To appear in CCBR, 2016

Cross-Database Face Antispoofing with Robust Feature Representation

Keyurkumar Patel[1], Hu Han[2,⋆], and Anil K. Jain[1]

[1] Department of Computer Science and Engineering,
Michigan State University, East Lansing, MI 48824, USA
[2] Key Lab of Intelligent Information Processing, Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
{patelke6,jain}@msu.edu,hanhu@ict.ac.cn

Abstract. With the wide applications of user authentication based on face recognition, face spoof attacks against face recognition systems are drawing increasing attentions. While emerging approaches of face antispoofing have been reported in recent years, most of them limit to the non-realistic intra-database testing scenarios instead of the cross-database testing scenarios. We propose a robust representation integrating deep texture features and face movement cue like eye-blink as countermeasures for presentation attacks like photos and replays. We learn deep texture features from both aligned facial images and whole frames, and use a frame difference based approach for eye-blink detection. A face video clip is classified as live if it is categorized as live using both cues. Cross-database testing on public-domain face databases shows that the proposed approach significantly outperforms the state-of-the-art.

Keywords: Face liveness detection, cross-database generalizability, deep texture feature, eye-blinking detection

1 Introduction

A number of studies on spoof attacks against face recognition (FR) systems show that state-of-the-art (SOTA) FR systems are still vulnerable [1,2,3,4]. Additionally, most FR systems are designed to represent the individual face images of the same subject in a way that minimizes the intra-person variations like illumination, pose, and modality [5]. These properties of FR systems, to some degree, reduce their ability to distinguish between live and spoof face images. Such vulnerabilities can lead to financial loss and security breaches when FR alone is used for authentication. As more security systems leverage face for user authentication, an increasing focus is shifting towards developing face antispoofing algorithms to prevent malicious users from gaining access.

An authorized user's face images or videos can be easily obtained covertly by a smartphone camera or from social media, and used to spoof a FR system.

⋆ Corresponding author. E-mail: hanhu@ict.ac.cn (H. Han). H. Han would like to thank NVIDIA for donating a GPU used in the research.

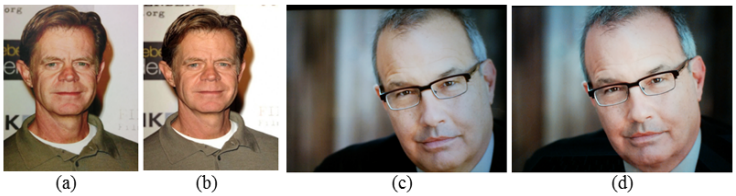

Fig. 1. Visual comparisons between spoof face images (a, c) captured from printed photos and real face images (b, d) indicate the challenges of face antispoofing.

Creating a realistic 3D face mask of an authorized user is also possible using multiple 2D face images. As spoof attacks are diverse in nature, a FR system may require a series of modules focusing on detecting each of the 2D and 3D attacks. In this paper, we focus on 2D attacks by photos and replays as they can be easily launched with low cost. However, even 2D face spoof detection alone is non-trivial. As shown in Fig. 1, even humans will find it difficult to distinguish between live and spoof face images.

A number of solutions to 2D face antispoofing have been proposed, but many of them do not generalize well to unconstrained scenarios. The generalization ability of face antispoofing approaches need to be significantly improved before they can be adopted by operational systems. Humans distinguish between a live face and a face photograph unconsciously by leveraging physical (*e.g.,* 3D shape, texture, and material) and behavioral (spontaneous facial motions like blink, sight, expression, etc.) cues that are generalizable. For example, human can easily distinguish between skin and other materials like paper and digital screen; facial motions are effective for identifying photographic attacks. These observations motivate us to use both physical and behavioral cues in designing our face antispoofing algorithms.

In the proposed approach, we integrate texture feature and eye-blink cue to achieve robust face antispoofing. We learn deep texture features from both aligned face images and whole video frames, because texture distortions in spoof face images exist in both face and non-face regions. We also design a frame difference based approach for efficient eye-blink detection. A face video clip is classified as live if it is categorized as live using both cues. The contributions of this paper include: (i) a novel texture representation using both aligned face images and whole video frames, highlighting general texture differences between live and spoof face images; (ii) a frame difference based approach for efficient eye-blink detection, complementing texture cue with low computational cost; and (iii) significantly improved cross-database testing performance than the SOTAs.

