UNSEEN FACE PRESENTATION ATTACK DETECTION WITH HYPERSPHERE LOSS

Preprint · January 2020

CITATIONS
0

READS
314

4 authors, including:

Haoliang Li
Nanyang Technological University
43 PUBLICATIONS **517** CITATIONS

SEE PROFILE

UNSEEN FACE PRESENTATION ATTACK DETECTION WITH HYPERSPHERE LOSS

Zhi Li *Haoliang Li* *Kwok-Yan Lam* *Alex Chichung Kot*

Nanyang Technological University, Singapore

ABSTRACT

Presentation attack is one of the main threats to face verification systems and attracts great attention of research community. Recent methods achieve great success in intra-database test. However, the problem is more complex in practical scenario as the type of attack could be unseen to system designers. In this paper, we formulate the face presentation attack detection task under an open-set setting and address with our proposed deep anomaly detection based method. The training process is end-to-end supervised by a novel hypersphere loss function and the decision making is directly based on the learned feature representation. We conduct extensive experiments on multiple prevailing databases and evaluate our implemented models by using various metrics. The results show our proposed method is effective against unseen types of attacks and superior to latest state-of-the-art.

Index Terms— unseen presentation attack detection, face anti-spoofing, anomaly detection, deep metric learning

1. INTRODUCTION

Face verification is one of the most prevailing authentication methods because of its high accuracy and convenience compared to other biometric based methods. It has successfully been applied to door access control, border control, and mobile device authentication areas.

Despite its ease of use, face image as a biometric modality has higher disclosure risk compared with fingerprint and iris image. Especially in the age of social media, attackers can easily access face images or videos of genuine users and generate spoofing artifacts such as printed photos, replayed videos and masks made of various materials. These artifacts could thereby being utilized to conduct presentation attacks (PAs) on face verification based systems.

Presentation attack, a.k.a. spoofing attack, is the main threats to face verification systems which hinders their application in scenarios where high-level security is required.

Most of previous work formulate face presentation attack detection (PAD) task as a close-set classification problem, which assumes all types of attack samples are seen at the training stage, and mark off genuine and spoofing samples from binary classification way. According to the discrepancy between genuine and known types of spoof samples, early research has designed various handcrafted features based on color textures [1], image quality [2] and distortion [3], with which they leverage binary classifiers for classification. After great success in other classification tasks, CNN has also been applied in face PAD task for feature extraction [4]. While achieving considerable success on intra-database test of multiple benchmark face PAD databases, performance of binary supervised methods degrades severely in cross-database test because of the domain shift. [5] makes use of depth map and rPPG signal as more robust cues for auxiliary supervision in training of CNN model. Despite of effectiveness against static printed photo and replayed video attacks, depth estimation is no longer in force under mask or make-up attacks. In addition, estimated rPPG signal is not robust to shaking. Recent work [6, 7] incorporate domain adaptation techniques in order to improve generalization ability of face PAD methods, which alleviates the over-fitting problem to some extent at the limitation of using some target domain samples. [8] incorporates client identity information to develop client-specific detector to improve the accuracy of individual as much as possible.

Almost all of the above research work is based on the assumption that we have sufficient knowledge about spoof artifacts and enough data samples could be used for model training. However, the practical scenario is much tougher and more complex. One of the realistic issues is that the attackers may employ new types of attack which we have never seen before. Hereby, there is necessity to analyse the performance of face PAD methods against unseen types of attacks. [9] is the first work that proposed to evaluate effectiveness of PAD methods against unseen types of attacks. Combined with different handcrafted features, outlier detection based methods are compared with conventional binary classification methods. Soon afterward, concatenated image quality feature [10] and color texture [11] are used as feature representation to compare the performance of different classifiers. [12] investigated the performance of client-specific solutions. To address insufficient diversity of spoof attack samples and limitation of handcrafted features, [13] collected a database containing 13 types of attack samples and proposed a method based on deep tree network (DTN) for hierarchical feature learning and classification which achieved state-of-the-art.

