FACE SPOOFING DETECTION USING LDP-TOP

Quoc-Tin Phan, Duc-Tien Dang-Nguyen, Giulia Boato, Francesco G. B. De Natale

Department of Information Engineering and Computer Science - University of Trento, Italy
quoctin.phan@unitn.it; {dangnguyen, boato}@disi.unitn.it; denatale@ing.unitn.it

ABSTRACT

In this paper, we propose a novel approach for face spoofing detection using the high-order Local Derivative Pattern from Three Orthogonal Planes (LDP-TOP). The proposed method is not only simple to derive and implement, but also highly efficient, since it takes into account both spatial and temporal information in different directions of subtle face movements. According to experimental results, the proposed approach outperforms state-of-the-art methods on three reference datasets, namely Idiap REPLAY-ATTACK, CASIA-FASD, and MSU MFSD. Moreover, it requires only 25 video frames from each video, i.e., only one second, and thus potentially can be performed in real time even on low-cost devices.

Index Terms— Face Anti-Spoofing, Local Derivative Pattern, Video Forensics

1. INTRODUCTION

Face recognition is one of the most commonly used techniques in applications of biometrics, e.g. access control, law enforcement, multimedia communication, human-computer interaction. Like other biometric modalities, however, a face recognition system can be attacked easily and at very low cost by two common attacks, namely *print attack* and *replay attack*. In print attacks, face spoofing is carried out by presenting a printed photo to a camera. In replay attacks, on the other hand, the attackers replay a previously recorded face image or video of a targeted user in order to spoof the biometric system. As attackers only need to acquire a printed photo or a video of the authorized user's face, with current technologies these types of attack can be carried out easily in both remote and logical access control systems protected by a face recognition system.

The literature presents a series of techniques for detecting face spoofing, which can be grouped into following categories. *Cue-based methods:* focusing on printed photo attacks. These methods capture important clues connected with vitality, such as eye blink [1], mouth movement [2] and head rotation [3]. However, it takes relatively long time to accumulate stable vitality features for face spoof detection. Additionally, these methods can be confused by other types of motion, e.g., background motion irrelevant to the facial aliveness or replayed motion in the attacked videos. *Data-driven-based methods:* exploited to detect both types of attack. These methods tend to use generic image-processing and computer vision algorithms, where they exploit mainly the motion information, like optical flow [4] or dynamic mode decomposition (DMD) [5], or texture information, like Local Binary Pattern (LBP) [6] or LBP-TOP [7]. However, they can be overfitted to one particular illumination and imagery conditions and hence do not generalize well to databases collected under different conditions. Recently, the closely related problem of discrimination between computer generated and natural human faces have been investigated by exploiting the differences in face geometry evolution [8], face dynamics [9], patterns in expressions [10], or tiny fluctuations in the appearance of a face [11]. Other methods exploit different sources with respect to 2D intensity image, such as 3D depth [12], IR image [13], and voice [14]. Nevertheless, these methods requires extra information, thus they have a narrower application range.

In this paper, we propose a data-driven-based method that exploits the high-order Local Derivative Pattern from Three Orthogonal Planes (LDP-TOP). The key innovation introduced here is the exploitation of the facial dynamic information in both space and time domain in different directions, capturing the subtle movements on the face. This approach not only improves the performance of the previous methods exploiting binary pattern, e.g. LBP-TOP [7], but also outperforms other state-of-the-art methods on most common datasets. Moreover, it requires only 25 frames, approximately a second, from each video, and potentially can be performed in real time even on low-cost devices.

The rest of this paper is organized as follows: Section 2 presents the proposed framework and describes in detail the processing steps of the LDP-TOP histograms extraction and concatenation. In Section 3 we depict the extensive experimental analysis, while some concluding remarks are drawn in Section 4.

