얼굴 스푸핑 방지를 위한 다중 양식에 관한 연구

오신모, 이효종
전북대학교 컴퓨터공학부
chenmou@jbnu.ac.kr, hlee@jbnu.ac.kr

A Study on Multiple Modalities for Face Anti-Spoofing

Chenmou Wu, Hyo Jong Lee*
Division of Computer Science and Engineering, Jeonbuk National University
*Corresponding author

Abstract

Face anti-spoofing (FAS) techniques play a significant role in the defense of facial recognition systems against spoofing attacks. Existing FAS methods achieve the great performance depending on annotated additional modalities. However, labeling these high-cost modalities need a lot of manpower, device resources and time. In this work, we proposed to use self-transforming modalities instead the annotated modalities. Three different modalities based on frequency domain and temporal domain are applied and analyzed. Intuitive visualization analysis shows the advantages of each modality. Comprehensive experiments in both the CNN-based and transformer-based architecture with various modalities combination demonstrate that self-transforming modalities improve the vanilla network a lot. The codes are available at https://github.com/chenmou0410/FAS-Challenge2021.

1. Introduction

As face recognition technology advances, more interactive intelligent applications are embracing this efficient and convenient payment. However, most existing facial recognition systems are at risk of presenting attacks, particularly in 3D mask attacks [1]. Therefore, advanced presentation attack detection (PAD) is imminent for the existing facial recognition system. Recently, both traditional image processing algorithms and deep learning (DL) methods have achieved major improvements in face anti-spoofing (FAS).

Similar to most pattern recognition issues, traditional algorithms adopted handcrafted features extraction cascaded with a trained classifier (SVM [2], Bayesian Classifiers [3]) for FAS problems. To be specific, the liveness-related features are firstly extracted by the classical handcrafted descriptors (*i.e.*, LBP [4], SIFT [5], HOG [6]) based on the prior knowledge, then a powerful feature is fused by feature cross methods and send into the classifier. However, the synthetic feature is still too shallow to provide efficient and robust information for unseen scenarios and unknown PAs.

Benefitting from several large datasets, the data driven method shines on the FAS. The convolutional neural network, as a most commonly used backbone, treats the FAS problem as a binary classification problem. The CNNs extract deeper texture features to distinguish the live and fake face supervised by a simple binary cross-entropy loss. Limited by the simple binary loss and the features only from the spatial domain (texture, edges, and corners), the network probabilistically leads to learning arbitrary pattern instead of anti-spoofing patterns and overfitting. To address this issue, several FAS methods add additional modalities (e.g., depth information, Infrared information, 3D information etc.) as auxiliary supervised label to improve the performance. Although these additional-modality-based approaches achieve greater performance in most presentation attacks, higher cost caused by a large number of additional-modality annotations should be considered.

Following the aforementioned discussion, we proposed to use self-transforming modalities rather than high-cost modalities to improve network performance. In this work, we first investigate the effectiveness of different self-transforming modalities including frequency-based and temporal-based modality. All these additional modalities are transformed from input RGB image or frames, which avoids additional annotations. In addition, we perform multiple modalities on the backbones of two concepts (Transformer [7] and CNN [8]), which aims to find the optimal architecture for modalities fusion based state of the arts.

2. Related Work

Previous FAS methods are mainly spitted in two categories. On the one hand, the detection of specific facial motions pattern is widely used to estimate PAD in early approaches [9]–[11], which is difficult to counter novel types of attacks. (*i.e.*, video replay). On the other hand, previous research captures the spoofing patterns based on handcrafted features and train a binary classifier to discriminate spoofing and living face [4]–[6].

Recently, data driven methods achieve great success on both frame level and video level face anti-spoofing. For the frame-based methods [12-15], the CNN models are pre-trained on a large face dataset first and fine-tuned on the FAS dataset in a binary-classification setting. In contrast, auxiliary modality supervised FAS methods [16-18] are more effective, which captures the more representative spoofing pattern with the support of supplementary information. For the video-based methods are proposed to extract the temporal [19] or rPPG [20] as an auxiliary modality fusing with the RGB features for PAD.