In this paper, we address face PAD task under open-set classification setting and aim to provide an effective method against unseen types of attacks. Our contribution in this work

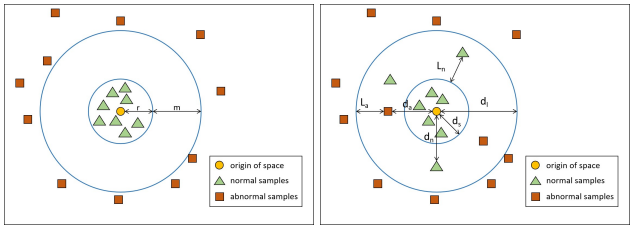
can be summarized as below:

- We propose a novel CNN-based method for face anti-spoofing against unseen type of attacks, which detects attacks directly on learned feature space with no need for additional classifiers to be trained.
- We design a novel hypersphere loss function for deep anomaly detection.
- We conduct extensive experiments to evaluate the performance of our proposed method. The results show our method outperforms latest state-of-the-art methods.

2. PROPOSED METHOD

Before introducing our proposed method, we firstly elaborate the task we work on. The problem that we aim to address is face PAD under open-set setting. Specifically, our task is to design a generalized face PAD method by using available genuine face and known types of attack samples. The method is expected to be applicable to unknown types of attacks.

According to the description above, we believe the problem is more appropriate to be treated as a supervised anomaly detection or outlier detection problem rather than the conventional binary classification problem since the available information about genuine faces and attacks are not asymmetric. Firstly, compared with genuine face samples, attacks are more variable. The shift between genuine face samples is mainly caused by the changes of illumination and resolution of face area. Besides, attack samples are diverse in forms and spoof media. This results in the situation that genuine face samples should naturally form a group and keep intra-class compactness which is not applicable to attack samples. Secondly, the task is under zero-shot classification setting due to the unknown or uncontrollability of attacks. Known attack samples are not as reliable as genuine face samples, and can only be used as reference or estimation of real attacks. Unbiased binary classification may result in over-fitting problem on training samples.



(a) Target Feature Space (b) Illustration of Distance and Loss

Fig. 1. Illustrations on Feature Space

2.1. Hypersphere Loss Function

With the aforementioned analysis, we start to design a distance-based loss function for supervised anomaly detection problem by describing the target feature space. Since the method is proposed from anomaly detection perspective, hereinafter, we use normal and abnormal samples to refer to genuine face and attack samples.

As is shown in Fig.1 (a), normal samples should be distributed near origin and converge to a hypersphere of radius r to maintain intra-class compactness. On the contrary, abnormal data samples are expected to be away from the smaller hypersphere by a predefined margin m to ensure inter-class separation between the normal and abnormal samples. In Fig.1 (b), loss for the jth normal sample L_n^j and abnormal samples L_a^j are computed in different ways. If we have N_n normal samples and N_a abnormal samples. Loss can be represented below.

$$L = \sum_{j=1}^{N_n} L_n^j + \sum_{j=1}^{N_a} L_a^j \tag{1}$$

As L is positive relevant to the number of data samples, if the amounts of samples for different types are not comparable, learned feature space will be biased. To avoid such kind of cases, we use $\frac{1}{N_n}$ and $\frac{1}{N_a}$ to eliminate the dependence on the amount of data samples. In addition, α_n and α_a are parameters to control the weights of loss contributed by samples of different types.

$$L = \frac{\alpha_n}{N_n} \sum_{j=1}^{N_n} L_n^j + \frac{\alpha_a}{N_a} \sum_{j=1}^{N_a} L_a^j \tag{2}$$

As shown in Fig. 1 (b), we represent loss by using the distance between data sample and its expected position and use the square of $L2$-norm for distance calculation.

$$L_n^j = \max(d_n^j - d_s, 0) \tag{3}$$

$$L_a^j = \max(d_l - d_a^j, 0) \tag{4}$$

where $d_n^j = \|f_n^j\|_2^2$, $d_a^j = \|f_a^j\|_2^2$, $d_s = r^2$, $d_l = (r+m)^2$, f_x^y is feature vector of the yth data sample of type x.