2. THE PROPOSED METHOD

Early work for detecting face spoofing based on LBP, e.g., [6, 15], shown that real faces contain different texture pat-

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

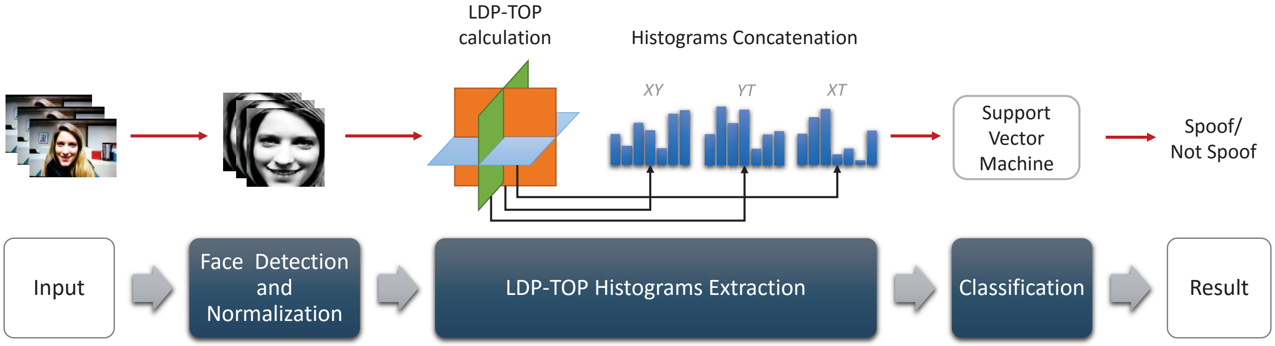


Fig. 1. Schema of the proposed method.

terns in comparison with fake ones. These techniques, however, analyze single frames, not considering the relationship of frames over time. In face spoofing attacks, the real faces are captured and replayed in front of the camera. Hence, texture analysis is a powerful technique to discriminate the face on the real world and the one on planar objects. Motion analysis is also significantly important for face spoofing detection and in combination with texture analysis can generate a powerful countermeasure. The attempt to extend LBP to image sequences, the so-called LBP-TOP, explored the spatial and temporal information in face spoofing detection [7]. Besides LBP, Local Derivative Pattern (LDP) [16] has been proposed as higher-order local binary descriptor and proved to have better performance in face recognition in comparison with LBP. Different from LBP, which encodes the relationship between the central point and its neighbors, the LDP templates extract higher-order local information by encoding various distinctive spatial relationships contained in a given local region, thus highlighting subtle changes on the face. Inspired from these work, we propose to extend LDP to a dynamic texture descriptor, thus exploiting the higher-order LDP from Three Orthogonal Planes, named LDP-TOP, to detect face spoofing attacks.

The proposed method contains three main steps (summarized in Fig. 1). In the first step (*face detection and normalization*), each video frame is gray-scaled and passed through a face detector. The detected faces are then geometrically normalized. In the second step (*LDP-TOP histograms extraction*), LDP operators [16] are applied on three orthogonal planes intersecting at the center of the XY, XT, and YT direction, where T is the time axis (the frame sequence), and then the extracted histograms are concatenated. Finally, in the last step (*classification*), Support Vector Machine (SVM) is applied to classify the extracted histograms and determine if the video input is spoofed or not.

For *face detection and normalization*, we first apply Viola-Jones method [17] to detect the face and keep the face bounding box stable over frames in order to capture subtle face moments properly. The extracted faces are then normalized and scaled to the resolution of $h \times w$.

Z_1	Z_2	Z_3
Z_8	Z_0	Z_4
Z_7	Z_6	Z_5

Fig. 2. 8 neighbors around Z_0

LDP-TOP histograms extraction is then applied on all normalized faces. Given a face region, the first-order derivative along a direction α is denoted as I_α. In this method, we consider $\alpha \in \{0°, 45°, 90°, 135°\}$. Let Z_0 be a point on the image $I(Z)$, and Z_i, $i = 1, \cdots, 8$ be the neighboring points around Z_0 (see Fig. 2). The four first-order derivatives at $Z = Z_0$ can be written as:

$$I_{0^\circ}(Z_0) = I(Z_0) - I(Z_4) \qquad I_{45^\circ}(Z_0) = I(Z_0) - I(Z_3)$$
$$I_{90^\circ}(Z_0) = I(Z_0) - I(Z_2) \qquad I_{135^\circ}(Z_0) = I(Z_0) - I(Z_1)$$

