Presentation Attack Detection by a Combination of Intrinsic Image Properties and a CNN

[1]Rodrigo Bresan, [2]Edmar Rezende, [1]Carlos Beluzo, [1]Tiago Carvalho
[1]Federal Institute of São Paulo (IFSP), Campinas, SP, Brazil
[2]University of Campinas (UNICAMP), Campinas, SP, Brazil
{rcbresan, edmar.rezende}@gmail.com, {cbeluzo, tiagojc}@ifsp.edu.br

***Abstract*—The usage of face recognition for biometric systems has become widely adopted, since it allows the usage of a trait that is accessible to most of the people. Despite important progress on the field of face recognition, there is still a lack of works whose focus consists on the detection of presentation attacks. Presentation attacks occur when an imposter presents a synthetic sample in order to impersonate a valid user. For face biometric systems, this kind of attack is performed using a photograph, by playing a video of the user (commonly known as replay attack) or by making usage of 3D masks. Hereby, we propose a low-cost solution to detect these kind of attacks without the need of extra hardware. Our hypothesis is based on the fact that, through the extraction of intrinsic image properties, such as depth, saliency and illumination, it is possible to distinguish between a real biometric sample and a synthetic one. Performed experiments show that the proposed method achieved HTER values of 41.64% and 3.88% in inter and intra protocols respectively, achieving near state-of-the-art results.**

I. Introduction

The adoption of physiological traits (i.e. face, iris and fingerprint) or behavioral characteristics (i.e. gait, typing rhythm), in order to identify or authenticate an individual, is denoted as biometrics. With the increasing adoption of biometric systems worldwide, from personal devices such as laptops and smartphones up to access to restricted areas, major challenges were posed in order to develop methods that are capable of distinguishing a real biometric sample from a synthetic one. The action of presenting a synthetic biometric sample to the acquisition sensor, in order to obtain access as a legitimate user is known in literature as presentation or spoofing attack.

The approach proposed in this work consists in building a low-cost method, without the need of extra hardware, by exploiting many of the intrinsic characteristics (such as depth, light and saliency properties) from a given biometric sample, as depicted in Figure 1. Our hypothesis is based on the assumption that these characteristics may contain telltales that indicate if a given biometric sample is real or not. Associated with a Convolutional Neural Network (CNN) for feature extraction and with a Support Vector Machine (SVM) for classification, our method is able to achieve close to state-of-the-art results without the necessity of laborious handcraft feature extraction step.

The major contributions of this work may be highlighted as follows: (1) usage of characteristics not yet explored, such as illumination and saliency in PAD; (2) decrease of laborious work demanded by handcrafted features extraction; (3) an HTER value of 3.88% and 41.64% for intra-dataset and inter-dataset evaluation protocols, respectively[1].

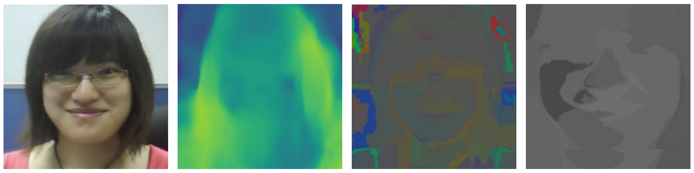
(a) Raw (b) Depth (c) Illumination (d) Saliency

Fig. 1. Different types of intrinsic properties estimated from a single frame.

II. Related Works

Techniques for presentation attack detection methods can be categorized in four major groups, according to Pan et al. [1]: user cooperation, user behavior modeling, data-driven and with the usage of additional hardware. The first approach is based on the cooperation between the user and the given authentication system, such as requesting the user to perform some specific movements, but at the cost of adding an additional time in the authentication process and also leaving behind some of the naturality of it. Methods that are based on the second approach rely on the user behavior itself (e.g. eye blinking, head movement) to detect fraudulent attacks. The existing methods based on this kind of approach are highly susceptible to fail when presented with video attacks (commonly referred as replay attacks). Techniques that are based on the data-driven characterization aim to find evidences that have correlation with an attack attempt, such as common local features. Finally, methods that use additional hardware (e.g. infrared cameras, depth sensors) aim to obtain more details about the scenery and thus be able to detect cues of a fraudulent access.