By substituting Eq. (3) and Eq. (4) into Eq. (2), we get the final hypersphere loss function below.

$$L = \frac{\alpha_n}{N_n} \sum_{j=1}^{N_n} \max(\|f_n^j\|_2^2 - r^2, 0) + \frac{\alpha_a}{N_a} \sum_{j=1}^{N_a} \max((r+m)^2 - \|f_a^j\|_2^2, 0) \tag{5}$$

Considering the inherent characteristics of anomaly detection task, the loss is calculated based on the distance from origin rather than the distance between two data samples like in typical triplet loss [14] to avoid additional triplet generation or hard triplet mining. Moreover, with our proposed loss function, the training process is supervised by a predefined

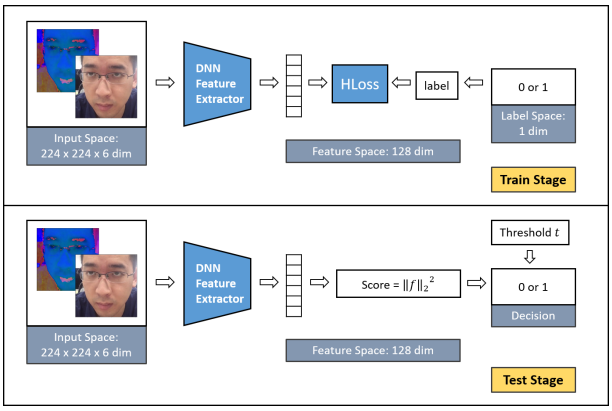


Fig. 2. Illustration on Architecture of Proposed Method

and static learning objective. While ensuring the stable convergence of the training process, we make sure the learned feature representation is calibrated with origin so that it can be directly used for decision making without additional classifiers to be trained.

2.2. Architecture of the Proposed Method

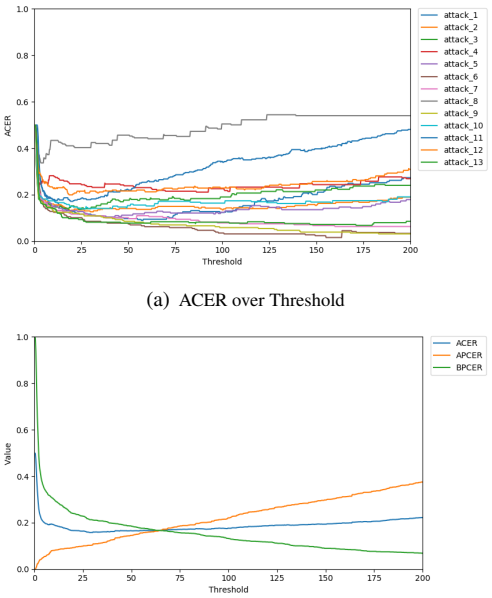
As shown in Fig. 2, we detect and crop face area of raw images in both RGB and HSV colorspace following [14] and concatenate the normalized images into a image cube sized of 224x224x6 as representation of data samples in input space. A DNN architecture is used as feature extractor to transform data samples to a 128 dimensional feature space. The target as to the training stage is to optimize the parameters of the feature extractor in order to find an optimal feature space for attacks detection. The training process is end-to-end supervised by hypersphere loss. At the test stage, after feature extraction of the trained DNN, the square of $L2$-norm of feature vector is compared with predefined threshold for decision making. Since the decision is determined on feature space directly, there is no need for additional classifiers to be trained.

3. EXPERIMENTAL RESULTS

3.1. Experimental Setup

3.1.1. Databases and Protocols

To evaluate our proposed method, we did extensive experiments on multiple databases: CASIA-FASD [15], Replay-Attack [16], MSU-MFSD [3], ROSE-Youtu [6] and SiW-M [13]. Our experiments followed the leave-one-out protocol adopted in [9, 13] where exclude 1 type of attack samples for test and all remaining types of attack samples along with genuine type samples are used for training.