3. Modalities Analysis

In this section, we first introduce three different modalities from temporal and frequency domain. Then, we do a primary analysis with visualization result.

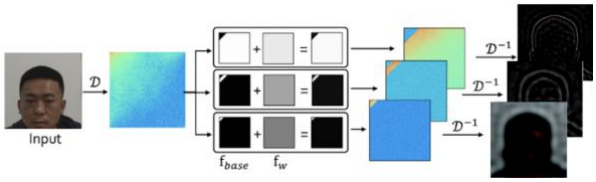
3.1 Temporal modality

Dynamic texture [21] and motion blurriness [22] cloud be useful for the FAS issue given that the significant temporal discrepancy has existed between the live faces and PAs. Instead of the optical flow, we adopt the Rank Pooling [23] based dynamic image to capture temporal cues as a temporal modality, which has been proved its superiority to regular optical flow [24].

Rank pooling first encodes consecutive frames K into feature vector as a rank group and use RankSVR [25] to learn the temporal discrepancy. The core function is defined below:

$$\underset{E}{argmin}\frac{1}{2}\|E(K)\|^2+\frac{2}{K(K-1)}\times\sum_{i>j}\theta_{ij} \qquad (1)$$
$$s.t.\ f^T\cdot\left(S_i-S_j\right)\geqslant 1-\theta_{ij},\quad \theta_{ij}\geqslant 0,$$

where E is a encoder that embeds consecutive K-frames into a laten vector f and θ_{ij} denotes a slack variable. In this work, we directly apply rank pooling on the source RGB frames to compute temporal modality online using K-frames (K=10 in this situation). Thus, temporal modality has the same size as input RGB frame.



(Figure 1) Illustration of the Frequency-aware Decomposition (FAD) to searching out salient frequency components. D denotes processing Discrete Cosine Transform (DCT). D^{-1} denotes processing Inversed Discrete Cosine Transform (IDCT).

3.2 Frequency Modality

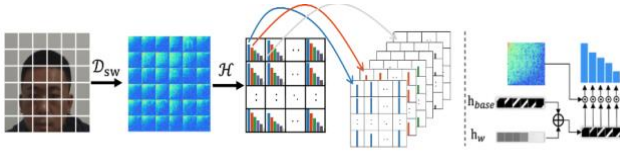
Frequency provides a complementary viewpoint where some subtle spoofing pattern could be well described. In this section, anti-spoofing clues are searched from two views of the frequency domain. Following the [26], the frequency-aware decomposition (FAD) and local frequency statistics (LFS) is applied on the input RGB image to generate frequency modalities.

FAD aims at learning the PAs patterns by frequency image decomposition. As shown in Figure 1, we first apply the Discrete Cosine Transform (DCT) [27] to obtain the frequency image $DCT(X)\in R^{H\times W\times 3}$. Then, three binary masks f_{base} are created manually to separate the frequency domain into low, middle, and high frequency bands. More specifically, the lower band covers the first 1 / 16 of the entire range, the middle band lies between 1 / 16 and 1 / 8 of the range, and the high band corresponds to the last 7 / 8 of the range. Finally, the decomposed image components F_i are calculated by

$$F_i=DCT^{-1}\{DCT(x)\odot\left[f^i_{base}+\sigma\left(f^i_w\right)\right]\},\ i=1,2,3$$
$$\sigma(z)=\frac{1-e^{-z}}{1+e^{-z}} \qquad (2)$$

where f^i_w are learnable filters and $\odot$ denotes element-wise product.

LFS aims at capturing feature that include shift invariance and local consistency of RGB image on the basis of explicitly frequency statistics. Given an RGB image, the localized frequency representations are first extracted by Sliding Window DCT (SWDCT), and then average frequency representations are computed with a series of frequency bands.