Specifically in data-driven group, Pinto et al. [2] proposed a method for detecting replay attacks through the analysis of visual rhythm. Other works have also been proposed exploring Common Local Features, such as Local Binary Patterns proposed by Maat et al. [3] in order to capture micro-texture patterns added in the fraudulent biometric sample during its acquisition, as well as HOG [4], [5] and DoG [6], [7], but

[1] The repository containing the algorithms used in this work is freely available at https://github.com/bresan/SpooPy. There you can find all the tools used for pre-processing, log files and reported results.

due to their nature, their results may be highly affected by illumination settings and camera devices. Schwartz et al. [8], through the extraction of face characteristics (such as color, texture and shape), proposed a method using Partial Least Squares (PLS) classifier in order to decide whether a given sample is genuine or not.

III. Proposed Method

To distinguish between real biometric samples and synthetic ones, we propose a new method based on the assumption that image intrinsic properties (such as depth, light properties and saliency) can provide relevant information for detecting presentation attacks.

Once these properties are extracted for each frame, we take advantage of transfer learning techniques using the ResNet50 architecture as a feature extractor, resulting in a set of bottleneck features. Finally, those features are used to feed a Machine Learning (ML) classifier, in order to detect if a given biometric sample is authentic or not. By classifying each frame as being authentic or not, we finally use a majority voting technique to decide whether the input sample is authentic. Figure 2 depicts an overview of proposed method.

A. Image Intrinsic Characteristics

After a simple pre-processing, where we crop the videos around presented faces and register all video frames (to provide a better alignment), the first step of the proposed method consists in estimate different types of representation for a given input, in a way to highlight different intrinsic characteristics of image.

1) Illuminant Maps: As proposed by Carvalho et al. [9], the usage of illumination maps is an effective indicator of photo editing. In the same way, our hypothesis is that a presentation attack tends to present a different illumination pattern when compared to valid access. Using Inverse Intensity-Chromaticity Space, the illuminant map from a given image can be calculated by the following equation, as proposed by Tan et al. [10]:

$$\chi_c(x) = m(x) \frac{1}{\sum_{i \in \{R,G,B\}} f_i(x)} + \gamma_c, \qquad (1)$$

where γ_c denotes the chromaticity of the illuminant in channel c, whereas $m(x)$ mainly captures geometric influences (i.e. surface orientation, camera and light position).

2) Depth Maps: As in illumination case, our hypothesis is that, when capturing the sample, if it is provided by an attack attempt, the depth will be different from a valid access mainly because it will be captured from a flat surface. This way, we estimate depth maps and use them as a way to provide a second set of intrinsic properties. The task of depth estimation in single images have shown very promising results with the usage of learning based methods. In the work presented by Godard et al. [11], a novel training method is proposed in order to perform single image depth estimation, without the need of ground truth depth data. This approach also uses a novel training loss, to deal with image reconstruction loss. As result, this method presented a higher performance when compared to previous works, showing best results even when compared to works that were trained using ground truth depth.

3) Saliency Maps: Similarly to depth maps, saliency maps are an additional intrinsic information evaluated in this work. Here, we estimate saliency maps using the method proposed by Zhu et al. [12], which proposes a robust background measure, called boundary connectivity. This measure is used to characterize the spatial layout of a given image regions with respect to its boundaries, showing a much higher robustness. Zhu et al. [12] also proposed a framework to integrate multiple low level cues, along with the background measure, in order to obtain clean and uniform saliency maps.

B. Bottleneck Features Extraction via ResNet50 Architecture and Transfer Learning

Once the intrinsic properties are extracted, the next step of the proposed method is to encode them into useful features, to be used individually as input to a machine learning classifier.

