Using LDP-TOP in Video-Based Spoofing Detection

Quoc-Tin Phan[1], Duc-Tien Dang-Nguyen[2], Giulia Boato[1], and Francesco G. B. De Natale[1]

[1] University of Trento, Italy
[2] Dublin City University, Ireland

Abstract. Face authentication has been shown to be vulnerable against three main kinds of attacks: *print, replay*, and *3D mask*. Among those, video replay attacks appear more challenging to be detected. There exist in the literature many countermeasures to face spoofing attacks, but a sophisticated detector is still needed to deal with particularly high-quality video based attacks. In this work, we perform analysis on the noise residual in frequency domain, and extract discriminative features by using a dynamic texture descriptor to characterize video based spoofing attacks. We propose a promising detector, which produces competitive results on the most challenging dataset of video based spoofing.

Keywords: Face Anti-Spoofing, Local Derivative Pattern, Video Based Attacks

1 Introduction

Among many applications of biometric authentication, face authentication has been considered as an efficient and reliable access control mechanism. A face authentication system works in less intrusive manner, which requires little cooperation from users. Thanks to the advances in face detection and recognition, a face authentication system can be flawlessly deployed on low-cost devices.

"Fingerprints cannot lie, but liars can make fingerprints." [7]

Like other biometric modalities, a face authentication system can be bypassed easily even at very low cost. We group such spoofing attacks into three main categories: i) *Print attacks*: the use of printed photo of an authorized user, ii) *Replay attacks*: a photo or a video of an authorized user is replayed on a digital screen, iii) *3D mask attacks*: the authorized user's face is simulated by 3D mask. The vulnerability of face authentication system to spoofing attacks has motivated plenty of proposed countermeasures in the past few years.

One of first attempts to detect print attacks is introduced in [1]. By analyzing the total amount of movements over video frames, print attacks can be effectively detected. This explains the great success, i.e., high performance detection accuracy [13], of motion-based methods. While print attacks leave clear and inevitable evidences, photos or videos replayed on a digital screen are more challenging to be detected [4, 15]. Replayed photos are generally in higher quality (e.g., color, contrast) compared with printed photos, and replayed videos can easily fool an authentication system since the

face in a high quality video and a real face are almost indistinguishable. By the advance of 3D printing techniques, another kind of face spoofing attack has been introduced in [5], the so-called 3D mask attacks. In [5] the vulnerability of face authentication system to spoofing with 3D masks is shown.

In this work, our concentration is placed on video-based attacks which refer to replaying a video on a digital screen. A number of approaches have been proposed [14, 9, 8, 2, 6, 11, 10], and we classify them into two main categories: *spatial domain* and *frequency domain* analysis.

Methods performing analyses on spatial domain take into account directly the pixel values of the suspected image. In [14], discriminative features characterizing spoofing attacks are extracted, such as specular reflection, blurriness, chromatic moment, and color diversity. Multiple classifiers are trained based on the concatenation of all extracted features and final decision is given by taking mutual information from multiple classifiers. Another approach exploiting dynamic texture descriptor has been introduced in [9]. In this work, the authors extend Local Derivative Pattern (LDP) to LDP on Three Orthogonal Planes (LDP-TOP) in order to capture highly detailed information on spatial domain as well as subtle face movements over frames. By analyzing image distortion artifacts such as reflection, color distribution, Moiré patterns (i.e., overlapping grids) and face shape deformation, the authors in [8] have developed a countermeasure to spoofing attacks on mobile phones. Most recently, the analysis on disparities between color texture of genuine faces and fake ones is investigated in [2]. The authors show that the use of YCbCr and HSV color spaces results in generally better detection accuracy.

Despite the success of methods relying on spatial domain analysis, discriminative features extract in spatial domain may become scene-dependent. Roughly speaking, instead of addressing only artifacts of spoofing attacks, spatial domain models learn also redundant information of the scene. Frequency domain analysis focuses on periodic patterns, i.e., Moiré patterns, which present as peaks in the spectrum image. Such periodic patterns are independent to image scene. In [6], the authors extract Moiré patterns from still images using a bandpass filter and make analysis on frequency domain. A large-scale dataset dedicated to video-based attacks is introduced in [11]. This is a challenging dataset containing huge number of videos recorded under different environmental conditions. The authors also provide a baseline method analyzing the noise residual of the video in terms of visual rhythms. Another method dedicated to noise analysis is mentioned in [10]. Instead of extracting only *low-level* features which are basically artifacts on the spectrum video, to reduce the sensitivity between intra- and extra-class variations, the authors propose to extract also mid-level features as *visual codebooks* from low-level features.