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

$$\begin{aligned} LDP^n_\alpha(Z_0) = \{ & f(I^{n-1}_\alpha(Z_0), I^{n-1}_\alpha(Z_1)), \\ & f(I^{n-1}_\alpha(Z_0), I^{n-1}_\alpha(Z_2)), \\ & \cdots, f(I^{n-1}_\alpha(Z_0), I^{n-1}_\alpha(Z_8)\} \end{aligned} \tag{1}$$

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

$$f(I^{n-1}_\alpha(Z_0), I^{n-1}_\alpha(Z_i)) =$$
$$\begin{cases} 0, & \text{if } I^{n-1}_\alpha(Z_i) \cdot I^{n-1}_\alpha(Z_0) > 0 \\ 1, & \text{if } I^{n-1}_\alpha(Z_i) \cdot I^{n-1}_\alpha(Z_0) \leq 0 \end{cases}, \quad i = 1, 2, \cdots, 8. \tag{2}$$

Eq. (1) encodes $(n-1)^{th}$-order gradient transitions, resulting in the n^{th}-order binary pattern on the local region. Given the binary patterns, we represent them in terms of 4 histograms, each describing a specific direction. This way, the final image histogram contains 4×2^8 bins.

We extract LDP histograms not only from the image plane, but also over time, i.e., taking into account also motion information, which can be observed from T_{ws} chronological-order frames under different time resolutions R. In our approach, the time window size T_{ws} is considered as the number of frames being counted, while the time resolution R

is defined as the temporal distance of considered frames. Suppose that f_i is the current frame, the next frame being counted will be f_j where $R = j - i$. In the case of multi-resolution, we denote $R = [1, r]$, the histograms where $R = 1, 2, \ldots, r$ are concatenated to form the final histogram. After this step, each input video is represented as a concatenated histogram of $3 \times 4 \times 2^8 = 3072$ bins in case of single time resolution, while the number of bins is $r \times 3 \times 4 \times 2^8$ in case of multi-resolution.

Finally, in *Classification* step, we apply SVM using LibSVM [18] with the Histogram Intersection Kernel [19] to classify the videos spoofed or not.

3. EXPERIMENTS

In this section, we report the experimental validation carried out in order to demonstrate how selected parameters affect the performance of LDP-TOP and to compare the effectiveness of the proposed method with respect to the state-of-the-art.

To have good comparisons, we performed our experiments on three common datasets: Idiap REPLAY-ATTACK [6], CASIA-FASD [20] and MSU MFSD [21][1]. The Idiap contains 1200 short videos in total, including three types of attacks: *print* (the operator presents printed photos in front of the carmera), *mobile* (digital photos or videos are displayed through iPhone screen), and *highdef* (high resolution photos or videos are displayed on iPad screen). The CASIA consists of 600 videos describing more complex attacks: *warped photo attack* (the operator attempts to warp the high quality photo and simulate the real face), *cut photo attack* (the operator hides behind and simulates eye blinking through the cut-out part of the photo), and *video attack* (high resolution videos are replayed on iPad). In CASIA, the subset including only high-quality videos is denoted as CASIA$_H$. MSU dataset provides publicly 280 videos of photo and video attack attempts to 35 subjects, including *printed photo* (the face is printed on A3 paper and presented in front of the camera) and *video replay* (the video is previously captured and replayed on Ipad and Iphone screen).