Instead of standing in laborious handcraft feature extraction process, our method take advantage of a combination between a robust Convolutional Neural Network (CNN) architecture and the transfer learning method.

As CNN architecture, we chose ResNet50 [13] architecture, which have presented very promising results into different tasks.

On the other hand, with the usage of transfer learning techniques, which consists in transferring the weights of a previously trained neural network [14], we avoid the necessity of training the whole network from scratch.

C. Top Classifier

By the end of the feature extraction step, we will have as artifact a 2,048 dimension feature vector for each frame. These features vectors for each of the extracted properties are then individually used as input for a SVM classifier [15]. With the results of the classifier, a fusion approach based on majority voting is used to obtain a final score to decide if a biometric sample is a presentation attack or a genuine access.

IV. Experiments and Results

In order to validate proposed method, different rounds of experiments were conducted using two public datasets containing both valid access attempts and presentation attacks. Besides that, two evaluation protocols, intra-dataset and inter-dataset, were also evaluated in order to address the efficacy of the method.

A. Datasets

In this work, two widely used datasets from literature were chosen to evaluate the proposed method.

1) CASIA: Proposed by Zhang et al. [16] and containing 600 videos, this dataset was created with the purpose of providing a diverse collection with many of the presentation attack types available. The attack videos were created from the genuine ones, in high resolution, simulating three different types of attacks: normal print attacks, print warped and, print with cut on eyes and video-based.

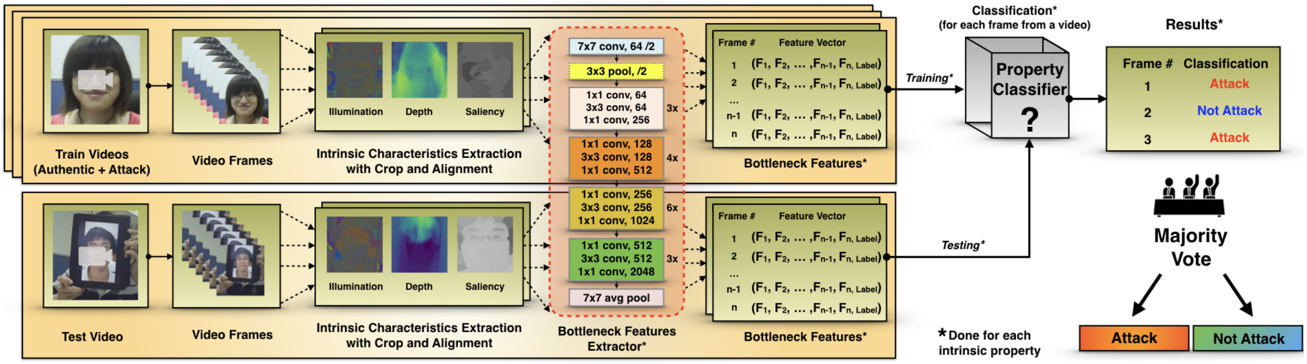

Fig. 2. Overview of proposed method.

2) Replay-Attack: The Replay-Attack Database consists of 1300 video clips of photo and video attack attempts to 50 clients, under different lighting conditions [17]. Three different types of presentation attacks are provided: video, mobile and print attacks, composing three different subsets: training (360 videos), development (360); testing (480 videos); and enrollment (100).

B. Experimental Protocols

The selected protocol for measuring the performance of the proposed method is made by using two of the recommended metrics by ISO/IEC 30107-3 [18]: (1) Bona fide Presentation Classification Error Rate (BPCER) and (2) Attack Presentation Classification Error Rate (APCER). The usage of both metrics is consolidated into two new ones: Equal Error Rate (EER) and Half Total Error Rate (HTER), since its massive adoption in the literature. The HTER value is calculated by the average of both APCER and BPCER measures, while the EER value is calculated by the threshold when the average of APCER and BPCER have the same value.