In this work, we select to analyze artifacts on frequency domain for some reasons. First, it is challenging to extract discriminative features on the spatial domain due to the contamination of the scene. Despite good detection performance on some benchmarking datasets, methods on the spatial domain tend to get overfitted on specific conditions of capturing. This results in the low generality in real applications where attempted attacks might be performed in various conditions. Moreover, most of methods on the spatial domain rely on the reliability of face detection and tracking algorithms, which are not always successful under poor conditions. We propose to analyze the spectrum video

by using a dynamic (spatial-temporal) texture descriptor. Dynamic texture descriptor is able to capture not only highly detailed information on spatial domain but also subtle changes over time. Thanks to the success of Local Derivative Pattern on Three Orthogonal Planes (LDP-TOP) [9], we select LDP-TOP as the descriptor. The main difference to [9] is that we use LDP-TOP to analyze discriminative textures of spectrum videos. Our proposed method outperforms two recent and closely related works on large-scale dataset of video-based attacks in terms of detection accuracy. By analyzing only the noise residual, we can skip face detection and tracking which require more computation and depend heavily on environmental conditions.

The paper is structured as follows: Section 2 presents in detail the schema of the proposed methods, and experimental analysis is described in Section 3. In Section 4, we draw some conclusions.

2 The proposed method

2.1 The recaptured artifacts

Video-based spoofing attacks can easily bypass the authentication system because the face in a high quality video is nearly indistinguishable with the real face. Nevertheless, when the camera records a digital display, the resulting video presents a number of visible artifacts:

– **Moiré pattern**: In recaptured videos, Moiré patterns occur in the form of visible periodic or almost periodic patterns in every video frame. Specifically, the sampling grid of the displaying device is overlaid by the sampling grid of acquisition device resulting the third grid pattern. Misalignment between the two devices also causes different observable forms of Moiré patterns. Shown in Figure 1 (a) is the Moiré pattern generated from two overlaid patterns containing parallel lines. Fig. 1 (b) is original image shown in Macbook Pro screen, Fig. 1 (c) and 1 (d) show images captured from Macbook Pro screen by HTC Desire HD phone and Apple Ipad Air, respectively. Note that the original image is taken from UVAD dataset [11].

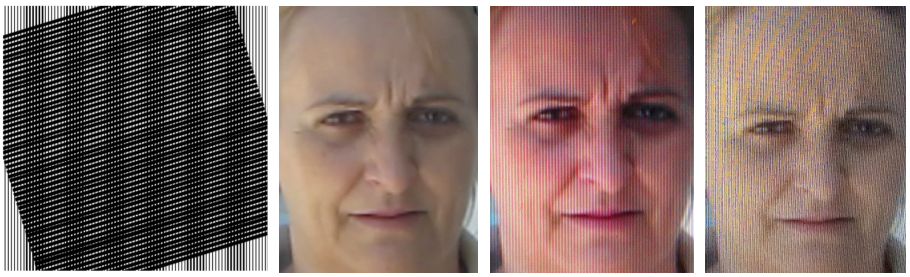

Fig. 1. Examples of Moiré patterns.

(a) (b)

Fig. 2. Example of flickering effect.

– **Flickering effect**: This effect corresponds to horizontal or vertical lines of equal spaces, caused by the desynchronization between the flashing frequencies of displaying and acquisition device. These noticeable lines might move vertically or horizontally over video frames. Shown in Fig. 2 (a) is an example of flickering effect observed when capturing the image from a screen display by the Olympus SP 800UZ. The alignment of these effects is highlighted in Fig. 2 (b).
– **Other artifacts**: Besides Moiré patterns and flickering effects, a recaptured video is generally blurred compared with the original video. The change in color tone can be also observed according to different acquisition devices.

Since aforementioned artifacts are independent to image scene, they should be isolated from the content of the image by using an image filter. Moiré pattern and flickering effect are almost periodic, they are characterized by high-energy peaks in the frequency domain. A practical way of detecting such artifacts is to perform Fourier transform of the suspected image, and apply a matched filter to the transformed image. This naive approach might result in many false detections. On the other hand, blurring artifact implies the increase of low frequency components which are challenging to be detected by simple thresholding.