(Figure 2) Illustration of the Local Frequency Statistics (LFS). SWDCT denotes applying Sliding Window Discrete Cosine Transform and H denotes gathering statistics on each grid adaptively. $\oplus$ indicates element-wise addition and $\odot$ indicates element-wise multiplication.

The whole process is shown in Figure 2. To capture detailed abnormal frequency distributions, LFS rearrange the frequency statistics in a multi-channel spatial map which shares the same distribution as the input image. The local statistics is collected from each frequency band which is each patch $p_i\in SWDCT(x)$ in Figure 2, and the statistics of each band are formulated by

$$y_i=\log_{10}\left\|p\odot\left[h^i_{base}+\sigma\left(h^i_w\right)\right]\right\|_1,\ i=\{1,2,\ldots,M\} \qquad (3)$$

The h^i_{base} is the base filter and h^i_w is the learnable filter. The local frequency statistics y_i of each patch is transposed as a $1\times 1\times M$ vector, which M is output channels.

4. Experiments

Evaluation Metrics. For a fair comparison, we calculate the Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER) and ACER followed the standard protocols and metrics provided by [28]. Furthermore, we also evaluate on accuracy, TPR and FPR three primary metrics. All metrics are formulated as follow:

$$TPR=TP/(TP+FN)$$
$$FPR=FP/(FP+TN)$$
$$APCER=FP/(TN+FP) \qquad (4)$$
$$BPCER=FN/(FN+TP)$$
$$ACER=(APCER+BPCER)/2.0$$

Dataset. We use HiFi-3D mask datasets [28] which includes 6 different attack type under different color and brightness. For both training set and test set, we adopt MTCNN [29] to detect face location to crop face into 256 x 256 pixel and normalized pixel value to [-1, 1]. Moreover, data augmentation strategies, random horizontal flip, color jitter on brightness, contrast, and saturation are specially utilized for training data.

Network detail. For the transformer-based architecture, we adopt Cvt-13 [7] that use sperate convolution instead of linear in vision in transformer. For the CNN-based architecture, we adopted SE-ResNet-18 architecture used in the ArcFace [8]. Both networks are initialized by Xavier initialization and we don't apply the pre-trained parameters on them. For the multiple modalities input, we directly connect channels since all modalities size is same and value lie in [-1,1].

Implementation detail. We implemented the networks

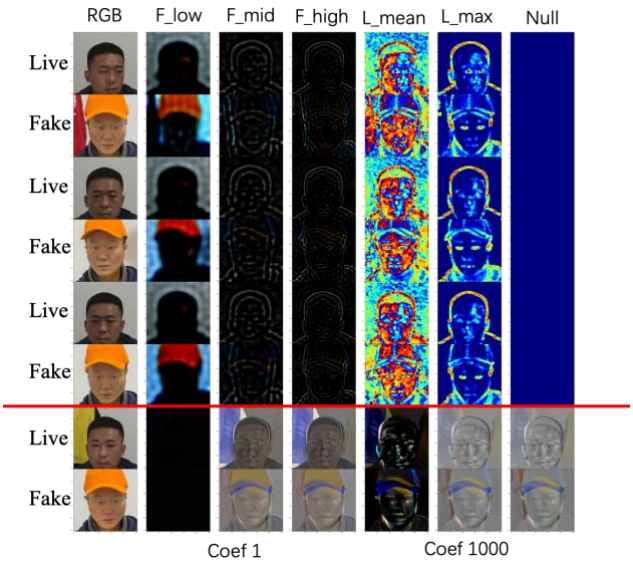
<Table 1> Comparison of Transformer-based network on the 3D-HiFI mask dataset