2 Related Work

Face antispoofing methods cover different categories from face motion analysis, texture analysis, image quality analysis to active methods [3,6]. In this paper, we briefly review related work on texture analysis and eye-blink detection.

2.1 Texture Analysis for Antispoofing

Face texture analysis based methods try to capture the texture differences between live and spoof face images from the perspective of surface reflection and material differences [7,8]. These methods can perform spoof detection using a single face image, and thus have relatively fast response. However, these methods may have poor generalizability, particularly under the relatively small training sets of public face spoof databases. Additionally, most of the face texture analysis based methods utilized hand-crafted features such as LBP. Antispoofing methods based on convolutional neural networks (CNNs) are rare, and limited to shallow CNNs [9]. Deep CNN models like CaffeNet [10] and GoogLeNet [11] were used for fingerprint antispoofing [12], but not for face antispoofing where the contactless image acquisition scenario of face is quite different. One important reason is that current face spoof datasets are small compared with the datasets of image classification and face recognition, which may easily lead to overfitting of CNN models.

2.2 Eye-Blink Detection for Antispoofing

Eye-blink is one of the spontaneous facial motions. On average, a human has one blink every 2–4 seconds) [13]. Hence, eye-blink is a strong evidence to differentiate live faces from face photographs, particularly when the input is a video clip. Eye-blink detection is typically formulated as an eye state (open and close) change problem given a video sequence [13,14]. Approaches involved in the above publications include undirected conditional graphical model [13], and conditional random fields (CRF) model [14]. These approaches demonstrated their effectiveness in video based face antispoofing, but both approaches require training, and again their performance in cross-database testing scenarios were not known. In this work, we present a non-learning based method for efficient eye-blink detection. The proposed eye-blink detection and the deep texture features work together to handle input of both video clips and single images.

3 Proposed Approach

The proposed approach consists of three main modules: deep generalized texture feature learning, image difference based eye-blink detection, and fusion strategy.

3.1 Deep Generalized Texture Features

Considering the success of CNNs in feature learning for image classification tasks, we choose to use CNNs such as CaffeNet [10] and GoogLeNet [11]. Although CNNs were used in [9], it is shallow, and thus does not leverage the powerful non-linear learning ability of deep CNNs. Additionally, most of the published methods assume that the input to an antispoofing system is only the facial region. However, in real applications, no matter for a live or spoof face presentation, the

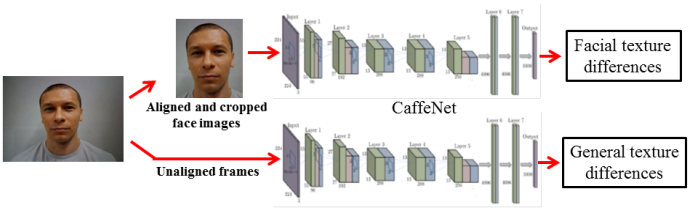

Fig. 2. Deep texture feature learning in the proposed antispoofing approach consists of learning features from both aligned faces and unaligned frames, which highlight facial and generalized image texture distortions in spoof face images, respectively.

input frames captured by the camera also contain non-face regions. Thus, texture distortions of spoof face image exist in the entire frame not only the facial region. Based on such an observation, we propose to learn texture features from both the aligned facial images and the whole frames under a deep CNN framework. This way, we believe the learned features will have better generalization ability. We formulated antispoofing as a live vs. spoof classification problem, and choose softmax loss

$$l(y, f) = -\log\left(\frac{e^{f_y}}{\sum_{j=1}^{m} e^{f_j}}\right), \tag{1}$$

where y is the sample's ground-truth label, and f denotes the output of the last fully-connected (fc) layer. We use the PittPatt SDK (acquired by Google in 2011) to detect the face and eye locations, and faces are aligned based on two-eye locations [15]. Both the aligned faces and whole frames are normalized into 256×256. During the task-specific fine-tuning, the last fc layer is set to have two nodes corresponding to two classes. Figure 2 gives an example using CaffeNet [10] model (with 5 conv layers and 3 fc layers) for feature learning.