2.2 Processing pipeline

The complete processing pipeline of the proposed method is presented in Fig. 3 and includes three main steps:

A. Noise extraction and spectrum calculation

Since all artifacts present entirely on every frame, we first define a *region of interest* (RoI) in the spatial domain. Taking into account RoI of size $w \times h$, we can reduce greatly the computation time of our proposed method. Denote the frame t-th of the video V as $V^{(t)}$, we extract its residual by applying a denoising filter $\mathcal{F}$ and subtracting the denoised frame from $V^{(t)}$ to obtain $V_r^{(t)}$.

$$V_r^{(t)} = V^{(t)} - \mathcal{F}(V^{(t)}), \quad 1 \leq t \leq M, \tag{1}$$

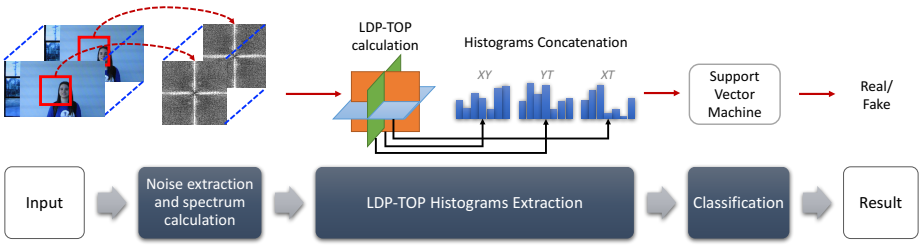

Fig. 3. Schema of the proposed method.

where M is the number of frames. The noise residual $V_r^{(t)}$ contains the noise pattern of frame t-th. Let $V_f^{(t)}$ denote the presentation of $V_r^{(t)}$ on the frequency domain. $V_f^{(t)}$ is calculated using 2D Discrete Fourier Transform (DFT).

$$V_f^{(t)} = DFT(V_r^{(t)}) \tag{2}$$

To get the final spectrum video V_s, we collect all Fourier spectrum $|V_f^{(t)}|$, and calculate their logarithmic scale.

$$V_s^{(t)} = \log(|V_f^{(t)}| + 1) \tag{3}$$

B. Histogram extraction

In this work, we treat the Fourier transformed video (spectrum video) as a three-dimensional texture map, and then apply a sophisticated local descriptor on the spectrum video in order to extract meaningful features. We select our previously proposed Local Derivative Pattern on Three Orthogonal Planes, the so-called LDP-TOP [9], as the descriptor thanks to its success in spoofing detection. It worths noting that in this work, we apply LDP-TOP to analyze discriminative textures of spectrum videos, i.e., not on the spatial domain as in [9].

Given the image I, the first-order derivative along each direction $\alpha = \{0°, 45°, 90°, 135°\}$ is denoted as I_α. Let Z_0 be a pixel, and $Z_i, i = 1, \cdots, 8$ be the neighboring pixels around Z_0. The four first-order derivatives at $Z = Z_0$ can be written as:

$$I_{0°}(Z_0) = I(Z_0) - I(Z_4) \quad I_{45°}(Z_0) = I(Z_0) - I(Z_3)$$
$$I_{90°}(Z_0) = I(Z_0) - I(Z_2) \quad I_{135°}(Z_0) = I(Z_0) - I(Z_1)$$

Z_1	Z_2	Z_3
Z_8	Z_0	Z_4
Z_7	Z_6	Z_5

Generally, the n^{th}-order directional LDP, $\text{LDP}_\alpha^n(Z_0)$, in direction α at $Z = Z_0$ is defined as:

$$\text{LDP}_\alpha^n(Z_0) = \{f(I_\alpha^{n-1}(Z_0), I_\alpha^{n-1}(Z_1), \cdots, f(I_\alpha^{n-1}(Z_0), I_\alpha^{n-1}(Z_8)\}, \tag{4}$$

where $I_\alpha^{n-1}(Z_0)$ is the $(n-1)^{th}$-order derivative in direction α at $Z = Z_0$, and $f(I_\alpha^{n-1}(Z_0), I_\alpha^{n-1}(Z_i))$ is defined as

$$f(I_\alpha^{n-1}(Z_0), I_\alpha^{n-1}(Z_i)) = \begin{cases} 0, & \text{if } I_\alpha^{n-1}(Z_i) \cdot I_\alpha^{n-1}(Z_0) > 0 \\ 1, & \text{if } I_\alpha^{n-1}(Z_i) \cdot I_\alpha^{n-1}(Z_0) \leq 0 \end{cases}, \quad i = 1, \cdots, 8. \quad (5)$$

Equation (4) encodes $(n-1)^{th}$-order gradient transitions, resulting the n^{th}-order binary pattern on the local region. Binary patterns are represented in 4 histograms, each describing a specific direction. This way, the final histogram contains 4×2^8 bins.

