Enhancing Computer Vision to Detect Face Spoofing Attack Utilizing a Single Frame from a Replay Video Attack Using Deep Learning

Aziz Alotaibi, Ausif Mahmood
University of Bridgeport
CT 06604 USA
e-mail: aalotaib@my.bridgeport.edu, mahmmod@bridgeport.edu

***Abstract*—Recently, automatic face recognition has been applied in many web and mobile applications. Developers integrate and implement face recognition as an access control into these applications. However, face recognition authentication is vulnerable to several attacks especially when an attacker presents a 2-D printed image or recorded video frames in front of the face sensor system to gain access as a legitimate user. This paper introduces a non-intrusive method to detect face spoofing attacks that utilize a single frame of sequenced frames. We propose a specialized deep convolution neural network to extract complex and high features of the input diffused frame. We tested our method on the Replay Attack dataset which consists of 1200 short videos of both real-access and spoofing attacks. An extensive experimental analysis was conducted that demonstrated better results when compared to previous static algorithms results.**

***Keywords*-face liveness detection; face detection; spoofing detection; replay attack dataset; anti-spoofing attacks**

I. INTRODUCTION

With the increased use of online authentication systems, automatic face recognition has attracted many developers to integrate and implement face biometrics as an access control into several web and mobile applications. Many computers and mobile phones have a built-in, a front facing camera, and; therefore, does not require additional sensor devices. In contrast, other biometric characteristics require an extra sensor such as fingerprint or iris scanning. Face recognition authentication requires no touch contact with any device which in the users' experience is convenient. However, face recognition authentication is vulnerable to several attacks that use print, video, or 3D mask of a valid user. Attackers can easily spoof the system by downloading a photo from the internet or by capturing a photo. Also, the attacker can penetrate the system by replaying a recorded video in front of a camera. In an attempt to address this issue, many researchers have proposed different methods to prevent the spoofing attacks.

Such anti-spoofing approaches are classified into two groups: the static approach and the dynamic approach.

The static approach is based on the analysis of a single static photo. In contrast, the dynamic approach is based on analyzing the temporal and spatial features of a sequence of input frames.

The remainder of this paper is organized as follows: Previous studies on face liveness detection are reviewed in Section II. Our proposed method is described in Section III. The discussion and performance evaluation on the Replay Attack dataset are presented in Section IV. Finally, our conclusions and a discussion of future work are presented in Section V.

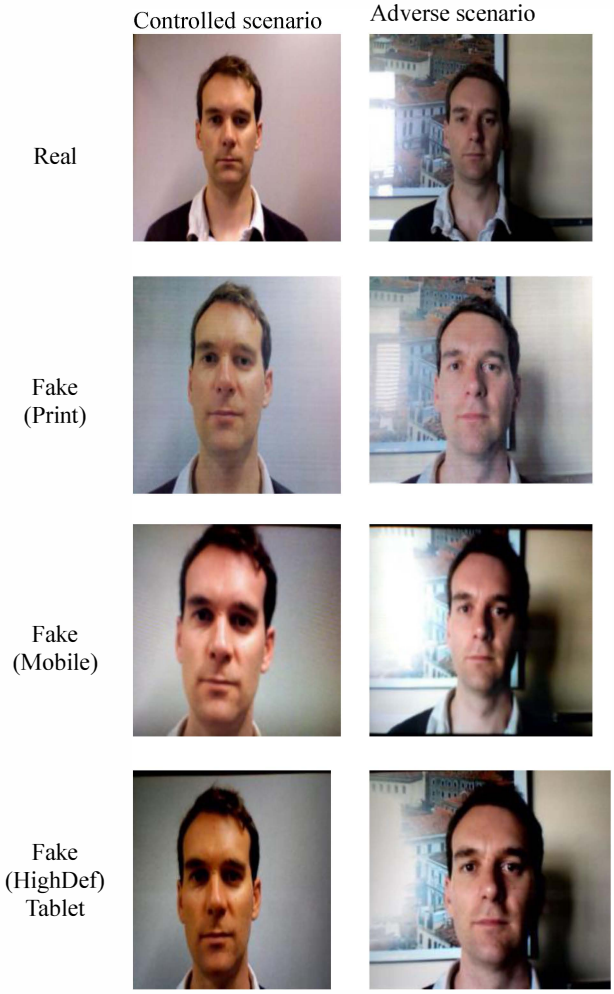

Figure 1. Examples of replay attack dataset (controlled and adverse scenario).

II. RELATED WORK

Face liveness detection approaches are classified into two main groups: the static approach and the dynamic approach.

