SWAPPED! Digital Face Presentation Attack Detection via Weighted Local Magnitude Pattern

Akshay Agarwal*, Richa Singh*, Mayank Vatsa*, Afzel Noore+
*IIIT-Delhi, +West Virginia University
*{akshaya, rsingh, mayank}@iiitd.ac.in, +{afzel.noore}@mail.wvu.edu

Abstract

Advancements in smartphone applications have empowered even non-technical users to perform sophisticated operations such as morphing in faces as few tap operations. While such enablements have positive effects, as a negative side, now anyone can digitally attack face (biometric) recognition systems. For example, face swapping application of Snapchat can easily create "swapped" identities and circumvent face recognition system. This research presents a novel database, termed as SWAPPED - Digital Attack Video Face Database, prepared using Snapchat's application which swaps/stitches two faces and creates videos. The database contains bonafide face videos and face swapped videos of multiple subjects. Baseline face recognition experiments using commercial system shows over 90% rank-1 accuracy when attack videos are used as probe. As a second contribution, this research also presents a novel Weighted Local Magnitude Pattern feature descriptor based presentation attack detection algorithm which outperforms several existing approaches.

1. Introduction

With success of face recognition systems, the research on face presentation attack detection has gained significant impetus. It's importance can be observed from the launch of a multi-million dollar IARPA project on biometric spoofing known as Odin[1]. While it is important to design algorithms for already known face presentation attacks such as print [2], replay [4], and 3D mask [19], it is equally important to test the vulnerabilities of face recognition systems and determine the weak points in the system that can be exploited with different types of attacks. In 2001, Ratha et al. [23] described eight points of attack on a biometric recognition system. Out of eight points, attack on points two, six, and seven can be performed digitally, where already acquired data is modified through software. Presentation attack, according to ISO/IEC IS 30107 [16] is defined as the attack which influences biometric data subsystem with the intention of causing interference in working of the system. Current research in face presentation attack detection (PAD) has primarily focused on physical methods of alterations such as print and silicone masks [19]. However, there are several ways to digitally alter the images as well, for instance, morphing [9] or retouching [3], which are relatively less explored in literature.

Figure 1: Morphing examples - the first row are constituent images and the images in the second row are morphed. Images are taken from http://www.morphthing.com/.

Figure 1 illustrates the effect of morphing. The top row shows original images and the bottom row shows the morphed images. The first two morphed images are created with the first three images and the image of Jennifer Aniston (second image) being common in both the morphed faces. It is interesting to note that the first morphed image shows features of Jennifer but the second image is almost impossible to trace back to Jennifer. This shows that morphing can be used to both elude and create duplicate identity. The effect of face morphing in enrollment was first introduced by International Organization for Standardization ISO 19792[2]. Ferrara et al. [9] in 2014 demonstrated the vulnerabilities of commercial face recognition systems towards morphed images. They also showed that these morphed images are

[1] https://www.iarpa.gov/index.php/research-programs/odin

[2] https://www.iso.org/standard/51521.html

challenging to be detected by face recognition experts as well as automatic algorithms [10].

Earlier, such thorough digital alterations could not be easily performed by non-tech savvy users. However, with recent explosion in the use of smartphones device and popularity of image processing applications such as Snapchat, it has become an item of entertainment for people of all ages and expertise. As per the statistics published by Snapchat[3], in the USA more than 60% of the population in the age range 18-34 years use Snapchat. Snapchat has a functionality of face swapping/switching, which is similar to morphing.

Inspired by the effectiveness of these mobile applications and limitation of face recognition algorithms, this research focuses on designing a novel algorithm to differentiate between digitally attacked images and original/non-tampered images. The contributions of this research are as follows:

- Since there is no publicly available face database for swapping/morphing, we first prepare a new database termed SWAPPED - Digital Attack Video Face Database. The database contains more than 600 videos created with the face swapping/switching feature of Snapchat along with more than 120 real videos.
- We next propose a novel algorithm for effectively differentiating between digital presentation attacks and original non-tampered videos/frames, using the proposed *Weighted Local Magnitude Pattern* feature descriptor.
- The comparison with existing state-of-the-art presentation attack detection algorithms showcases the efficacy of the proposed algorithm for digital presentation attack detection.