(a) ACER over Threshold

(b) Average APCER, BPCER and ACER over Threshold

Fig. 3. Performance of Our Method Tested on SiW-M

3.1.2. Evaluation Metrics

For fair comparison with state-of-the-art methods, we report the results by using the same metrics as previous work. Area under curve (AUC) is adopted as evaluation metric on CASIA-FASD, Replay-Attack, MSU-MFSD and ROSE-Youtu to report the overall performance of models under varied threshold. On SiW-M database, in addition to EER, we evaluate models by ACER, which is the average of APCER and BPCER, at a fixed threshold for all sub-experiments following [13].

3.1.3. Implementation Details

Since our method doesn't use sequential cues, we uniformly sample 30 frames of images from each video clip, which is followed by face detection and cropping. Cropped face area is normalized to standard sized (224 x 224) images in both RGB and HSV colorspaces. We select ResNet18 based CNN for feature representation to transfer data samples into 128-dimensional feature space and the square of $L2$-norm of feature vector is used as anomaly score compared with predefined threshold. The proposed method is implemented in PyTorch and trained with fixed learning rate 10^{-6} and batch size 30. We set α_n and α_a in Eq. (5) as 1 and 1, respectively. Parameters r and m which control compactness and separation are set to 2 and 10, respectively. Threshold for decision making is fixed to 28 for all experiments on SiW-M.

Table 1. AUC(%) of Models Tested on CASIA-FASD, Replay-Attack and MSU-MFSD Database

Method	CASIA-FASD			Replay-Attack			MSU-MFSD			Overall
	Video	Cut Photo	Warped Photo	Video	Digital Photo	Printed Photo	Printed Photo	HR Video	Mobile Video	
SVM_{OC} + BSIF	70.7	60.7	95.9	84.3	88.1	73.7	64.8	87.4	74.7	78.7±11.7
SVM_{RBF} + LBP	91.5	91.7	84.5	99.1	98.2	87.3	47.7	99.5	97.6	88.6±16.3
NN + LBP	**94.2**	88.4	79.9	99.8	95.2	78.9	50.6	99.9	93.5	86.7±15.6
Deep Tree Network	90.0	97.3	97.5	99.9	99.9	**99.6**	**81.6**	99.9	97.5	95.9±6.2
Ours	92.7	**97.5**	**98.0**	**99.9**	**99.9**	98.7	80.2	**99.9**	**99.3**	**96.2±6.1**

Table 2. Experimental Results Tested on SiW-M

Metric (%)	Method	Attack Type													Overall
				Mask Attacks					Makeup Attacks			Partial Attacks			
		Replay	Print	Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Imperson.	Cosmetic	Funny Eye	Paper Glasses	Partial Paper	
		1	2	3	4	5	6	7	8	9	10	11	12	13	
ACER	SVM_{RBF} + LBP	20.6	18.4	31.3	21.4	45.5	11.6	13.8	59.3	23.9	16.7	35.9	39.2	11.7	26.9±14.5
	Auxiliary	16.8	6.9	19.3	**14.9**	52.1	8.0	12.8	55.8	13.7	**11.7**	49.0	40.5	**5.3**	23.6±18.5
	Deep Tree Network	**9.8**	**6.0**	15.0	18.7	36.0	**4.5**	**7.7**	48.1	11.4	14.2	19.3	**19.8**	8.5	16.8±11.1
	Ours	11.6	13.0	**13.9**	23.5	**12.0**	8.6	11.2	**40.3**	**10.9**	13.1	**17.9**	20.8	8.2	**15.8±8.6**
EER	SVM_{RBF} + LBP	20.8	18.6	36.3	21.4	37.2	7.5	14.1	51.2	19.8	16.1	34.4	33.0	7.9	24.5±12.9
	Auxiliary	14.0	4.3	**11.6**	**12.4**	24.6	7.8	10.0	72.3	10.1	**9.4**	21.4	**18.6**	**4.0**	17.0±17.7
	Deep Tree Network	**10.0**	**2.1**	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	**19.8**	20.5	8.8	16.1±12.2
	Ours	13.2	14.0	18.1	24.0	**12.4**	**3.1**	**6.2**	**34.8**	**3.1**	16.3	21.4	21.7	9.3	**15.2±9.0**

