Detecting Presentation Attacks from 3D Face Masks under Multispectral Imaging

Jun Liu, Ajay Kumar
Department of Computing, The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
csjunliu@comp.polyu.edu.hk, csajaykr@comp.polyu.edu.hk

Abstract

Automated detection of sensor level spoof attacks using 3D face masks is critical to protect integrity of face recognition systems deployed for security and surveillance. This paper investigates a multispectral imaging approach to more accurately detect such presentation attacks. Real human faces and spoof face images from 3D face masks are simultaneously acquired under visible and near infrared (multispectral) illumination using two separate sensors. Ranges of convolutional neural network based configurations are investigated to improve the detection accuracy from such presentation attacks. Our experimental results indicate that near-infrared based imaging of 3D face masks offers superior performance as compared to those for the respective real/spoof face images acquired under visible illumination. Combination of simultaneously acquired presentation attack images under multispectral illumination can be used to further improve the accuracy of detecting attacks from more realistic 3D face masks.

1. Introduction

Safeguarding the integrity of the biometrics system is critical to avail benefits offered by the biometrics systems for *e-governance*, *e-business* and a range of surveillance applications. Presentation of fake biometrics sample, that can generate closely matching replicas of the real biometrics, requires little efforts and biometrics systems are more vulnerable to such sensor level attacks from fraudulent biometrics samples. Biometrics systems using different modalities have been challenged with a range of attacks resulting from fraudulent fingerprints reconstructed from latent prints of subjects, iris stamps manufactured from the covertly acquired iris or eye images or 3D silicon masks manufactured from multiple 2D images of real subjects. Such challenges have resulted in the development of anti-spoofing techniques to detect such sensor-level attacks before their authentication. This paper focuses on such problem and its scope has been limited to the development of more accurate method to detect disguised or fake face images from 3D face masked subjects.

Automated detection of spoof-biometrics sample images is widely regarded as two-class classification problem. Significant intra-class variations among the real human faces and sophistication in the generation and/or presentation of highly similar fraudulent biometrics sample poses severe challenges in the accurate detection of fake faces.

2. Related Work

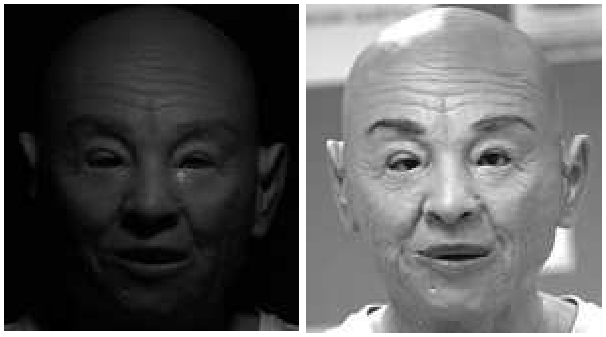

Figure 1. Illustration of a typical realistic 3D face mask, with opened eye and mouth instance from real subject, under multispectral imaging used in our work. Image acquired under near-infrared illumination appears in the *left* while respective image under visible illumination appears on the *right*.

Recent research efforts on the development of more accurate and effective anti-spoofing techniques are fueled by the tremendous growth in demand for safeguarding the integrity of deployed biometrics system. Reference [27] provides a detailed survey on a range of methods developed for detecting presentation attacks on the face recognition based surveillance and recognition systems. Displaying the face image of subject or his photo (2D spoof attacks), replay of video depicting face of the subject (2.5D spoof attacks) or presentation of a 3D face replica representing real subject (3D spoof attacks) are popular methods to thwart integrity of face recognition systems using such sensor level presentation attacks. Among these three categories of attacks, presentation attacks using the realistic 3D face masks is widely considered to be more challenging for the detection and is also the focus of our work in this paper.

Automated detection of 3D face masks to preserve integrity of 2D face recognition systems has attracted increasing attention in the literature. Manjani *et al*. [3] have recently introduced a new silicon face mask based database in public domain and also developed a deep dictionary learning based approach to derive sparse representation of features for the spoof face detection. This is a promising attempt to accurately detect silicon face masks using visible illumination images and will further facilitate much needed research efforts in this area. In addition to the texture based [2], [8]-[10], [29] cues, there have also been attempts to incorporate motion based cues to differentiate real faces from the spoof faces. Usage of eye blinks and mouth movements by Tirunagari *et al*. [26] or usage of Eulerian motion magnification by Bharadwaj *et al*. [13] are examples of some representative works using such approaches. Multiple cures are expected to enhance the accuracy of detecting presentation attacks and such approaches have also attracted attention of researchers. Usage of texture and motion based cues by Siddiqui *et al*. [24] is one such representative example of detecting spoof faces using multiple cues.