978-1-5090-0880-3/16/$31.00 ©2016 IEEE

A. *Static Techniques*

Several methods are proposed to address the spoofing attack problems that utilize a single static photo. Static approaches are based on an analysis of a 2D single photo. Also, it is non-intrusive interaction which is convenient for most users. Tan et al. [1] proposed two techniques to extract the most significant features which are an extension to the sparse logistic regression model. Li et al. [2] analyzed the structural texture of the 2D image and of the 3D image to detect the spoofing attacks. Authors observed that the reflections of light on 2D and 3D surfaces result in different frequency distributions. In [3], Peixoto et al. detected the spoofing attacks by applying a Different of Gaussian (DoG) filter that consists of two Gaussian filters with different standard deviations to remove lighting variations and noise from the 2D image input. Zhang et al. [4] used multiple DoG filters to extract the high-frequency features from the input image to distinguish a fake image from a real image. Maatta et al. [5] analyzed the 2D image using the multi-scale local binary pattern (LBP) to extract the texture information . Then authors computed a concatenated histogram that was fed into a support vector machine (SVM) classifier to detect face liveness. Chingovska et al. [6] applied features method based on the LBP operator and their variation to capture the textural information of the input image. Moreover, Kim et al. [7] computed the diffusion speed of a 2D single image to extract the difference in the illumination characteristics of both 2D and 3D input images. Authors applied a local speed pattern (LSP) to extract information features that were fed into a linear SVM classifier. Yang et al. [8] proposed a component-based face recognition coding approach where the authors extracted the micro-textures from twelve regions of the facial components to detect the spoofing attacks.

B. *Dynamic Techniques*

Dynamic approaches are based on utilizing spatial and temporal features using more than one frame of the input image, typically, these approaches are slow and more computationally expensive. Further, some of the dynamic techniques rely on the users to follow some instructions, but not all users may cooperate in this respect. Pan et al. [9] recognized the behavior of spontaneous eye blinking using a non-intrusive method based on a unidirectional conditional graphic framework to detect spoofing attacks. Singh et al. [10] proposed a framework to detect face liveness by identifying eye and mouth movements using the Haar classifier. Kim et al. [11] utilized camera focus to capture variations between pixel values from two sequential images. Bharadwaj et al. [12] presented a new spoofing detection framework for video sequences using motion magnification. They proposed a configuration LBP and motion estimation to extract the features. Tirunagari et al. [13] used a recently developed mathematical method called the dynamic mode decomposition (DMD) algorithm to capture and extract dynamic visual information from the input video. The information features in the dynamic visual are obtained using the LBP, and then fed into a SVM classifier.

III. PROPOSED METHOD

The motivation behind our approach is to detect the spoofing attacks using a nonlinear diffusion based on the additive operator splitting (AOS) scheme with a large time size to obtain the sharp edges and preserve the boundary locations of the input image. Our proposed method utilizes only one frame from sequenced frames of replayed video attacks to detect spoofing attacks. The single frame is captured from different mediums such as print photographs, mobile screens, and high definition (Tablet) screens that requires less processing times. Due to the recent success reported in [7], we applied a non-linear diffusion to extract shape edges and corners contained in the input frames. Then, we used a specialized deep convolution network to detect the texture surface and the edges in order to distinguish a fake image from a real image.

A. *Nonlinear Diffusion*

Nonlinear diffusion is applied to obtain the sharp edges and preserve the boundary locations. Unlike linear diffusion that blurs important features, including edges, and dislocates the edges as it smooths a finer scale to a coarser scale [14] [15]. The nonlinear diffusion, called anisotropic diffusion, proposed by Perona and Malik [16] suffers from regularization. Weickert [17] proposed a semi-implicit scheme to address this issue as given :

$$(I_k)^{t+1} = \sum_{l=1}^{d} (m\, I - \tau\, d^2\, A_l)^{-1}\, I_k^t$$

where k denotes the number of channels and d is the input dimension. I represents the identity matrix, and A_l is the diffusion in the vertical or horizontal direction. In a case where $l = 2$ (2D), the equation is given [18] :

$$(I_k)^{t+1} = (2\, I - \tau 4\, A_x)^{-1} I_k^t + \left(2\, I - \tau 4\, A_y\right)^{-1} I_k^t$$

We extract the information features from the image surface by computing the diffusion speed as given in [7] [19]:

$$I(x,y) = |\log(I^0(x,y) + 1) - \log(I^l(x,y) + 1)|$$

(a) (b) (c) (d)

Figure 2. (a) and (b) Original face of real access – Controlled scenario (c) (d) Diffusion speed map scaled from [0, 255].