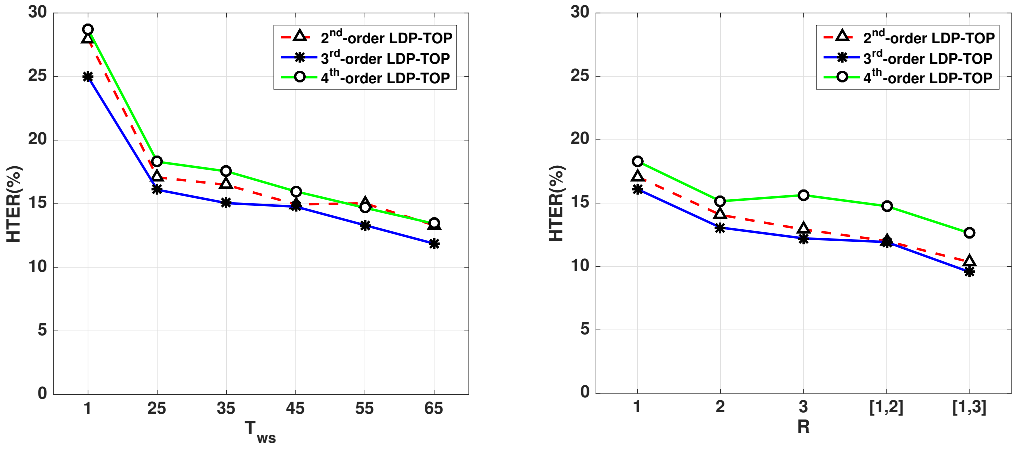
In the performance measurement, we report statistics in terms of Equal Error Rate (EER), Half Total Error Rate ($HTER$), and True Positive Rate (TPR). The term EER is defined as the value on the Detection Error Trade-off (DET) curve - plotting False Acceptance Rate (FAR) on the x-axis and False Acceptance Rate (FRR) on the y-axis - where FAR and FRR are equal. EER can be used to give a threshold-independent performance measurement. $HTER$, on the other hand, can be applied as threshold-dependent performance measurement, and is defined as the average of FRR and FAR. More details about these measurements can be found in [22].

Face locations are provided in the MSU and Idiap datasets. For CASIA, we used CAMShift package from Matlab Toolbox, which is based on Viola-Jone algorithm [17] and Continuously Adaptive Mean Shift (CAMShift) [23], to detect face locations since they were not provided. All experiments were performed on a PC with CPU Intel Core i5 - 2.5 GHz, Ram 8 GB DDR3 with Matlab 2015b installed. The Matlab code of the proposed method can be obtained via `http://mmlab.disi.unitn.it/codes/LDP-TOP/`.

In the first experiment, we observed the effectiveness of the second-order, third-order, and fourth-order LDP-TOP under different time window sizes T_{ws}. Fig. 3 (a) represents the behavior of the second-order, third-order, and fourth-order LDP-TOP on the challenging dataset CASIA. It can be seen that LDP-TOP ($T_{ws} > 1$) significantly outperformed normal LDP ($T_{ws} = 1$). This result confirmed the important role of the temporal information in countering face spoofing attacks. Moreover, the performance increased gradually with the time window size. As mentioned in [16], high-order derivative descriptors not only capture more detailed information but also noise presented in an image. This fact has been verified clearly in Fig. 3, in which the third-order LDP-TOP resulted in best detection ability while the higher-order (herein is the fourth-order) LDP-TOP performed less efficiently. Next, we observed the behavior of the second-order, third-order, and fourth-order LDP-TOP by varying the time resolution while keeping the time window size constant, where $T_{ws} = 25$. The reason of choosing $T_{ws} = 25$ is that the experimental videos are not sufficient in length (some CASIA videos last around 2 seconds). The performance of multi-resolution LDP-TOP where $R = [1, r]$ is generally higher than the case of single resolution where $R = r$. Best results were recorded with $R = [1, 3]$ on the CASIA, see Fig. 3 (b).

