silicone mask.

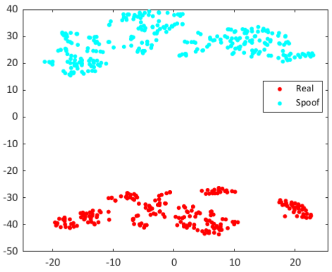


Figure 5. t-SNE plot of CHIF feature obtained from all the frames of one real and spoof video.

5. Conclusion

This paper presents a feature engineering algorithm referred to as *'Convoluted Histogram Image Features (CHIF)'*. The proposed CHIF uses the properties of convolution neural network which learns the discriminative filters through non-linear mapping. These filters are used to compute local image descriptor in the form of histogram image descriptors. These convoluted histogram features are useful in extracting textural information and able to boost the presentation attack detection (PAD) performance. Using silicone mask presentation attack database (SMAD), comparison with existing PAD algorithms utilizing deep features of dictionary learning and CNN on face PAD task shows the strength of the proposed CHIF approach. The proposed method also outperforms existing algorithms which utilize linearly learned filters for texture classification as well as a pre-trained deep CNN model. In future, we plan to extend the proposed algorithm to be robust against adversarial attacks [13] as well as fooling PAD attempts [1, 2].

6. Acknowledgement

A. Agarwal is partly supported by Visvesvaraya PhD Fellowship. M. Vatsa and R. Singh are partly supported from the Infosys Center for AI at IIIT-Delhi. M. Vatsa is also partially supported by the DST, Government of India through the Swarnajayanti Fellowship.

References

[1] A. Agarwal, A. Sehwag, M. Vatsa, and R. Singh. Deceiving face presentation attack detection via image transforms. In *IEEE BigMM*, 2019.

[2] A. Agarwal, A. Sehwag, M. Vatsa, and R. Singh. Deceiving the protector: Fooling face presentation attack detection algorithms. In *IEEE ICB*, 2019.

[3] A. Agarwal, R. Singh, and M. Vatsa. Face anti-spoofing using haralick features. In *IEEE BTAS*, pages 1–6, 2016.

[4] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. Swapped! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE IJCB*, pages 659–665, 2017.

[5] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore. Face presentation attack with latex masks in multispectral videos. In *IEEE CVPRW*, pages 275–283, 2017.

[6] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh. Computationally efficient face spoofing detection with motion magnification. In *IEEE CVPRW*, pages 105–110, 2013.

[7] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. In *IEEE BTAS*, pages 1–7, 2018.

[8] Z. Boulkenafet, J. Komulainen, and A. Hadid. Face spoofing detection using colour texture analysis. *IEEE TIFS*, 11(8):1818–1830, 2016.

[9] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, pages 1–7, 2012.

[10] N. Erdogmus and S. Marcel. Spoofing face recognition with 3d masks. *IEEE TIFS*, 9(7):1084–1097, 2014.

[11] J. Galbally and S. Marcel. Face anti-spoofing based on general image quality assessment. In *ICPR*, pages 1173–1178, 2014.

[12] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[13] G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *IJCV*, 127(6-7):719–742, 2019.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[16] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *ICPR*, pages 1363–1366, 2012.

[17] 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*, 13(10):2639–2652, 2018.

[18] J. Liu and A. Kumar. Detecting presentation attacks from 3d face masks under multispectral imaging. In *IEEE CVPRW*, pages 47–52, 2018.

[19] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao. 3d mask face anti-spoofing with remote photoplethysmography. In *ECCV*, pages 85–100. Springer, 2016.

[20] S.-Q. Liu, X. Lan, and P. C. Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In *ECCV*, pages 558–573, 2018.

[21] S.-Q. Liu, P. C. Yuen, X. Li, and G. Zhao. Recent progress on face presentation attack detection of 3d mask attacks. In *Handbook of Biometric Anti-Spoofing*, pages 229–246. 2019.

[22] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008.

[23] J. Määttä, A. Hadid, and M. Pietikinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, pages 1–7, 2011.

[24] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa. Evading face recognition via partial tampering of faces. In *IEEE CVPRW*, 2019.

[25] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE TIFS*, 12(7):1713–1723, 2017.

[26] S. Mehta, A. Uberoi, A. Agarwal, M. Vatsa, and R. Singh. Crafting a panoptic face presentation attack detector. In *IEEE ICB*, 2019.

[27] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. In *ECCV*, pages 404–420. Springer, 2000.

[28] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE T-PAMI*, 24(7):971–987, 2002.

[29] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *IEEE ICCV*, pages 1–8, 2007.

[30] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, pages 41.1–41.12, 2015.

[31] R. Shao, X. Lan, and P. C. Yuen. Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing. *IEEE TIFS*, 14(4):923–938, 2019.

[32] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Face anti-spoofing with multifeature videolet aggregation. In *ICPR*, pages 1035–1040, 2016.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE CVPR*, pages 1–9, 2015.

[35] D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE TIFS*, 10(4):746–761, 2015.

[36] X. Li, J. Komulainen, G. Zhao, Pong-Chi Yuen, and M. Pietikinen. Generalized face anti-spoofing by detecting pulse from face videos. In *ICPR*, pages 4244–4249, 2016.

[37] Z. Xu, S. Li, and W. Deng. Learning temporal features using lstm-cnn architecture for face anti-spoofing. In *IAPR ACPR*, pages 141–145, 2015.

[38] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014.

[39] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.