Given the small sizes of existing face spoof databases, training such deep CNNs is prone to overfitting. Therefore, we design a generic-to-specific transfer learning scheme for network training. As shown in Fig. 3, the designed generic-to-specific learning first pre-train the network for classification in a related image classification domain. Such a pre-training assists in the network training by providing a reasonable initialization, which is usually better than random initialization. We then fine-tune the network w.r.t. face classification allowing the network to enhance face-specific feature learning, because we believe some face texture patterns could be shared between face classification and face antispoofing. Finally, task-specific fine-tuning with live and spoof face images is performed to build the capability of distinguishing between live and spoof face images. As we stated early, during the task-specific fine-tuning, we used both aligned face images and whole frames allowing feature learning of both face-specific texture features and generalized texture textures. Our experiments verified the effectiveness of such a deep texture feature learning.

3.2 Image Difference Based Eye-blink Detection

An eye-blink activity is determined by the eye states ($\mathbb{S}$) in sequential frames ($\mathbb{F}$) of a video, where $\mathbb{F} = \{f_1, f_2, \cdots, f_n\}$, and $\mathbb{S} = \{s_1, s_2, \cdots, s_n\}$. For simplicity,

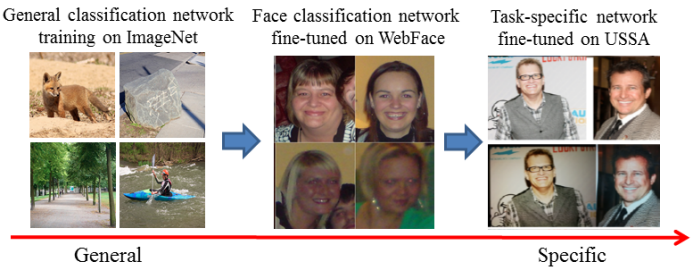

Fig. 3. The general-to-specific deep transfer learning strategy utilizes large databases of image and face classification (*e.g.,* ImageNet and WebFace) and a relatively small face spoof database (USSA).

we define two eye states, *i.e.,* $s_i \in \{o : open; c : close\}$. An eye-blink is reported if there is a state change in $\mathbb{S}$, either from open to close or from close to open. In our approach, we solve the eye-blink detection problem by detecting the state changes (*e.g.,* , $o \rightarrow c$ or $c \rightarrow o$) without explicitly predicting $\hat{\mathbb{S}}$. Formally, we look at the difference image ($\mathbb{D}$) between two image frames, and predict whether an eye state change ($\mathbb{G}$) occurs $\psi : \mathbb{D} \rightarrow \mathbb{G}$, where ψ is a mapping function from $\mathbb{D}$ to $\mathbb{G}$; $\mathbb{D} = \{d_j = (f_{j+1} - f_j)\}_{j=1}^{n-1}$; and $\mathbb{G} = \{g_j \in \{0 : no\ change; 1 : change\}\}_{j=1}^{n-1}$.

We crop the left and right eye regions separately into a 40×50 pixels. We dynamically threshold the periocular region and the pupil of the eyes to get binary eye images, and calculate the eye difference images d_j between successive frames. The following rule is used to determine the eye state changes

$$g_j = \psi(d_j) = \begin{cases} 1 & \text{if } \mathcal{V}(d_j) \geq T \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where $\mathcal{V}(\cdot)$ calculates the percentage of pixels in a difference image d_j; as the difference images d_j is binary, the function $\mathcal{V}(\cdot)$ can simply count the percentage of white pixels; $\psi(\cdot)$ is the threshold mapping function, and the threshold $T = 0.1$ is choose based on our experimental verifications.

Such a simple eye-blink detection method may be affected by illumination, but it utilizes both the spatial information and the contextual information between sequential frames, making it fairly robust. The proposed method is also efficient in calculation ($\sim$100 fps on Windows 7 with Intel Core2 3.0G CPU and 8G RAM). Thus, it is possible to run the algorithm on commodity smartphones.