A second set of experiments was run to compare the proposed method with state-of-the-art approaches in two contexts: *intra-dataset*, where the training and testing sets are taken from the same dataset, and *cross-dataset*, where the training and testing sets are taken from different datasets. In intra-dataset test, we followed the evaluation protocol defined on the Idiap where the decision threshold δ is selected on EER on the development set. Since the development set is not defined in CASIA and MSU, we applied a k-fold cross-validation with $k = 5$ and reported the average statistics over 5 iterations, as in previous work [5, 6, 7]. We compared best results of our approach (obtained at $T_{ws} = 25$, $R = [1, 3]$, $h = w = 64$ on the Idiap, and $h = w = 128$ on the CASIA, CASIA$_H$ and MSU) with the best results of the two methods exploiting binary pattern: LBP [6] and LBP-TOP [7]. Furthermore, we also compared our work with the two recent work: DMD+LBP [5], and IDA [21]. Shown in Table 1 and 2 are the results measured in terms of $HTER$ and EER, respectively. According to these results, the proposed method outperforms all other approaches on all datasets. It is worth noticing that for a fair comparison, we took into consideration

[1] For simplicity, the Idiap REPLAY-ATTACK, CASIA-FASD and MSU MFSD dataset are referred as the Idiap, CASIA and MSU, respectively as in the rest of this paper.



(a)

(b)

Fig. 3. Performance of the 2^{nd}-order, 3^{rd}-order, and 4^{th}-order LDP-TOP on the CASIA under different time window sizes (a) and different time resolutions (b).

Table 1. Intra-dataset performance comparison of the proposed LDP-TOP with SoA methods in terms of $HTER(\%)$.

Method	Idiap	CASIA	MSU	CASIA$_H$
LBP [6]	15.16	18.17	–	–
LBP-TOP [7]	7.60	20.71	–	–
DMD+LBP [5]	3.75	21.75	–	–
IDA [21]	7.41	–	–	–
LDP-TOP	1.75	9.56	7.70	5.13

Table 2. Intra-dataset performance comparison of the proposed LDP-TOP with SoA methods in terms of $EER(\%)$.

Method	Idiap	CASIA	MSU	CASIA$_H$
IDA [21]	–	–	8.58	12.9
LDP-TOP	2.50	8.94	6.54	7.45

only methods focusing on the region of the face, thus we discarded the case where the method in [5] was applied on the entire frame.

In cross-dataset test, we followed the cross-dataset testing protocol proposed in [21] by analyzing Idiap versus MSU, and CASIA$_H$ versus MSU. The final result is the average of TPR collected in each iteration of 4-fold validation at $EER = 0.1$. Only the print-attack videos, as suggested in [21], were selected in this experiment. Table 3 reveals that the proposed method performed generally better than the methods using the combined texture features Gabor+HOG+LBP [15] and DoG+LBP [24] which were reported by [21], and less than IDA [21] in cross-dataset test. This can be explained as LDP-TOP is still a data-driven-based method, thus it should be combined with other information to be generalized on cross datasets. However, in the context of practical authentication systems, the real clients' faces are previously captured and learnt by the system, hence the advantage of the proposed method is undeniable in detecting face spoofing.

Table 3. Cross-dataset performance in terms of $TPR(\%)$ at $FAR = 0.1$. The name in each column denotes the name of the training dataset followed by the testing one.

Method	Idiap - MSU	MSU - Idiap	CASIA$_H$ - MSU	MSU - CASIA$_H$
Gabor+HOG+ LBP [15]	24.70	23.80	3.0	6.9
DoG+LBP [24]	15.10	48.80	10.6	4.1
IDA [21]	61.10	69.10	26.9	9.1
LDP-TOP	33.12	45.12	11.88	5.83

Beside the accuracy, the computational cost of anti-spoofing techniques is also important. The approach in [21] requires the whole frame sequence and the average processing time of 0.26 second per frame, then for instance, the total processing time is $0.26 \times 5 \times 25 = 32.5$ seconds for a 5-second-length video ($\approx$ 25 frames per second). In contrast, the proposed method only requires $\approx$ 25 frames per video, and the total average processing time is respectively 1.5, 3, and 4.5 seconds with $R = 1, [1, 2], [1, 3]$ for a video of arbitrary length.

4. CONCLUSIONS

We proposed a novel approach for face spoofing detection using the high-order Local Derivative Pattern from Three Orthogonal Planes. This method outperformed state-of-the-art work on the three common datasets in face spoofing: Idiap, CASIA, and MSU. It is also simple and computationally efficient, thus being suitable for real time processing and low-cost devices. Although the proposed method can be applied in cross datasets, future extension will exploit additional features to overcome the overfitting problem of data-driven-based methods.