Convolutional neural networks were introduced [16] about 25 years ago and have demonstrated tremendous success in a range of problems in biometrics [1], [19]-[20]. We also incorporate such deep learning based approach to investigate detection of spoof 3D faces using multispectral imaging. Therefore these networks are briefly introduced in section 3 while section 4 details on various network configurations investigated in our experiments. The experimental protocols, database acquisition, and results are presented in section 5. Finally, the key conclusions from this work are summarized in section 6.

3. Convolutional Neural Network

Material composition, texture, surface reflectance, illumination and shape of 3D surface from presented faces can generate a variety of features. These heterogeneous features can be self-learned using convolutional neural networks (CNN) [18] and were used in our investigation. CNN is a type of neural network, which mainly uses convolution pattern (multiple layers) to connect the neurons. This network can be viewed as a combination of linear and nonlinear image processing operations. Usually, a CNN network is composed of different kinds of layers, such as convolution layer, pooling layers, ReLU layers, normalization layer, fully connected layers, and loss layers. The network will learn the parameters automatically through forward propagation and backward propagation. A good architecture is a combination of different layers, which can help to extract different unique features when different inputs are provided.

3.1. Convolutional Layer

Convolution layer consists of filter banks, which is activated or learned through multiple forward propagation and backward propagation. The operation in this process can summarized from the following.

$$f(\mathrm{x}) = \sum_{x \subset I} I(x) \cdot W(x) \tag{1}$$

where *f* represents the output, *I* is the input image, *x* is the pixel in the region of filtering region while *W* represents the filter parameter.

3.2. Pooling Layer

Pooling layer is a non-linear operation typically represents the down-sampling operation to reduce the dimension of feature and the complexity of entire network. Among several choices of pooling layers in the literature, such as average pooling (AVE), max pooling (MAX), L2-norm pooling, *etc.* max pooling was used in this work because it cannot only increases the nonlinearity of the entire network, but also reduces the computational complexity. Max pooling operating can be expressed as

$$f(\mathrm{x}) = \max(x_1, \dots, x_i) \tag{2}$$

where x_i, represents the image pixel in the region of operation.

3.3 ReLU, Normalization, and Fully Connected Layer

Since the overall network is a nonlinear operation, ReLU (Rectified Linear Units) layer is used to increase the nonlinear properties of the CNN network. The local response from the normalization layers represents *lateral inhibition* by normalizing local input region responses. Our work used LRN (local response normalization) operation in the normalization layers which can be defined as.

$$f(\mathrm{x}) = \left(1 + \frac{\alpha}{n} \sum_{x \subset I} x^2\right)^{\beta} \tag{3}$$

where *x* denotes input of current layer (or activated value of current neuron) while α and β represents two parameters which can control the scale of normalization. The fully connected layer is the inner product layer where the neurons have full connection to neurons in the previous layer, as in a typical neural network.

3.4 Loss Layer

Finally, the loss layer is incorporated to penalize learned output with true output. In this layer, different loss function is used to penalize this deviation, *e.g.* Softmax loss function is used for predicting multi-classes, Sigmoid loss function to map different input probability values in [0, 1], or the Euclidean loss function to penalize the real value and learned value. We used *softmax* loss function to interpret the outputs as the genuine and impostor class probabilities:

$$f(\mathrm{x}) = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}, j = 1, \dots, \mathrm{K} \qquad (4)$$

where K represents the number of classes and x is the feature vector response from previous layer. The cost function with respect to single sample appears in following:

$$J(\mathrm{W}, \mathrm{b}) = \frac{1}{2} \left\| h_{(W,b)}(x) - y \right\|^2 \qquad (5)$$

where J represents the deviation of CNN output and ground truth, x is the input of CNN network, w is the weight of each regression parameters, and b denotes for bias parameter. Thus, the objective of CNN network is to perform optimization to reduce $J(\mathrm{W}, \mathrm{b})$ during training phase.

4. CNN Configurations

Several CNN architectures were considered in our experiments to ascertain the performance for detecting presentation attacks from more realistic 3D face masks under multispectral imaging. These architectures are briefly summarized in the following.

4.1 Network 1

In this network, there are 3 convolution layers and 1 fully connected layer. There are two output from the fully connected layer. Therefore the classification result (mask or real) can be get directly from the *softmax* loss function.