We consider the time window size T_{ws} ($T_{ws} \leq M$) as the number of chronological-order frames. Only the first T_{ws} frames are taken into account for histogram extraction. In Fig. 3, three planes XY, XT, YT are pair-wise orthogonal, where XY corresponds to a frame of the spectrum video. XT, YT refer to horizontal and vertical planes. As a result, we end up three 2D texture maps of size $w \times h$, $w \times T_{ws}$, and $h \times T_{ws}$.

Histogram of LDP-TOP is the concatenation of three LDP histograms from three orthogonal planes. Finally we end up a feature vector of dimension $3 \times 4 \times 2^8$.

C. Classification

We use Support Vector Machine (SVM) [3] to learn and detect video-based attacks. Specifically, we embed the Histogram Intersection Kernel (HIK) as the kernel of SVM. HIK was introduced in [12] to compare color histograms. A HIK between two histogram a and b is simply defined as:

$$K(a, b) = \sum_{i=1}^{n} \min(a_i, b_i), \quad a_i \geq 0, b_i \geq 0. \quad (6)$$

In the next section, we present how parameters are experimentally selected and give some insights on the effectiveness of the proposed method.

3 Experimental analysis

3.1 Parameter selection

We validate the effectiveness of the proposed method on the Unicamp Video-Attack Database (UVAD) [11]. This dataset contains valid access and attempted attack videos of 404 different identities. Each video is recorded at high quality, 30 frames per seconds, and 9 seconds long. The resolution of all videos is fixed to 1366×768, where the face appears approximately in middle of the frame. Six cameras have been used to record real access videos. Each person is recorded by only one camera, but in different scenarios (different backgrounds, lighting conditions and places), generating 808 real access videos in total. For attempted attacks, real access videos are displayed in seven different display screens and recaptured by the same set of cameras used before. Finally, the recapturing process produces $16,268$ attempted attack videos.

We evaluate our method using videos from all six cameras: Sony, Kodak, Olympus, Nikon, Canon, and Panasonic. The training set contains real access and attempted attack videos from Sony, Kodak and Olympus, resulting in 344 real access and 3528 attempted attack videos. On the other hand, real access and attempted attack videos from Nikon, Canon and Panasonic are used for testing purpose, resulting in 60 real access

Fig. 4. The first row presents example video frames of real access in outdoor (the first two images) and indoor (the last two) condition. The second row presents example video frames of attempted attacks in outdoor (the first two images) and indoor (the last two) condition.

and 6356 attempted attack videos. This setup is applied in [10]. Specifically, we select 300 (30 real access and 270 attempted attacks) samples of the training set to serve as the development set which is used for decision threshold estimation. Some example video frames are shown in Fig. 4 [3].

We report statistics mainly in terms of Half Total Error Rate (HTER), and Area Under the Curve (AUC). The decision of *positive* or *negative* is simply made by comparing the output score of the test sample with a decision threshold. This threshold directly causes two kinds of error: False Rejection Rate (FRR) referring to rejecting real faces, and False Acceptance Rate (FAR) referring to accepting spoofed faces. HTER is applied as threshold-dependent performance measurement, and is defined as the average of FRR and FAR. On the other hand, AUC is calculated as the area under the ROC curve, and is invariant to decision threshold.

In order to select the best configuration in our proposed method, we run experiments on UVAD under various settings. We test the effectiveness of three denoising filters: Median, Gaussian and Wiener, and use second-order and third-order LDP-TOP to extract features from spectrum videos. We consider the RoI of 256×256 pixels which is located in the center of every frame. We set the time window size T_{ws} to 100. The window size of all denoising filters are identically 7×7, and Gaussian standard deviation is set to 2. Fig. 5 depicts DET (Detection Error Tradeoff) curves of all configurations.