2. SWAPPED!! Proposed Digital Attack Database

Existing research on face morphing as presentation attack focuses on images only. To extend the scope of digital attacks on face videos, we present the proposed digital presentation attack database collected as part of this research. This is the first of its kind presentation attack database prepared using one of the most used chat application on mobile devices.

There are several open source algorithms available for creating morphed images, however, Snapchat is one of the most popular and easily accessible tool for morphing or swapping face images. Since it is easy to navigate through the app, non-technology savvy users can also efficiently use it to create various kinds of altered images, swapping being one of the popular ones. A video where a woman swaps her face with Kardashians has been viewed more than $21,000$ times in a week[4]. Even after being easy to operate, the face switch/swap feature is effective enough to change the properties of the face completely that by just looking at the altered face, it is difficult to determine whether it is real or not.

The database is collected using the face switching/swapping feature of Snapchat which works in the following manner[5]: First the face is detected using the Viola-Jones face detector [27]. To make the change more accurate and precise, key point location of the facial features such as eye, mouth, and face boundary are detected using Active Shape Model (ASM) [5]. Once the facial keypoints are detected, a 3D mesh is generated which fits the face properly and can move in real time with changes in face. The facial keypoints are detected from both the faces and the central region is swapped from one image to the other. The boundary is then seamlessly blended to create the new swapped/stitched face image.

The database consists of two parts: bonafide faces and altered faces. Since this research and the Snapchat feature is more prevalent on mobile phones, the bonafide/genuine images are captured using mobile phones. For every user, at least one video of around six seconds is captured using the front camera. In total, 129 bonafide videos are captured from 110 individuals in two months. These videos are captured in unconstrained environment such as natural outdoor, hallway, and inside the office. Faces present in the videos are detected using Viola-Jones face detector and normalized to 296×296 pixels. As summarized in Table 1, after face detection, the bonafide subset contains more than $30,000$ face frames. Figure 2(a) shows sample images from the bonafide set captured in different illumination and background conditions.

To prepare the altered/attacked videos using face swap, two good quality frontal face images of 84 subjects are captured in semi-controlled environment. Samples of these images are shown in Figure 2(b). These images are termed as the input gallery for face swapping/switching. To create a swapped video, Snapchat application requires the users to select the host video/image and an image with which they want to perform the face swap/switch. Using host videos from 31 subjects and overlap images from 84 subjects, 612 presentation attack videos are prepared. Samples of swapped faces are shown in Figure 2(c). Similar to bonafide faces, the altered/attacked faces are normalized to size 296×296. Characteristics of the proposed database are summarized in Table 1.

Table 2 characterizes the existing and the proposed attack research with respect to the input source and environment. As shown, the proposed database is prepared with

[3]http://wallaroomedia.com/snapchat-statistics-updated-2017/

[4]https://tinyurl.com/k6nfly9

[5]https://tinyurl.com/lla8sat

Table 1: Characteristics of the proposed attack database.

Data Type	# Subjects	# Videos	# Detected Faces
Real	110	129	30,728
Attack	31	612	1,04,052

Table 2: Types of attack on face recognition system.

Attack	**Video**	**Unconstrained**	**Image**
Print Attack [2]	✓	✓	×
Replay Attack [4]	✓	✓	×
3D Mask [19]	✓	✓	×
Morphing [22]	×	×	✓
Proposed	✓	✓	✓

unconstrained image sources and contains both images and videos. To promote further research in this important research problem, the database will be released publicly[6].

2.1. Protocol and Performance Metrics

Along with the database, we also define a benchmark protocol that can be used to report and compare results. To summarize, the bonafide subset of the database contains 129 videos from 110 subjects and presentation attack subset contains a total 612 videos from 31 subjects.