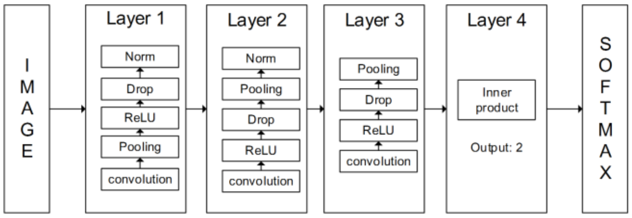

Figure 2: Network configuration 1.

4.2 Network 2

In this setting, the configurations of network is the same as network 1. However, the difference is initialization method is used. The initialization of weights in this network is also Gaussian distribution with zero mean, but the variance of the output in each layer is equal to one. This initialization method is referred to as "*msra*" [25].

In order to let the variance of the output equal to one, the initialization method is according to Gaussian distribution.

$$\mathrm{N}\left(0, \frac{2}{n}\right), with\ n = k^2 \cdot d \qquad (6)$$

where k is the filter size in current layer and d is the number of filters in last layer. Under this situation, the variance of the outputs in each convolution layer or fully connected layer is one.

4.3 Network 3

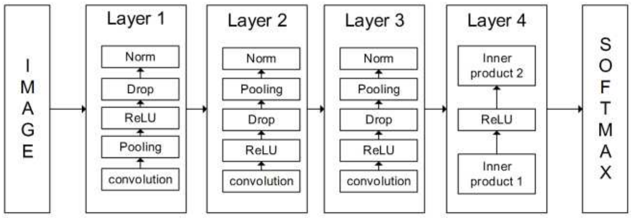

Figure 3: Network configuration 3.

In this network, in order to increase the nonlinear properties of network, two fully connected layers are added in the last part. Besides, one ReLU layer is also inserted between these two fully connected layers to increase the nonlinearity of the whole network.

4.4 Network 4

In order to overcome over-fitting, one drop layer between the last two fully connected layers is also added for this configuration. In addition, in order to test the impact of number of convolution layers on the overall performance. Our experiments evaluated the performance when the number of convolution layer is set as two.

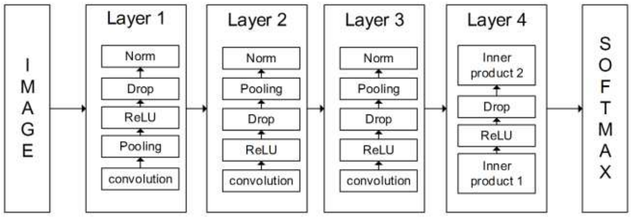

Figure 4: Network configuration 4.

4.5 Network 5

In this network setting, the number of convolution layer is fixed to two. Therefore we can use this setting the evaluate the influence of convolution layer on the overall performance.

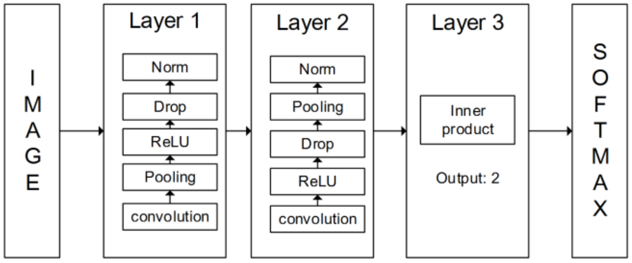

Figure 5: Network configuration 5.

4.6 Siamese Network

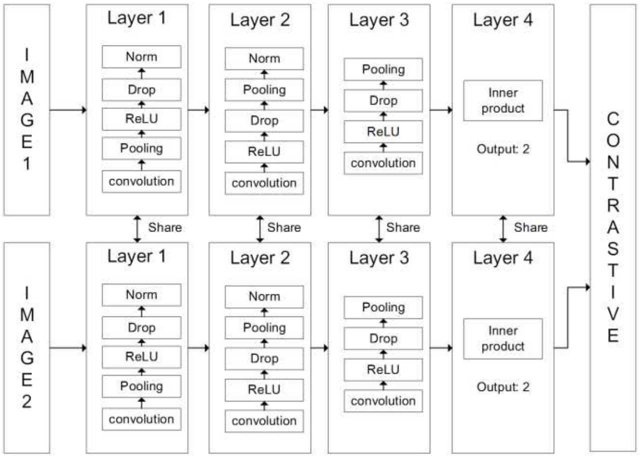

Figure 6: Siamese network used in our experiments.

The general idea of Siamese network is to train two networks at the same time. The purpose is to reduce the distance between these two networks' output if the input pair belongs to the same category, and increase the distance between these two networks' output if pair belongs to different categories.