As can be seen in Fig. 5, Gaussian filter allows best error tradeoff in both of second-order and third-order LDP-TOP. More specifically, third-order LDP-TOP can produce better detection performance compared with second-order LDP-TOP. Median filter is particularly powerful in removing outlier pixels (i.e., salt and pepper noise) and preserving edges, but it appears less efficient in our analysis. Numeric results on UVAD are shown in Table 1.

[3] The set of identities involved in real access videos are disjoint with those involved in attempted attacks. That is why we do not show the attempted attack of the same identity.

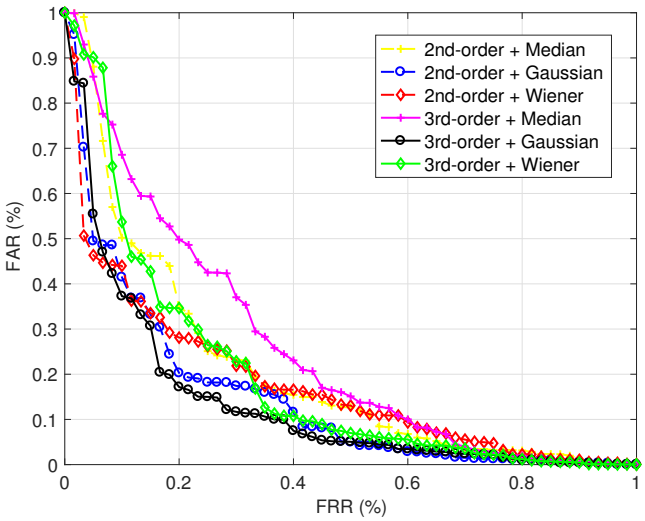

Fig. 5. DET curves under different configurations.

Table 1. Results the proposed method on UVAD.

(a) 2nd-order LDP-TOP

	HTER (%)	AUC (%)
Median	29.39	80.01
Gaussian	25.45	86.19
Wiener	30.15	83.24

(b) 3rd-order LDP-TOP

	HTER (%)	AUC (%)
Median	33.34	74.93
Gaussian	23.69	87.58
Wiener	24.25	81.70

3.2 Comparisons with State-of-the-Art approaches

In this section, we discuss about performance comparisons on UVAD. Since our method is dedicated to noise analysis, we compare our results mainly with the two recent and closely related works: Visual Rhythm (VR) [11] and Visual Codebooks (VC) [10].

VR based approach captures noise signatures in terms of 2D maps which are basically generated by traversing the spectrum video in horizontal, vertical and zigzag directions. This can be considered as a baseline method in UVAD dataset. The authors in [11] report results in a different dataset configuration where the size of testing set is much smaller than our consideration which is described in the previous section. In order to make more thorough and coherent analysis, we run visual rhythm extraction on our dataset configuration by using the implementation provided by the authors. We select Gaussian as the denoising filter and Gray-Level Co-Occurrence Matrix (GLCM) as the texture descriptor since this is the best reported configuration in [11]. Table 2 depicts detection performance of VR on UVAD. As can be seen, VR reaches its best performance with visual rhythm extracted from vertical direction.

According to VC, low-level features are extracted from small cuboids in the spectrum video, and the authors apply Bag-of-Visual-Word to map onto a more discrimina-

Table 2. Results of VR on UVAD.

	HTER (%)	AUC (%)
Horizontal	52.01	51.02
Vertical	28.09	73.48
Zigzag	41.28	71.77

Table 3. Results of all methods on UVAD.

	FAR (%)	FRR (%)	HTER (%)
Correlation [1]	81.60	14.56	48.06
LBP [4]	27.41	66.04	46.72
VR [11]	44.52	**11.67**	28.09
VC [10]	44.73	15.00	29.87
Proposed method	**7.38**	40.00	**23.69**

tive mid-level representation. We report all statistics in [10] since those were collected in the same dataset configuration.

Shown in Table 3 are performance comparisons with VR and VC. In our method we choose the best configuration in which the Gaussian denoising filter and third-order LDP-TOP are applied. Compared with VR and VC, our proposed method achieves lower detection error, as shown in Table 3. Moreover, we consider only the RoI of size 256×256, which is much smaller than the entire volume of the video. There exist plenty of methods contributing to spatial domain analysis, however, reproducing all results of them on UVAD is beyond the scope of this paper. We emphasize more on methods on the frequency domain and show here only results of baseline methods on the spatial domain [1, 4]. It is evident that methods in [1, 4] perform poorly in detecting video based spoofing attacks.