Modality combination	ACC (%)	APCER	BPCER	ACER	TPR	FPR
RGB	85.77	0.1170	0.1693	0.1431	0.8307	0.1170
YUV	84.37	0.1273	0.1866	0.1569	0.8134	0.1273
RGB & Temporal	83.04	0.1523	0.1890	0.1707	0.8110	0.1523
RGB & FAD	87.79	0.1205	0.1240	0.1222	0.8760	0.1205
RGB & LFS	88.72	0.0829	0.1437	0.1133	0.8569	0.0829
RGB & 2FM	88.97	0.1002	0.1231	0.1116	0.8769	0.1002

<Table 2> Comparison of CNN-based network on the 3D-HiFI mask dataset

Modality combination	ACC (%)	APCER	BPCER	ACER	TPR	FPR
RGB	86.12	0.1090	0.1599	0.1344	0.8507	0.1170
YUV	83.33	0.1384	0.1872	0.1628	0.8112	0.1273
RGB & Temporal	84.51	0.1453	0.1901	0.1677	0.8221	0.1523
RGB & FAD	87.65	0.1257	0.1293	0.1275	0.8690	0.1305
RGB & LFS	88.91	0.0783	0.1334	0.1058	0.8551	0.0769
RGB & 2FM	89.22	0.0910	0.1131	0.1020	0.8834	0.0902

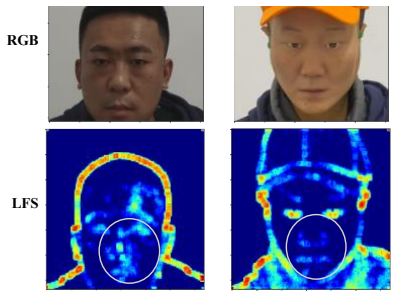
with Pytorch on 4 Nvidia 2080Ti GPUs. For the Cvt, the initial learning rate and weight decay are 3e-4 and 5e-4, respectively. The maximum epochs are 30 and learning rate is decayed every 5 epochs. For the SE-ResNet, the initial learning rate and weight decay are 1e-4 and 5e-4, respectively. The maximum epochs are 60 and learning rate is decayed every 10 epochs. Both two models are trained with Adam optimizer and batch size is 48.



(Figure 3) Visualization of three modalities. We visualize modalities on live face and 3D mask with a hat sample. F-low, F-mid and F-high are the three frequency bands of the FAD modality. L-mean and L-max mean the average and maximum of the channel M. The temporal modality is below the red line. The confidence of SVR is set 1and 1000.

Modality Analysis based on visualization. The temporal modality, FAD modality and LFS modality is shown in figure 3. We found FAD modality work well on the extra decoration on the face and temporal modality can capture eye and facial motion. Only by visualization, the difference between spoofing face and live face is the most obvious of the LFS modality. As shown in Figure 4, the frequency modality based on the LFS can clearly show the difference between the liveness and fakeness in texture, where forage face has abnormal smooth texture which is not exist on the live face.

Impact of Multiple Modality for Deep Models. Here we fully evaluate the performance of two models (SE-ResNet-18 and Cvt-13) with various modality combination, which are shown in Table I and Table II. Only the accuracy and TPR are higher better, and the other metrics is lower better. We first evaluate on different color space, which RGB achieve better performance on both CNN and Transformer based network. Same as the visualization result, LFS modality performs better than the FAD in frequency domain. FAD and LFS modality used together to obtain the best results.



(Figure 4) Detail visualization of LFS modalities.

5. Conclusion

This paper verifies the importance of considering various modalities on deep-based face anti-spoofing. We utilize the FAD method and LFS method on frequency domain and rank pooling on temporal domain to generate three modalities for FAS. Based on the visualization, we analyze the advantage of three modalities. Finally, we demonstrate the superiority of our approach experimentally.

Acknowledgment

This research was supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Education (GR 2019R1D1A3A03103736).