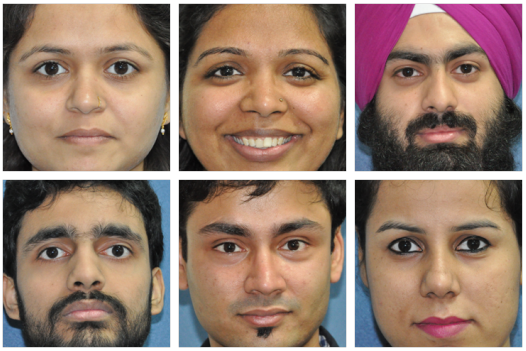
Out of these videos, the real subset is divided into three random folds, where two folds contain 40 videos from 40 subjects. The third fold contains 49 videos from 30 subjects. In the attack subset, the number of videos are large and hence it is divided into 10 folds. Each fold of the attack subset contains 60 videos corresponding to 3 subjects except the last fold which contains 72 videos from 4 subjects. Therefore, for evaluation a total of 3×10 iterations are performed. At any time only one fold is used for training and the remaining folds are used for testing.

The performance of the presentation attack detection features is reported in terms of Equal Error Rate (EER) and Average Classification Error Rate (ACER). EER is defined as the point where the Bonafide Presentation Classification Error Rate (BPCER) is equal to the Attack Presentation Classification Error Rate (APCER). BPCER is the percentage of bonafide faces which are incorrectly classified as attack/altered faces while APCER is the percentage of attack faces which are incorrectly classified as bonafide faces. To calculate the BPCER and APCER on the test set, a threshold value is selected based on the EER of the development set[7]. ACER is then computed as the average of BPCER and APCER.

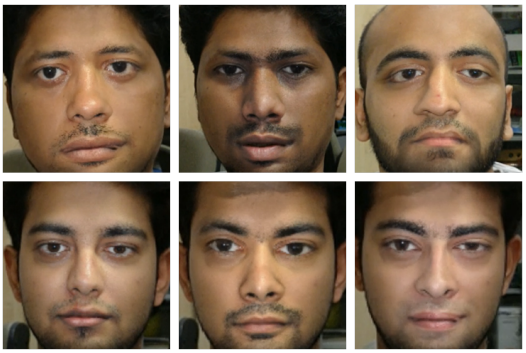
[6]http://iab-rubric.org/resources.html

[7]In this research, half of the training set is used as the development set.

(a) Sample bonafide image set



(b) Sample images used for face overlap



(c) Morphed images from Snapchat

Figure 2: **Sample images from the proposed SWAPPED database.**

$$ACER = \frac{BPCER + APCER}{2} \times 100 \qquad (1)$$

3. Effect of Swapping Attack on Face Recognition

To evaluate the effectiveness of the face swap feature as an attack on the face recognition system we have performed two different experiments: 1) Various iOS devices are now equipped with the face unlock feature. Thus, the first experiment is face unlocking on iPhone and 2) face identification

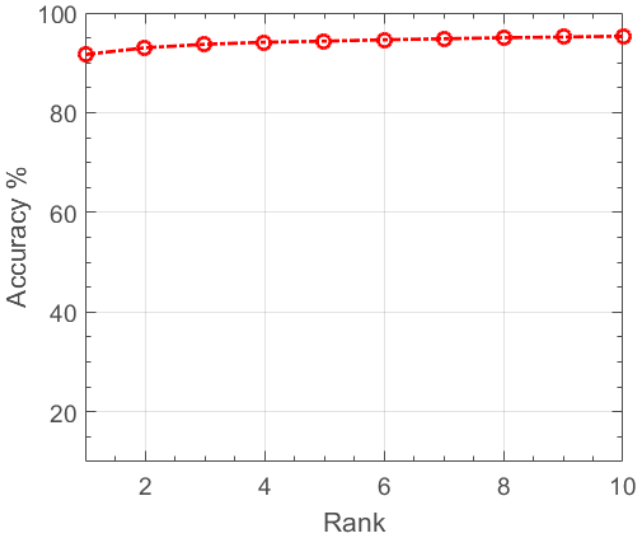


Figure 3: CMC plot for face identification using COTS.

using a Commercial-Off-The-Shelf System (COTS), FaceVACS[8]. In the first experiment to unlock the iPhone, video of the swapped face prepared using an image of the genuine person who is enrolled in the mobile device is displayed in front of the mobile camera. It is interesting to observe that the face recognition algorithm in the iPhone is unable to detect the attack and hence unlocks every time. This shows the vulnerability of face recognition in mobile devices to digital attacks.