B. *Convolution Neural Network*

Machine learning has been successfully used in several different applications such as object detection [20],

handwriting recognition [21] face detection [22], and face recognition [23]. The convolution neural network (CNN) was introduced by LeCun et al. [20] [21]. Our specialized convolution neural network is designed to obtain the local and complex features. Our proposed deep convolution consists of six layers. The first five layers are convolutional and subsampling layers, and the last layer is the output layer, as shown in Figure 3.

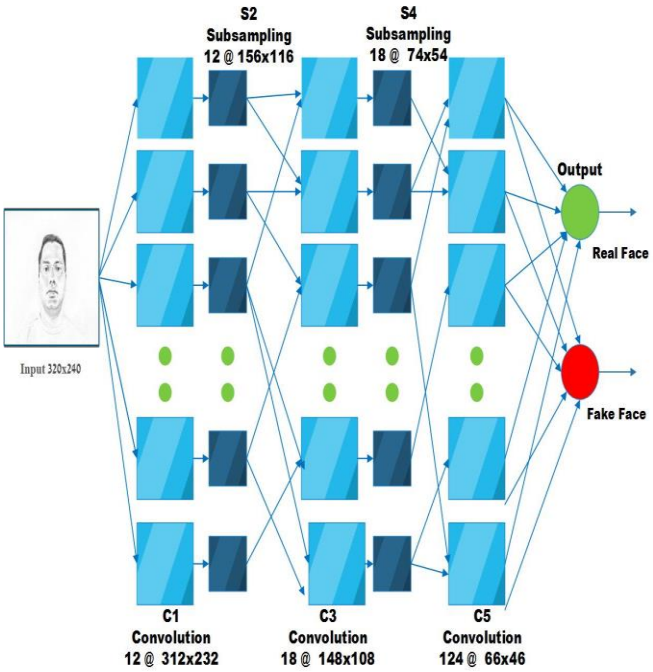

Figure 3. Our proposed convolution neural network architecture.

The input image, not counted in the CNN layer, has a size of 320 × 240 pixels. Layer C1 is the first convolution layer and consists of twelve feature maps. Each unit in the feature map is a result of connecting a 9 × 9 neighbor in the input image. The new size of the feature map is 312 × 232 pixels. Layer S2 is a subsampling layer with twelve feature maps of 156 × 116 pixels. Each feature map in the subsampling layer is connected to an average kernel 2 × 2 neighborhood from the previous corresponding feature map in C1. The average 2 × 2 kernel is non-overlapping. Therefore, the size of the feature map in S2 is half the size of the feature map in C1. C3 is a convolution layer composed of eighteen feature maps of 148 × 108 pixels. Each feature map takes inputs from two random feature maps from the previous S2 subsampling layer. All two feature maps from subsampling are connected to only one 9 × 9 kernel. Layer S4 is a subsampling layer with eighteen feature maps of 74 × 54 pixels. Each feature map in the subsampling is connected to an average 2 × 2 kernel neighborhood from the previous corresponding feature map in C3. The 2 × 2 kernel is non-overlapping over the input and reduces the subsampling to half the size of the input. C5 consists of 124 feature maps of 66 × 46 pixels. Each unit in the feature map is a result of connecting the 9 × 9 neighbor in the input. Each feature map takes two random feature map from the previous S4 subsampling layer. Finally, the last layer is the output layer, a fully connected layer. The output of each feature map is normalized or squashed between 1 and -1 using a hyperbolic tangent activation function called Tanh, which helps with the backpropagation learning. All weights (w) and biases (b) are randomly initialized between -1 and 1. We used a small learning rate with a value of 0.005 to help the network learn quickly, especially to update the weight in the backpropagation. The last layer is the fully connected layer; we used the softmax activation function as a classifier

IV. PERFORMANCE EVALUATION

A. *Dataset*

Replay-Attack Database [6] was released in 2012. It consists of 1200 short videos of 50 different subjects divided into 200 real-access videos and 1000 spoofing attack videos. Each subject recorded a short video with a resolution of 320 x 240 pixels under two different conditions: (1) the controlled condition contained a uniform background and (2) the adverse condition contained a non-uniform background and day-light illumination. The spoof attacks were generated by using one of the following scenarios: (1) print using hard copy, (2) mobile using iPhone screen, and (3) high definition using iPad screen. Each spoof attack video was captured in two different attack modes: hand-based attacks and fixed support attacks [32]. The Replay-Attack database is divided into three subsets: training, development, and testing.