Reference

[1] S. Jia, G. Guo, and Z. Xu, “A survey on 3D mask

presentation attack detection and countermeasures," Pattern Recognit., vol. 98, Feb. 2020, Art. no. 107032.
[2] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in Proc. IEEE Int. Conf. Image Process., Quebec City, QC, Canada, 2015, pp. 2636–2640.
[3] X. Cao, D. Wipf, F. Wen, G. Duan and J. Sun, "A practical transfer learning algorithm for face verification", Proc. Int. Conf. Comput. Vis., pp. 3208-3215, Dec. 2013.
[4] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "LBP Top based countermeasure against face spoofing attacks," in Proc. Asian Conf. Comput. Vis., 2012, pp. 121–132.
[5] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," IEEE Trans. Inf. Forensics Security, vol. 11, pp. 2268–2283, 2016.
[6] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face antispoofing," in Proc. IEEE Biometrics Theory Appl. Syst., Arlington, VA, USA, 2013, pp. 1–8.
[7] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan and L. Zhang, "Cvt: Introducing convolutions to vision transformers," arXiv preprint arXiv:2103.15808, 2021.
[8] J. Deng, J. Guo, J. Yang, N. Xue, I. Cotsia and S. P. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
[9] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in 'liveness' assessment," IEEE Trans. Inf. Forensics Security, vol. 2, no. 3, pp. 548–558, Aug. 2007.
[10] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, Computationally efficient face spoofing detection with motion magnification," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Jun. 2013, pp. 105–110.
[11] W. Kim, S. Suh, and J.-J. Han, "Face liveness detection from a single image via diffusion speed model," IEEE Trans. Image Process., vol. 24, no. 8, pp. 2456–2465, Aug. 2015.
[12] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In IPTA, pages 1–6, 2016.
[13] Keyurkumar Patel, Hu Han, and Anil K Jain. Cross-database face anti-spoofing with robust feature representation. In Chinese Conference on Biometric Recognition, pages 611–619, 2016.
[14] Anjith George and S´ebastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In International Conference on Biometrics, number CONF, 2019.
[15] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face despoofing: Anti-spoofing via noise modeling. In Proceedings of the European Conference on Computer Vision (ECCV), pages 290–306, 2018.
[16] Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In 2017 IEEE International Joint Conference on Biometrics (IJCB), pages 319–328, 2017.
[17] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 389–398, 2018.
[18] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, Guoying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5295-5305.
[19] A. Yang, P. M. Esperanc¸a, and F. M. Carlucci, "Nas evaluation is frustratingly hard," arXiv preprint arXiv:1912.12522, 2019.
[20] A. Zela, J. Siems, and F. Hutter, "Nas-bench-1shot1: Benchmarking and dissecting one-shot neural architecture search," arXiv preprint arXiv:2001.10422, 2020.
[21] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. Ho, "Detection of face spoofing using visual dynamics," TIFS, vol. 10, no. 4, pp. 762–777, 2015.
[22] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang, and X. Feng, "Replayed video attack detection based on motion blur analysis," TIFS, vol. 14, no. 9, pp. 2246–2261, 2019.
[23] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," TPAMI, vol. 39, no. 4, pp. 773–787, 2016.
[24] J. Wang, A. Cherian, and F. Porikli, "Ordered pooling of optical flow sequences for action recognition," in WACV. IEEE, 2017, pp. 168–176.
[25] A. J. Smola and B. Sch¨olkopf, "A tutorial on support vector regression," Statistics and computing, vol. 14, no. 3, pp. 199–222, 2004.
[26] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In ECCV.
[27] Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE transactions on Computers 100(1), 90{93 (1974)
[28] Ajian Liu, Chenxu Zhao, Zitong Yu, Jun Wan, Anyang Su, Xing Liu, Zichang Tan, Sergio Escalera, Junliang Xing, Yanyan Liang, Guodong Guo, Zhen Lei, Stan Z. Li, Du Zhang, "Contrastive Context-Aware Learning for 3D High-Fidelity Mask Face Presentation Attack Detection", arxiv, 2021.
[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignmentusing multitask cascaded convolutional networks," IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1499–1503, Oct. 2016.