In the second experiment, face identification is performed using a COTS system. Another set of frontal images is collected from the individuals whose images are used for creating the swapped videos. These images comprise the gallery for face identification. From each of the attack videos, 30 random frames are used as the probe for face identification experiment. Figure 3 shows the CMC curve obtained for this experiment. The results show that 90% of the time, attack images are matched to enrolled gallery images at rank-1.

4. Proposed Digital Presentation Attack Detection Algorithm

It is our assertion that digital alterations generally perform smoothing and blending to minimize the irregularities due to the differences in the source frame. This reduces the difference in neighboring pixel values. Ojala et al. [20] have reported that sometimes more than 90% of the texture surfaces are uniform. Based on this intuition, we propose a new attack detection algorithm for detecting digital attacks. The algorithm is based on a novel feature encoding method termed as *Weighted Local Magnitude Patterns*. Similar to local binary pattern (LBP), the proposed descriptor encodes the differences between a center pixel and it's neighbors. However, instead of binarizing them, it assigns the weights inversely in proportion to the difference from the center pixel.

Figure 4 illustrates the steps involved in the proposed attack detection algorithm. The input image is first tessellated into multiple blocks of size 3×3. For each patch, the absolute difference between the center pixel and its neighborhood pixels are calculated. Since there are eight neighborhood pixels, there are eight difference values. The difference values are sorted in ascending order. Instead of binarizing the absolute differences, the sorted values are multiplied with 2^p, where $p = 0, ..., 7$ for 8 neighborhood values.

The motivation for sorting and multiplying is to give higher weight to the pixel which has a value similar to the center pixel. The final output value is then mapped to a value in the range of 0 to 255 (i.e., any value greater than 255 is set of 255). Finally, a histogram feature vector is calculated. The output images using the proposed feature descriptor are shown in Figure 5 along with the corresponding output obtained by LBP. It can be observed that the output images of the proposed feature/algorithm retains the high-frequency information while reducing the low-frequency information. With images obtained from Snapchat's swapped/switched feature, facial keypoint regions such as eye, mouth, and nose are the most affected while center region is well blended. This is clearly highlighted in the output images of the proposed algorithm. Therefore, we postulate that for morphing related attacks, the proposed feature is better at detecting alterations than existing feature descriptors. The extracted feature vectors from training data are provided to the supervised Support Vector Machine (SVM) [26] classifier to learn the presentation attack detection model.

5. Experiments

The performance of the proposed algorithm is demonstrated on the SWAPPED digital attack video face database. Various methods are proposed for image tampering and watermark detection such as JPEG ghosts [8] and JPEG double quantization patterns [24]. In the case of tampering, the image retains most of the original information, while in the proposed digital presentation attack, original image content changes completely to different content. Literature of face presentation attack detection [12, 25] has shown state-of-the-art performance using texture features. Therefore, we have compared the performance of the proposed algorithm with seven different textural feature based algorithms: LBP* [18], Rotation Invariant Uniform LBP (RIULBP)* [20], Complete LBP