TABLE I. REPLAY-ATTACK DATABASE

Type	Training Fixed |hand	Development Fixed | hand	Test Fixed | hand	
Genuine face	60	60	80	200
Print-attack	30 + 30	30 + 30	40 + 40	100 + 100
Phone-attack	60 + 60	60 + 60	80 + 80	200 + 200
Tablet-attack	60 + 60	60 + 60	80 + 80	200 + 200
Total	360	360	480	1200

B. *Discussion and Analysis*

In this subsection, we discuss and analyze our approach for detecting replay video attacks utilizing only one frame of a single video as shown in Figure 4. Utilizing one frame instead of utilizing 375 frames of a single video reduces the time required for processing which in the users' experience is convenient. When we apply the AOS-based diffusion scheme to obtain sharp edges and surface textures, we found out that real access frame and high definition (Tablet) frame have similar edges and texture which makes it hard for our specialized convolution neural network to distinguish between them. Re-capturing the replay video twice destroys the sharp edges and changes the pixel locations. After conducting several experiments with different time step values, we determined that a time step of (τ = 100) yields the

best result when iterating five times (L =5). When using a larger time step (one greater than τ = 100) the most important features fade out such as the edges and location. Moreover, we also tested the impact of the number of iterations on the classification result. We conducted four experiments using four different iterations (5, 10, 15, and 20) while holding the time step constant at τ = 100. However, increasing the number of iterations from 5 to 10 blurs the face and consumes additional time. The iteration L = 5 yields a HTER of 10%, whereas iterations of L = 10 and L = 20 yield accuracy rates of 14.625% and 17.375 %, respectively as shown in Figure 5. Our proposed specialized CNN has proven to be powerful in extracting not only the sharp edges but also the texture information of a single frame. The trained kernels are able to detect features that help to distinguish the speed-diffused frame. After visualizing the first convolution layer, there is a clear difference in the real and fake diffused frames (e.g., the eye, nose, lips, and cheek regions). The real face has more edges and distinct corners around the eyes and lips, where the fake face has fewer edges and flat surfaces.

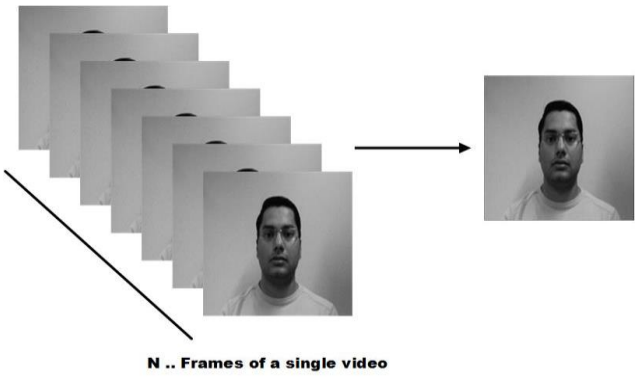

Figure 4. One frame of a short video of real access.

TABLE II. HTER (%) OF CLASSIFICATION FOR THE REPLAY ATTACK DATASET

	test
LBP $_{3*3}{}^{u2} + x^2$ [6]	34.01%
LBP $_{3*3}{}^{u2}$ + LDA [6]	17.17%
LBP $_{3*3}{}^{u2}$ + SVM [6]	15.16%
LBP + SVM [5]	13.87%
DS-Local Speed Pattern [7]	12.50%
Our proposed approach	10%

C. *Performance Evaluation Using the Replay Attack*

We computed the half total error rate (HTER) to assess the statistical significance of the performance of our proposed approach [24]. The HTER is half of the sum of the false rejection rate (FRR) and false acceptance rate (FAR), as shown below:

$$\mathbf{HTER} = \frac{\mathbf{FRR} + \mathbf{FAR}}{2}$$

where FRR is the number of false rejections divided by the total number of clients and FAR is the number of false acceptances divided by the total number of imposters.

TABLE III. HTER (%) OF CLASSIFICATION WITH DIFFERENT PARAMETERS USING THE REPLAY ATTACK DATASET

τ	L	Accuracy	τ	L	Accuracy
40	5	17.125	100	5	10.00
60	5	13.125	120	5	15.875
80	5	13.75	140	5	14.625

TABLE IV. PERFORMANCE EVALUATION FOR DIFFERENT NUMBERS OF ITERATIONS. THE TIME STEP IS FIXED AT A VALUE OF 100

τ	L	Accuracy	τ	L	Accuracy
100	1	16.5	100	6	11.875
100	2	13.5	100	7	11.875
100	3	15.75	100	8	12.625
100	4	14.875	100	9	11
100	5	10.00	100	10	14.625

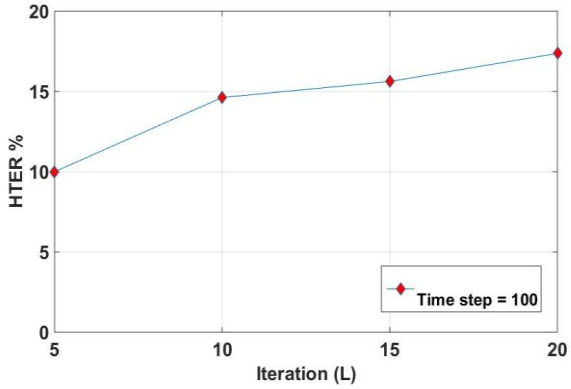

Figure 5. 12 Performance evaluation for different numbers of iterations. The time step is fixed at a value of 100.