5. REFERENCES

[1] L. Sun, G. Pan, Z. Wu, and S. Lao, “Blinking-based live face detection using conditional random fields,” in *ICB*, 2007, vol. 4642, pp. 252–260.

[2] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigün, “Real-time face detection and motion analysis with application in ”liveness” assessment,” *TIFS*, vol. 2, no. 3-2, pp. 548–558, 2007.

[3] W. Bao, H. Li, N. Li, and W. Jiang, “A liveness detection method for face recognition based on optical flow field,” in *IASP*, 2009, pp. 233–236.

[4] A. Anjos, M. Chakka, and S. Marcel, “Motion-based counter-measures to photo attacks in face recognition,” *Biometrics IET*, 2013.

[5] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, “Detection of Face Spoofing Using Visual Dynamics,” *TIFS*, vol. 10, no. 4, pp. 762–777, 2015.

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

[7] T. Freitas Pereira, J. Komulainen, A. Anjos, J. De Martino, A. Hadid, M. Pietikäinen, and S. Marcel, “Face liveness detection using dynamic texture,” *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 2, 2014.

[8] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “3D-Model-Based Video Analysis for Computer Generated Faces Identification,” *TIFS*, vol. 10, no. 8, pp. 1752–1763, 2015.

[9] D.-T. Dang-Nguyen, V. Conotter, G. Boato, and F. G. B. De Natale, “Video forensics based on expression dynamics,” in *GlobalSIP*, 2014, pp. 161–166.

[10] D.-T. Dang-Nguyen, G. Boato, and F. G. B. De Natale, “Identify computer generated characters by analysing facial expressions variation,” in *WIFS*, 2012, pp. 252–257.

[11] V. Conotter, E. Bodnari, G. Boato, and H. Farid, “Physiologically-based detection of computer generated faces in video,” in *ICIP*, 2014, pp. 248–252.

[12] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, “Face liveness detection using 3D structure recovered from a single camera,” in *ICB*, 2013, pp. 1–6.

[13] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, “Face liveness detection by learning multispectral reflectance distributions,” in *FG*, 2011, pp. 436–441.

[14] G. Chetty, “Biometric liveness checking using multimodal fuzzy fusion,” in *FUZZ-IEEE*, 2010, pp. 1–8.

[15] J. Määttä, A. Hadid, and M. Pietikäinen, “Face spoofing detection from single images using texture and local shape analysis,” *IEEE Biometrics, IET*, vol. 1, no. 1, pp. 3–10, 2012.

[16] B. Zhang, Y. Gao, S. Zhao, and J. Liu, “Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor,” *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 533–544, 2010.

[17] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *CVPR*, vol. 1, pp. 511–518, 2001.

[18] C.-C. Chang and C.-J. Lin, “Libsvm: A library for support vector machines,” *TIST*, vol. 2, pp. 27:1–27:27, 2011.

[19] A. Barla, F. Odone, and A. Verri, “Histogram intersection kernel for image classification,” in *ICIP*, 2003, vol. 3, pp. 513–516.

[20] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S.Z. Li, “A Face Antispoofing Database with Diverse Attacks,” in *ICB*, 2012, pp. 1–7.

[21] D. Wen, H. Han, and A. K Jain, “Face spoof detection with distortion analysis,” *TIFS*, vol. 10, no. 4, pp. 746–761, 2015.

[22] I. Chingovska, A. Anjos, and S. Marcel, “Anti-spoofing: Evaluation Methodologies,” in *Encyclopedia of Biometrics*, pp. 41–45. Springer, 2nd edition, 2014.

[23] G. R. Bradski, “Real time face and object tracking as a component of a perceptual user interface,” in *WACV*, 1998, pp. 214–219.

[24] N. Kose and J.-l. Dugelay, “Classification of captured and recaptured images to detect photograph spoofing,” *ICIEV*, pp. 1027–1032, 2012.