[8]http://www.cognitec.com

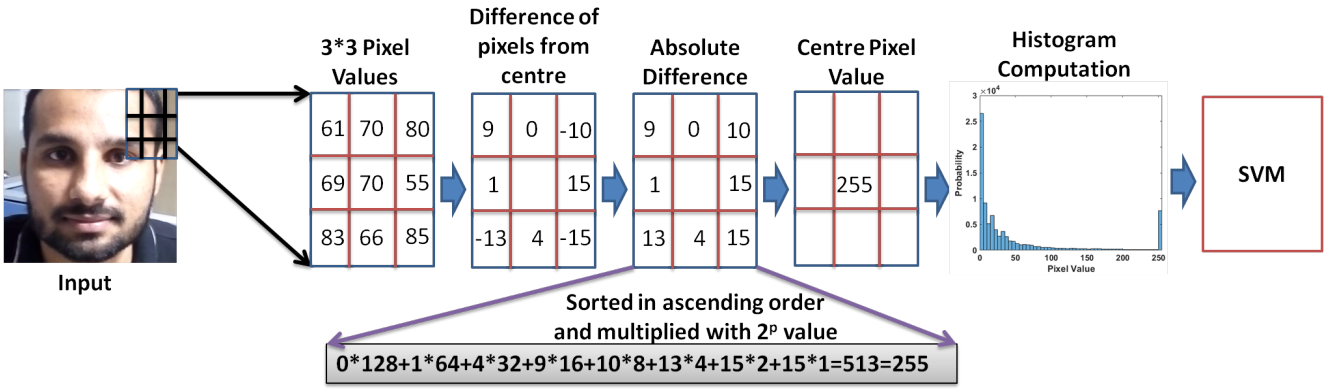


Figure 4: Illustrating the steps involved in the proposed attack detection pipeline.

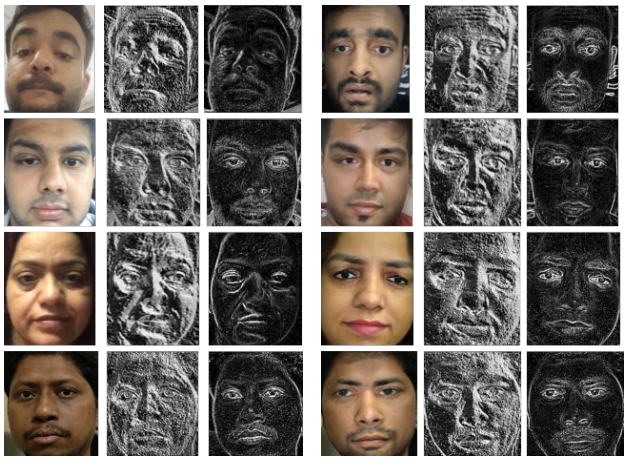


Figure 5: Illustrating the features obtained for real and altered samples. (a) input image, (b) LBP feature, and (c) proposed feature image.

(CLBP)* [14], Uniform LBP (ULBP)*, LPQ [21], BSIF [17], and Combination of Redundant Discrete Wavelet Transform (RDWT) [11] with Haralick [15] proposed in [1]. The codes of starred algorithms are taken from http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab.

5.1. Results and Analysis

The protocol defined in Section 2.1 is used in the experiments. Since the proposed database contains videos, the results can be measured both in terms of video classification and frame classification. In case of videos, the entire video is classified as bonafide or attack whereas for frame based, every frame is classified as bonafide or attack. The score of the video is calculated as the average of all the scores corresponding to frames of that video.

Table 3: Average classification results (%) of the proposed and existing algorithm for video and frame based presentation attack detection on the proposed database.

Input	Features	EER	ACER
Video	LBP* [6]	21.7 ± 6.1	21.3
	ULBP* [7]	24.5 ± 6.0	22.7
	RIULBP* [20]	24.7 ± 4.9	23.5
	CLBP* [13]	24.5 ± 6.1	24.8
	Haralick+RDWT [1]	25.6 ± 7.2	24.5
	BSIF [22]	25.2 ± 9.1	24.9
	LPQ [18]	22.9 ± 5.2	23.9
	Proposed	$\mathbf{18.2 \pm 5.6}$	**18.1**
Frame	LBP* [6]	27.1 ± 4.3	27.3
	ULBP* [7]	29.0 ± 3.4	28.6
	RIULBP* [20]	28.7 ± 3.7	28.7
	CLBP* [13]	28.7 ± 3.8	28.8
	Haralick+RDWT [1]	28.9 ± 4.8	28.4
	BSIF [22]	30.2 ± 7.0	30.2
	LPQ [18]	28.7 ± 4.0	30.4
	Proposed	$\mathbf{24.5 \pm 5.1}$	**25.4**

Table 3 and Figure 6 show the results of the proposed and existing features for digital presentation attack detection by face swapping. The analysis of the results in Table 3 are listed below:

- The proposed features yield the average EER value of 18.2% and 24.5% for video and frame based attack detection respectively. While the bonafide presentation classification error rate is 6.2% which is the least among all the algorithms, the attack presentation classification error rate is 29.5%.
- The proposed features show an improvement of 16% in terms of EER from the second best-performing feature i.e. LBP for video based attack detection.