For each one of the selected datasets, two approaches were used for evaluation: intra-dataset, where the presented method was evaluated within the same dataset; and inter-dataset, where one dataset was used for training, and another different dataset was used for test.

C. Experimental Setup

Proposed method have been implemented using Python 3.4 with Keras 2.2.0 [2] and TensorFlow 1.0.1 [3]. All the experiments have been performed in a virtual machine (VMWare), inside a shared container data center. Our virtual machine setup has 16 cores processors, 100 GB of RAM and 1 TB of storage, with Ubuntu 16.04.4 LTS Desktop 64 bits installed.

D. Intra-Dataset Evaluation

In this section, the results of the intra-dataset evaluation are presented for each of the datasets. The protocol recommended by each author was used to evaluate the performance of the method proposed in this work.

TABLE I
INTRA-DATASET RESULTS (HTER VALUE) FOR THE CASIA DATASET

Attack Type	Illumination	Depth	Saliency
Print	8.33	13.05	**6.38**
Tablet	**5.00**	10.27	**5.00**
Cut	10.00	17.50	**6.94**
Overall	**3.88**	33.33	14.81

TABLE II
INTRA-DATASET RESULTS (HTER VALUE) FOR THE REPLAY ATTACK DATASET

Attack Type	Illumination	Depth	Saliency
Print	**1.25**	16.87	11.87
Highdef	**1.56**	30.83	10.62
Mobile	**0.62**	22.81	5.83
Overall	**5.50**	31.62	12.52

1) CASIA Dataset: Table I presents obtained results for three different attacks. Our overall result (using all attack types in test set) was obtained using the illumination maps, with an HTER value of 3.88%. Using this property we also obtained HTER values of 8.33% for printing attacks individually and 5.00% for tablet attacks. These results confirm our hypothesis that the illumination map from a given authentic biometric sample differs from a synthetic one. For saliency properties we also attained expressive results, with an HTER value of 5.00% on the video-based attacks when reproduced with a tablet.

2) Replay Attack Dataset: Table II presents results for intra-dataset protocol using Replay Attack dataset. Using illumination maps for mobile attacks, proposed method obtained the smallest HTER for the intra-dataset evaluation, with the HTER value of 0.62%. Expressive results were also achieved using the illumination maps for the print and high definition attacks, with HTER values of 1.25% and 1.56%, respectively. As overall result, the best performance was attained using the illumination properties, with an HTER value of 5.50%. Using the saliency properties, an HTER value of 5.83% was obtained for the attacks reproduced through a mobile phone.

[2] https://keras.io
[3] https://www.tensorflow.org

TABLE III
PERFORMANCE RESULTS (HTER VALUE) FOR INTER-DATASET EVALUATION

Methods	Replay Attack	CASIA
Pinto et al. [19]	49.72	47.16
Yang et al. [20]	41.36	42.04
Patel et al. [21]	31.60	-
Depth	48.00	43.33
Saliency	47.78	52.79
Illumination	41.64	50.18

E. Inter-Dataset Evaluation

For biometric systems that make usage of face characteristics, the ability of being adaptable from one given dataset to another is crucial for real world applications. In this section, we present the obtained results for the inter-dataset evaluation protocol, when one dataset was used for training and another was used for test. Table III presents achieved results.

For the Replay Attack, the best results were obtained when using the illumination maps, achieving an HTER of 41.64% when trained on the CASIA dataset, followed by HTER values of 48.00% and 47.48% for depth and saliency maps, respectively.

Results reaching near state-of-the-art approaches were achieved on the CASIA dataset, achieving an HTER value of 43.33% when using the depth maps, the second best when compared to previous methods [19], [20]. For the saliency and illumination maps, HTER values of 52.70% and 50.18% were attained, respectively.