3.3 Voting Fusion

No matter if a spoof attack is from a photograph or video, the camera captures a video sequence. Therefore, it is reasonable to make a decision per video clip, *e.g.,* 1, 5, or 10 sec. in real applications. Thus, for the deep texture features (face-specific texture feature and generalized texture feature), a voting scheme can be used to generate a decision per video clip. This decision can be then combined with the output of eye-blink detection algorithm to generate a final decision.

Table 1. Comparisons with the STOAs in cross-database testing (HTER in %).

Database	CompRep [6]	JointCT [18]	IDA [3]	VisCod [2]	Proposed
Replay-Attack	29.3	16.7	26.9	34.4	12.4
CASIA FASD	35.4	37.6	43.7	38.5	31.6

We choose to classify a face video clip as live, if both algorithms' outputs are live. Such a fusion scheme may cause a few false rejections of live faces, but our experiments show that the false reject rate is still in a reasonable level while the false accept rate is significantly reduced than the SOTA.

4 Experimental Results

We use three public-domain face spoof databases for evaluations: Idiap Replay-Attack [16], CASIA FASD [17], and MSU USSA [6], and follow their protocols of training, validation, and testing sets.

4.1 Comparisons with the STOA

Most of the earlier publications on face antispoofing did not report performance under cross-database testing. The baselines here are mainly the SOTAs in recent years. **CompRep** [6] analyzed the image distortions in spoof images, and provided a complementary representation of LBP and color moment. **JointCT** [18] extracted a joint color-texture feature from the luminance and the chrominance channels using LBP. **IDA** [3] proposed multiple image quality features for image distortion analysis of spoof face images. **VisCod** [2] performed face spoofing detection through visual codebooks of spectral temporal cubes. **SpoofNet** [9] is an early attempt of using a shallow CNN with three layers for antispoofing of face, fingerprint, and iris, but only *intra-database* performance is reported. So it is used as the baseline of intra-database testing.

USSA dataset is quite new, so most of the baselines did not report cross-database performance on USSA. So for the proposed approach, we follow the settings in [6], and use USSA as the training set for deep texture feature learning. Replay-Attack and FASD are used for testing. We set the base_lr as 0.01, use a polynomial learning rate policy with a power of 0.5. The momentum is set to 0.9, and the weight decay is set as 0.0002.

As shown in Table 1, the proposed approach achieves much lower HTERs than the best of the STOAs on both Replay-Attack and FASD databases. Figure 4 shows some live vs. spoof detection results by the proposed approach. Visualization of the images that have positive responses to the same nodes of the second-last fc layer show that proposed approach learns generalized texture features such as photo-holding, non-natural illumination, and non-natural skin texture (see Fig. 5).

To compared with the intra-database testing performance of SpoofNet [9] on Replay-Attack, we further fine-tune the proposed approach using the training and development sets of Replay-Attack. The proposed approach achieves 0.5% HTER, which is more than 30% HTER reduction compared with SpoofNet [9].

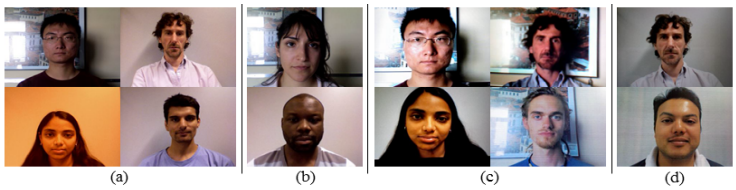

Fig. 4. Face spoof detection results on Replay-Attack: (a) true accept and (b) false reject of live face images, and (c) true reject and (d) false accept of spoof face images.

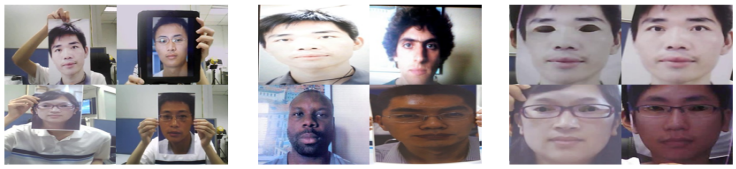

Fig. 5. Visualization of the images that have positive responses to the same nodes of the second-last fc layer indicates the generalized texture feature learning process.