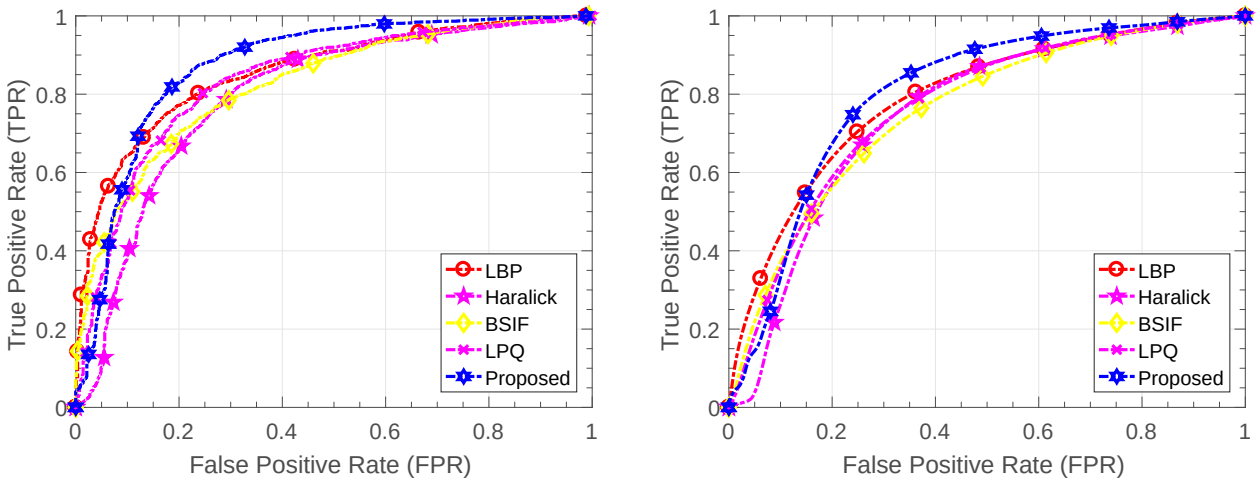


Figure 6: ROC curve for video (left) and frame (right) based presentation attack detection on the proposed database.

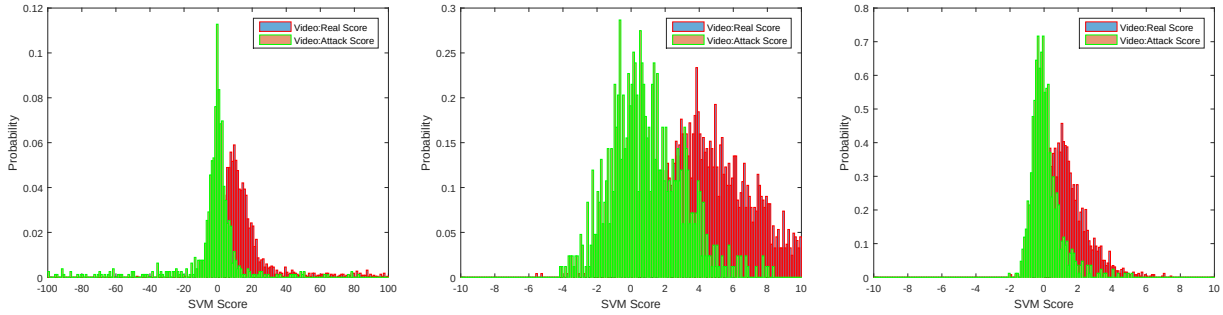


Figure 7: SVM score distribution of real and attack videos: (a) Proposed, (b) LBP, and (c) LPQ.