V. CONCLUSION AND FUTURE WORK

In this work, three different intrinsic properties (depth, illumination and saliency) from a given biometric sample were evaluated in order to detect a presentation attack. Taking advantage of transfer learning techniques and a robust CNN architecture, the proposed method was capable of reaching near state-of-the-art results in different scenarios, with an HTER of 3.88% and 5.50% for intra-dataset evaluation on the CASIA and Replay Attack Datasets, respectively, when using the illumination properties.

On the evaluation of inter-dataset protocols, which is the most challenging one in the literature, close to state-of-the art results were achieved for the CASIA using depth maps, with an HTER of 43.33%. For the Replay Attack dataset, when using the illumination maps, we attained an HTER value of 41.64%.

For future works, the study of new properties that may reveal cues for PAD is intended, as well as the evaluation of fusion approaches, in order to provide better results and insights on techniques to recognize these kind of attacks.

ACKNOWLEDGMENT

We would like to thank the financial support Initiation Program (PIVICT, IFSP - Campinas), to São Paulo Research Foundation (FAPESP)(2017/12631-6), to the National Council for Scientific and Technological Development - CNPq (423797/2016-6), and to NVIDIA for the donation of a TITAN XP GPU to be used on this research.

REFERENCES

[1] G. Pan, Z. Wu, and L. Sun, "Liveness detection for face recognition," in *Recent Advances in Face Recognition*, K. Delac, M. Grgic, and M. S. Bartlett, Eds. Rijeka: IntechOpen, 2008, ch. 9. [Online]. Available: https://doi.org/10.5772/6397

[2] A. d. S. Pinto, H. Pedrini, W. Schwartz, and A. Rocha, "Video-based face spoofing detection through visual rhythm analysis," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, Aug 2012, pp. 221–228.

[3] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 International Joint Conference on Biometrics (IJCB)*, Oct 2011, pp. 1–7.

[4] J. Komulainen, A. Hadid, and M. Pietikäinen, "Context based face anti-spoofing," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, Sept 2013, pp. 1–8.

[5] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *2013 International Conference on Biometrics (ICB)*, June 2013, pp. 1–6.

[6] B. Peixoto, C. Michelassi, and A. Rocha, "Face liveness detection under bad illumination conditions," in *2011 18th IEEE International Conference on Image Processing*, Sept 2011, pp. 3557–3560.

[7] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Proceedings of the 11th European Conference on Computer Vision: Part VI*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 504–517. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888212.1888251

[8] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *2011 International Joint Conference on Biometrics (IJCB)*, Oct 2011, pp. 1–8.

[9] T. J. d. Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. d. R. Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1182–1194, July 2013.

[10] R. T. Tan, K. Ikeuchi, and K. Nishino, "Color constancy through inverse-intensity chromaticity space," in *Digitally Archiving Cultural Objects*. Springer, 2008, pp. 323–351.

[11] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[12] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 2814–2821.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *CoRR*, vol. abs/1411.1792, 2014. [Online]. Available: http://arxiv.org/abs/1411.1792

[15] N. M. Nasrabadi, "Pattern recognition and machine learning," *Journal of electronic imaging*, vol. 16, no. 4, p. 049901, 2007.

[16] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face anti-spoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*, March 2012, pp. 26–31.

[17] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," 2012.

[18] "Information technology – Biometric presentation attack detection – Part 3: Testing and reporting," International Organization for Standardization, Geneva, CH, Standard, Mar. 2017.

[19] A. Pinto, H. Pedrini, M. Krumdick, B. Becker, A. Czajka, K. W. Bowyer, and A. Rocha, "Counteracting presentation attacks in face, fingerprint, and iris recognition," *Deep Learning in Biometrics*, p. 245, 2018.

[20] J. Yang, Z. Lei, and S. Z. Li, "Learn Convolutional Neural Network for Face Anti-Spoofing," *ArXiv e-prints*, Aug. 2014.

[21] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Biometric Recognition*, Z. You, J. Zhou, Y. Wang, Z. Sun, S. Shan, W. Zheng, J. Feng, and Q. Zhao, Eds. Cham: Springer International Publishing, 2016, pp. 611–619.