Table 3. AUC(%) of Models Tested on ROSE-Youtu

Method	Attack Type						Overall
	Photo				Video		
	Print	Crop.	Eyes.	Half	Leno.	Mac	
SVM_{RBF} + LBP	67.5	**96.8**	97.3	89.6	94.6	67.7	85.6±14.2
Ours	**80.5**	95.6	**98.9**	**96.5**	**95.3**	**80.7**	**91.3±8.3**

3.2. Results and Analysis

3.2.1. Experiment on CASIA-FASD, Replay-Attack and MSU-MFSD

Following the evaluation protocol 1 in [9], we train and test our models under the same experimental setup as previous work. We compare the performance of our method against replayed video and printed photo attacks with 4 state-of-the-art methods proposed in [9, 17, 11, 13]. As shown in Tab.1, our proposed method achieves the best performance on $6/9$ sub-experiments and the overall performance outperforms the most recent deep learning-based method on average value of AUC and standard deviation.

3.2.2. Experiment on SiW-M

As stated in [13], the variety of attack types is limited in CASIA-FASD, Replay-Attack and MSU-MFSD databases, where only replayed videos and printed photos are included. To verify the effectiveness against novel types of attacks, we test our models on recently released SiW-M database which contains 13 types of attack samples. We report the performance by ACER, EER and compare with 3 methods proposed in [9, 5, 13] as shown in Tab.2. Our proposed method achieves best overall performance on ACER and EER. Specifically, it outperforms previous state-of-the-art by 6%, 6% on average values and reduces standard deviation by 23%, 26% respectively. Smaller standard deviations show our method is more generalized to different types of attacks. Especially, our method is more effective on transparent and obfuscation attacks, on which previous methods perform poor. Fig.3 shows performance of our proposed models over threshold which can be adjusted according to security level of different application scenarios.

3.2.3. Experiment on ROSE-Youtu

As is known, even for common video and photo forms of attacks, they may be conducted by using different spoof media. The variation of spoof media also causes domain shift is proven in conventional face PAD research. It's improper to treat such kind of attacks as one known type although they are in the same form. Therefore, we also test on ROSE-Youtu database, which contains 4 types of printed photos and videos replayed on 2 types of screens. Performance of our method is compared with SVM_{RBF} + LBP [17] in Tab.3. The result shows our proposed method outperforms baseline method by 7% on average value of AUC. In addition, the smaller standard deviation shows that our method has better generalization ability against different types of attacks.

4. CONCLUSION

In this paper, we design and implement a novel method to detect face presentation attacks of unknown type. We train a deep CNN-based feature extractor with our proposed hypersphere loss and detect attack samples directly on learned feature space without additional classifiers to be trained. Extensive experiments on prevailing databases by multiple evaluation metrics show the effectiveness and superiority of our proposed method compared with recent state-of-the-art.

5. REFERENCES

[1] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE International Conference on Image Processing (ICIP)*, Sep. 2015, pp. 2636–2640.

[2] 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*, vol. 23, no. 2, pp. 710–724, Feb 2014.

[3] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, April 2015.

[4] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014.

[5] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot, "Learning generalized deep feature representation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 10, pp. 2639–2652, Oct 2018.

[7] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, July 2018.

[8] I. Chingovska and A. R. dos Anjos, "On the use of client identity information for face antispoofing," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 787–796, April 2015.

[9] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.

[10] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *2018 International Conference on Biometrics (ICB)*, Feb 2018, pp. 75–81.

[11] F. Xiong and W. AbdAlmageed, "Unknown presentation attack detection with face rgb images," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Oct 2018, pp. 1–9.

[12] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Spoofing attack detection by anomaly detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 8464–8468.

[13] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition*, Cham, 2015, pp. 84–92, Springer International Publishing.

[15] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*, March 2012, pp. 26–31.

[16] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*, Sep. 2012, pp. 1–7.

[17] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 612–618.