3.3 Computational complexity

Considering the computational complexity of the method, we provide here the analysis of the most impactful steps. Denote N the total number of pixels in RoI, $N = w \times h$. The complexity of the filtering step is $\mathcal{O}(N)$. Fast Fourier Transform requires $\mathcal{O}(N \times \log(N))$, and third-order LDP-TOP requires $\mathcal{O}(N)$ computations. Finally, the computational complexity is bounded to $\mathcal{O}(T_{ws} \times N \times \log(N))$, where T_{ws} is simply the number of considered frames.

Our testing is run on the computer with the following configuration: Intel(R) Xeon(R) CPU E5-2630 v3 2.40GHz; 64 Gb of RAM; Linux Ubuntu 14.04 LTS 64 bit installed. By considering small RoI, it is shown that the proposed approach can be applied for real-time applications.

The Matlab implementation can be obtained via:
`github.com/quoctin/anti_video_spoofing`.

4 Conclusions

We have proposed a novel approach for detecting video based spoofing attacks. A recaptured video basically contains discriminative artifacts such as blurring, Moiré patterns and flickering effects. Those signatures are present in the frequency domain and can be analyzed by using dynamic texture descriptor. Thanks to the superiority of Local Derivative Pattern on Three Orthogonal Planes (LDP-TOP), we achieve promising results compared with related works. Future extension of this work will be devoted to designing sophisticated filters for extracting aforementioned artifacts. We will also find a mechanism to reduce the dimension of feature vectors.

References

[1] Anjos, A., Marcel, S.: Counter-Measures to Photo Attacks in Face Recognition: a public database and a baseline. In: International Joint Conference on Biometrics 2011 (Oct 2011)

[2] Boulkenafet, Z., Komulainen, J., Hadid, A.: Face Spoofing Detection Using Colour Texture Analysis. IEEE Trans. Information Forensics and Security 11(8), 1818–1830 (Aug 2016)

[3] Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intelligent Systems and Technology 2, 27:1–27:27 (2011)

[4] Chingovska, I., Anjos, A., Marcel, S.: On The Effectiveness of Local Binary Patterns in Face Anti-spoofing. In: International Conference of Biometrics Special Interest Group. pp. 1–7 (Sept 2012)

[5] Erdogmus, N., Marcel, S.: Spoofing Face Recognition With 3D Masks. IEEE Trans. Information Forensics and Security 9(7), 1084–1097 (July 2014)

[6] Garcia, D.C., de Queiroz, R.L.: Face-Spoofing 2D-Detection Based on Moiré Pattern Analysis. IEEE Trans. Information Forensics and Security 10(4), 778–786 (April 2015)

[7] Marcel, S., Nixon, M.S., Li, S.Z.: Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks. Springer Publishing Company, Incorporated (2014)

[8] Patel, K., Han, H., Jain, A.K.: Secure Face Unlock: Spoof Detection on Smartphones. IEEE Trans. Information Forensics and Security 11(10), 2268–2283 (Oct 2016)

[9] Phan, Q.T., Dang-Nguyen, D.T., Boato, G., De Natale, F.G.B.: Face Spoofing Detection using LDP-TOP. In: IEEE International Conference on Image Processing. pp. 404–408 (Sept 2016)

[10] Pinto, A., Pedrini, H., Schwartz, W.R., Rocha, A.: Face Spoofing Detection Through Visual Codebooks of Spectral Temporal Cubes. IEEE Trans. Image Processing 24(12), 4726–4740 (Dec 2015)

[11] Pinto, A., Schwartz, W.R., Pedrini, H., d. R. Rocha, A.: Using Visual Rhythms for Detecting Video-Based Facial Spoof Attacks. IEEE Trans. Information Forensics and Security 10(5), 1025–1038 (May 2015)

[12] Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision 7(1), 11–32 (1991)

[13] Tirunagari, S., Poh, N., Windridge, D., Iorliam, A., Suki, N., Ho, A.T.S.: Detection of Face Spoofing Using Visual Dynamics. IEEE Trans. Information Forensics and Security 10(4), 762–777 (April 2015)

[14] Wen, D., Han, H., Jain, A.K.: Face Spoof Detection With Image Distortion Analysis. IEEE Trans. Information Forensics and Security 10(4), 746–761 (April 2015)

[15] Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A Face Antispoofing Database with Diverse Attacks. In: 5th IAPR International Conference on Biometrics (ICB). pp. 26–31 (March 2012)