4.2 Evaluations of Individual Modules

Face vs. non-face texture features. The face-specific texture feature alone achieves 17.5% HTER on Replay-Attack, which outperforms most of the SOTA methods except for JointCT [18]. The generalized texture feature learning from the whole frames is new in face antispoofing studies. This feature alone achieves 31.2% HTER on Replay-Attack. Through the generalized texture features seem to be much less discriminative than the face-specific texture features, the integration of these two features shows impressive complementarity, leading to 13.8% HTER. We also utilized different networks such as CaffeNet [10] and GoogLeNet [11], but we notice GoogLeNet outperforms CaffeNet.

Eye-blink detection. We evaluate the proposed eye-blink detection algorithm on the ZJU eye-blink dataset [13]. The proposed approach achieves 98.8% accuracy, which outperforms the SOTA accuracy of 95.7% [13] on ZJU dataset. Additionally, the proposed eye-blink detection is much faster than [13] (100 fps vs. 20 fps). After the eye-blink cue is fused with the deep texture features, HTER on Replay-Attack is further reduced from 13.8% to 12.4% (mainly reducing false acceptance of photograph attacks). These results show that the proposed eye-blink detection is helpful for assisting in photograph attack detection.

5 Summary and Conclusion

Cross-database face antispoofing replicates real application scenarios, and is a challenging problem for biometrics antispoofing. We propose a robust feature representation integrating deep texture features and eye-blink cue as countermeasures against presentation attacks of photographs and replays. Texture features are learned in a generic-to-specific way from both aligned face images and whole fames, which provide impressive complementarity. Eye-blink detection algorithm based on image difference is proposed to assist in texture cue with low

computational cost. Cross-database testing on MSU USSA, Idiap Replay-Attack, and CASIA FASD databases shows that the proposed method outperforms the state-of-the-art. Our future work includes jointly learning spatial-temporal representation for antispoofing.

References

1. S. Marcel, M. S. Nixon, and S. Z. Li, editors. *Handbook of Biometric Anti-Spoofing: Trusted Biometrics under Spoofing Attacks*. New York: Springer, 2014.
2. A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *IEEE Trans. Image Process.*, 24(12):4726–4740, Dec. 2015.
3. D. Wen, H. Han, and A. K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Security*, 10(4):746–761, Apr. 2015.
4. D. F. Smith, A. Wiliem, and B. C. Lovell. Face recognition on consumer devices: Reflections on replay attacks. *IEEE Trans. Inf. Forensics Security*, 10(4):736–745, Apr. 2015.
5. H. Han, S. Shan, X. Chen, S. Lao, and W. Gao. Separability oriented preprocessing for illumination-invariant face recognition. In *Proc. ECCV*, pages 307–320, 2012.
6. K. Patel, H. Han, and A. K. Jain. Secure face unlock: Spoof detection on smartphones. *IEEE Trans. Inf. Forensics Security*, 11(10):2268–2283, Oct. 2016.
7. J. Li, Y. Wang, T. Tan, and A. K. Jain. Live face detection based on the analysis of fourier spectra. In *Proc. SPIE*, pages 296–303, 2004.
8. T. F. Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *Proc. ICB*, pages 1–8, 2013.
9. D. Menotti, G. Chiachia, A. Pinto, W. Robson Schwartz, H. Pedrini, A. Xavier Falcao, and A. Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Trans. Inf. Forensics Security*, 10(4):864–879, Apr. 2015.
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.
11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, pages 1–9, 2015.
12. C. Wang, K. Li, Z. Wu, and Q. Zhao. A DCNN based fingerprint liveness detection algorithm with voting strategy. In *Proc. CCBR*, pages 241–249, 2015.
13. G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *Proc. ICCV*, pages 1–8, 2007.
14. M. Szwoch and P. Pieniazek. Eye blink based detection of liveness in biometric authentication systems using conditional random fields. In *Proc. ICCVG*, pages 669–676, 2012.
15. H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1148–1161, Jun. 2015.
16. I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *Proc. IEEE BIOSIG*, pages 1–7, 2012.
17. Z. Zhuang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *Proc. ICB*, pages 26–31, 2012.
18. Z. Boulkenafet, J. Komulainen, and A. Hadid. Face anti-spoofing based on color texture analysis. In *Proc. ICIP*, pages 2636–2640, 2015.