In this network configuration, two independent networks can be trained at the same time. Image 1 and image 2 are two different inputs for these two networks. In the training process, the parameters can be shared in each separate layer, such as convolution layers in network 1 to convolution layers in network 2, fully connected layer in network 1 to fully connected layer in network 2. Thus, the complexity of the network can be reduced. In our work, parameters in convolution layers are shared with each other network. The loss function used in Siamese network is Contrastive loss function, which can be expressed as.

$$E = \frac{1}{2N}\sum_{n=1}^{N}(y)d^2 + (1-y)\max(margin - d, 0)^2 \quad (7)$$

where d is distance between two different outputs, $d = \|a_n - b_n\|_2$. a_n and b_n are the outputs of two networks. If a_n and b_n belongs to the same category, then $y = 0$, otherwise, y=1. In this situation, E can penalize the distance between these two categories. When real face is exposed to NIR and visible light separately, there is significant difference in respective images. However, when 3D face masks are exposed to NIR and visible light separately, the corresponding image difference is relatively smaller. Therefore two images (NIR and visible) from the real face are assumed to belong to different categories, and the images (NIR and visible) form 3D face mask are assumed to belonging to same categories.

4.7 Combination of Visible and NIR Image

Simultaneously acquired visible and NIR images can both be used to enhance the performance for the detection of presentation attacks and was investigated in this work. Firstly, we use network configuration 4, which can offer (section 5) the best performance from trained network by individually using NIR and visible images. The feature vectors from the last layer, each from the NIR and visible input images, were combined/concatenated to jointly describe the NIR-Visible image pair input for the classification. A two-class SVM classifier was trained, using the data from training phase. This SVM was used for the performance evaluation using the test phase data.

4. Experiments and Results

The focus of this work is on the comparative performance evaluation from the images acquired under NIR and visible illumination, and from the joint usage of such simultaneously acquired real and 3D face mask images. Therefore we acquired such database under indoor environment. In the following sections, a brief description is firstly provided for the employed database. This is followed by experimental results using the network configurations and architectures detailed in section 3.

A. The Database

The database employed in this work is composed of NIR and visible image pairs. Each image pair in this database therefore has two images, which are acquired at the same time, *i.e.*, One image from NIR camera, and other from visible RGB camera. The schematic diagram representing our image acquisition setup appears in the following figure.

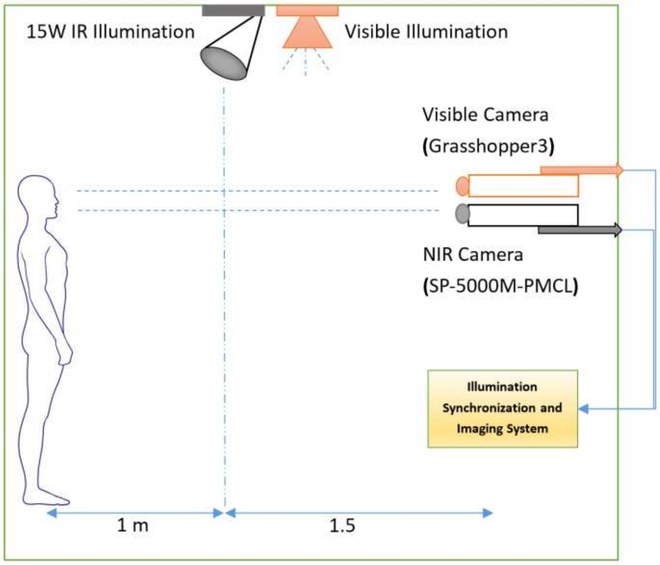

Figure 7: Image acquisition setup for multispectral imaging.

We acquired images from a total 13 different 3D face masked subjects and 9 different real subjects. The images in figure 1 shows the 3D face masked image samples from our database. The first six 3D face masked subjects are used for the training and the rest seven 3D face masked subjects are used in the test phase for the performance evaluation. The size of *acquired* NIR image is 2560×2048 pixels, and the size of acquired visible image is 3376×2704 pixels. The region of interest from each of the acquired face image were automatically detected using a robust publicly accessible face detector from [21]. We employed two different protocols during the performance evaluation and these are summarized in the following Table.

Table 1: Protocols used in our experiments.