- It is interesting to observe that the combination of Haralick+RDWT, which yields lowest EER on physical spoofing database, provides the highest EER value of 25.6% in video based attack detection. In the case of frame based attack detection BSIF feature yields the lowest performance.

- The proposed feature histogram incorporates sorting in ascending order. However, when the difference values are sorted in descending order, the EER increases from 18.2% to 25.5% and from 24.5% to 29.3% for video and frame based detection, respectively.

- LBP histogram feature provides the second lowest ACER value both for video and frame based detection. Analyzing the results of correctly classified and misclassified samples of both the algorithms show that the proposed algorithm tends to misclassify bonafide samples that contain spectacles. Figure 7 shows the SVM score distribution of bonafide and attack videos.

6. Conclusion

This paper extends the research of presentation attack from physical attacks to digital attacks and presents a database of swapped faces prepared using Snapchat application. The proposed face attack shows the vulnerability of the face recognition system both in mobile phones and commercial system. A new digital presentation attack detection algorithm is proposed using a novel descriptor, termed as Weighted Local Magnitude Patterns. The proposed algorithm achieves lower error rates compared to existing texture feature based approaches. As future research, we plan to extend the proposed algorithm to further reduce the errors, particularly by incorporating temporal information in presentation attack detection.

7. Acknowledgement

This research is partially supported by MEITY, Government of India and Visvesvaraya PhD Fellowship.

References

[1] A. Agarwal, R. Singh, and M. Vatsa. Face anti-spoofing using Haralick features. In *8th International Conference on Biometrics Theory, Applications and Systems*, 2016.

[2] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *International Joint Conference on Biometrics*, Oct. 2011.

[3] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, Sept 2016.

[4] I. Chingovska, A. Anjos, and S. Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *International Conference of the Biometrics Special Interest Group*, 2012.

[5] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[6] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. LBP-TOP based countermeasure against face spoofing attacks. In *Asian Conference on Computer Vision*, pages 121–132. Springer, 2012.

[7] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *International Conference on Biometrics*, 2013.

[8] H. Farid. Exposing digital forgeries from jpeg ghosts. *IEEE Transactions on Information Forensics and Security*, 4(1):154–160, 2009.

[9] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *International Joint Conference on Biometrics*, 2014.

[10] M. Ferrara, A. Franco, and D. Maltoni. *On the Effects of Image Alterations on Face Recognition Accuracy*, pages 195–222. Springer International Publishing, 2016.

[11] J. E. Fowler. The redundant discrete wavelet transform and additive noise. *IEEE Signal Processing Letters*, 12(9):629–632, 2005.

[12] J. Galbally, S. Marcel, and J. Fierrez. Biometric antispoofing methods: A survey in face recognition. *IEEE Access*, 2:1530–1552, 2014.

[13] Y. Guo, G. Zhao, and M. Pietikinen. Discriminative features for texture description. *Pattern Recognition*, 45(10):3834 – 3843, 2012.

[14] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[15] R. M. Haralick and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, (6):610–621, 1973.

[16] Information technology - biometric presentation attack detection. Standard ISO/IEC 30107-1:2016 - Part 1 - Framework.

[17] J. Kannala and E. Rahtu. BSIF: Binarized Statistical Image Features. In *21st International Conference on Pattern Recognition*, pages 1363–1366, 2012.

[18] J. Määttä, A. Hadid, and M. Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *International Joint Conference on Biometrics*, 2011.

[19] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security*, 2017.

[20] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[21] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. In *International Conference on Image and Signal Processing*, pages 236–243. Springer, 2008.

[22] R. Raghavendra, K. B. Raja, and C. Busch. Detecting morphed face images. In *IEEE 8th International Conference on Biometrics Theory, Applications and Systems*, 2016.

[23] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614–634, 2001.

[24] J. A. Redi, W. Taktak, and J.-L. Dugelay. Digital image forensics: a booklet for beginners. *Multimedia Tools and Applications*, 51(1):133–162, 2011.

[25] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha. Face anti-spoofing with multifeature videolet aggregation. In *23rd International Conference on Pattern Recognition*, pages 1035–1040, 2016.

[26] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[27] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.