V. CONCLUSION

We introduced an effective approach to address the problem of face spoofing attacks using a static frame of sequenced frames. We applied an AOS-based schema with a large time step size to generate the speed-diffused image. Applying large time step parameter helps in extracting the sharp edges and texture features in the input image. Fake face images had fewer sharp edges and flattened surfaces around the eyes, nose, lips and cheek regions when we recaptured the input video twice, which destroys the sharp edges and changes the pixel locations. In contrast, real face frame had sharp edges and rounded surfaces, especially around the nose and lips. Previous approaches used handcrafted features, such as the LBP algorithm, to extract the information features from the diffused image. In contrast, this paper uses a specialized deep convolution network. Our

proposed CNN architecture was able to extract the local and complex features from the diffused image. Our best classification result has a HTER of 10 % when applying a time step of (τ = 100) and setting the number of iterations to (L = 5). Using a large time step destroys the edges. Our future work will explore applying the sparse auto-encoder to obtain a diffused frame. Thus the overall architecture will be an autoencoder generating diffused frame to be fed to our deep CNN network.

REFERENCES

[1] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision–ECCV 2010*, ed: Springer, 2010, pp. 504-517.

[2] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," in *Defense and Security*, 2004, pp. 296-303.

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

[4] Z. Zhiwei, Y. Junjie, L. Sifei, L. Zhen, Y. Dong, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Biometrics (ICB), 2012 5th IAPR International Conference on*, 2012, pp. 26-31.

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

[6] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG - Proceedings of the International Conference of the*, 2012, pp. 1-7.

[7] K. Wonjun, S. Sungjoo, and H. Jae-Joon, "Face Liveness Detection From a Single Image via Diffusion Speed Model," *Image Processing, IEEE Transactions on,* vol. 24, pp. 2456-2465, 2015.

[8] Y. Jianwei, L. Zhen, L. Shengcai, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Biometrics (ICB), 2013 International Conference on*, 2013, pp. 1-6.

[9] P. Gang, S. Lin, W. Zhaohui, and L. Shihong, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1-8.

[10] A. K. Singh, P. Joshi, and G. C. Nandi, "Face recognition with liveness detection using eye and mouth movement," in *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on*, 2014, pp. 592-597.

[11] K. Sooyeon, Y. Sunjin, K. Kwangtaek, B. Yuseok, and L. Sangyoun, "Face liveness detection using variable focusing," in *Biometrics (ICB), 2013 International Conference on*, 2013, pp. 1-6.

[12] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally Efficient Face Spoofing Detection with Motion Magnification," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013, pp. 105-110.

[13] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of Face Spoofing Using Visual Dynamics," *Information Forensics and Security, IEEE Transactions on,* vol. 10, pp. 762-777, 2015.

[14] J. Weickert, B. M. T. H. Romeny, and M. A. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering," *Image Processing, IEEE Transactions on,* vol. 7, pp. 398-410, 1998.

[15] J. Canny, "A Computational Approach to Edge Detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. PAMI-8, pp. 679-698, 1986.

[16] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 12, pp. 629-639, 1990.

[17] J. Weickert, B. T. H. Romeny, and M. Viergever, "Efficient and reliable schemes for nonlinear diffusion filtering," *Image Processing, IEEE Transactions on,* vol. 7, pp. 398-410, 1998.

[18] J. Ralli, "PDE Based Image Diffusion and AOS," 2014.

[19] E. H. Land and J. McCann, "Lightness and retinex theory," *JOSA,* vol. 61, pp. 1-11, 1971.

[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, pp. 2278-2324, 1998.

[21] B. B. Le Cun, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990.

[22] C. Garcia and M. Delakis, "Convolutional face finder: a neural architecture for fast and robust face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 26, pp. 1408-1423, 2004.

[23] S. Lawrence, C. L. Giles, T. Ah Chung, and A. D. Back, "Face recognition: a convolutional neural-network approach," *Neural Networks, IEEE Transactions on,* vol. 8, pp. 98-113, 1997.

[24] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.