	Training Phase		Test Phase	
	Attack Subjects/Images	Real Subjects/Images	Attack Subjects/Images	Real Subjects/Images
Protocol 1	6/7408	4/1622	6/9408	4/1736
Protocol 2	6/7408	4/1622	7/9686	5/1987

B. Experimental Results

In order to evaluate the performance for detecting spoof faces, ISO/IEC 30107-3 [11] recommends the following metrics for the evaluation of *Presentation Attack Detection* (PAD): (a) *Attack Presentation Classification Error Rate* (APCER), which means the rate of attacks classified as real presentations and (2) *Normal Presentation Classification Error Rate* (NPCER), which indicate the rate of real faces classified as attack faces. Besides, *Average Classification Error Rate* (ACER), which can be computed from the average of previous two errors, is also presented in the in our experimental results. We trained all our models using Caffe [28] on a single NVIDIA GTX670 platform. We firstly present experimental results using the NIR and Visible image set separately. This is followed by the experimental results using the combination of jointly acquired NIR and visible illumination images.

Table2: Experimental results using NIR database.

Image Type	Protocol	Network Configuration	APCER (%)	NPCER (%)	ACER (%)
NIR	Protocol 1	Network 2	0	100	50
NIR	Protocol 1	Network 3	5.0	0	2.5
NIR	Protocol 1	Network 4	0.16	0	**0.08**
NIR	Protocol 2	Network 2	0	100	50
NIR	Protocol 2	Network 3	4.98	12.38	8.68
NIR	Protocol 2	Network 4	0.2	12.63	**6.41**

Table 3: Experimental results using visible database.

Image Type	Protocol	Network Configuration	APCER (%)	NPCER (%)	ACER (%)
Visible	Protocol 1	Network 2	0.06	2.9	1.5
Visible	Protocol 1	Network 3	0.04	0.58	**0.31**
Visible	Protocol 1	Network 4	0.07	0.9	0.5
Visible	Protocol 2	Network 2	0.06	15.2	7.6
Visible	Protocol 2	Network 3	0.04	13.1	**6.6**
Visible	Protocol 2	Network 4	0.07	13.4	6.8

Table 4: Experimental results from jointly using NIR and visible images.

Image Type	Protocol	Network Configuration	APCER (%)	NPCER (%)	ACER (%)
NIR & Visible	Protocol 2	Network 4	0	12.63	6.32

The experimental results or our attempts using the Siamese network that can combine NIR and visible images are illustrated in table 5. In the training set, two images (NIR and visible) of real face are labelled in different categories. On the contrary, the images (NIR and visible) from 3D face masked subjects are labelled to the same categories.

Table 5: Experimental results from Siamese architecture using NIR-Visible image pairs.

Train Attack Subjects/Images	Train Real Subjects/Images	Test Attack Subjects/Images	Test Real Subjects/Images	APCER (%)	NPCER (%)	ACER (%)
6/7408	4/1622	7/9686	5/1987	99	0	49.5

C. Comparative Results

This part summarizes the comparative results for the 3D face masks detection using two different protocols used in our experiments.

Table 6: Experimental results from "NIR-Visible database" using protocol 1.

Method	Image type	Protocol	Best Result ACER (%)
CNN	NIR	protocol 1	**0.08**
CNN	Visible	protocol 1	0.31
CNN	NIR & Visible	protocol 1	**0.05**

Table 7: Experimental results from "NIR-Visible database" using protocol 2.

Method	Image type	Protocol	Best Result ACER (%)
CNN	NIR	protocol 2	**6.41**
CNN	Visible	protocol 2	6.8
CNN	NIR & Visible	protocol 2	**6.32**

5. Conclusions and Future Work

This paper has investigated comparison and combination of performance for automatically detecting 3D face masked subjects from simultaneously acquired near-infrared and visible illumination images. This work required us to develop such imaging setup and acquire database for the experiments. The experimental results presented in section 4 indicate that near-infrared imaging can offer superior performance than those based on visible imaging. Our experimental results also suggest that the combination of simultaneously acquired near-infrared and visible images (multispectral imaging) can be used to further improve the performance for the detection of 3D face masked subjects.

Our experiments also indicate that the performance from Siamese network configuration is not encouraging, *i.e.* poor, as compared with those from the other CNN configurations. This can be possibly due to lack of adequate training data, and the nature of training strategy employed in our experiments. Extensive data augmentation and usage of superior learning strategy, *i.e.* usage of instance normalization instead of batch normalization, is expected to further improve performance and is part of further work. It can also be observed from the results in table 5 and table 6 that the performance improvement from the (feature-level) combination of near infrared and visible images is not very significant and therefore our results should be considered preliminary but encouraging. Usage of better normalization of feature vectors, data augmentation, or better fusion strategy is expected to further improve the performance and is part of our ongoing work.

References

[1] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Trans. Info. Forensics and Security*, vol. 10, no. 4, pp. 864-879, 2015.

[2] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," *Proc. 6th Intl. Conf. Image Processing Theory Tools and Applications, IPTA 2016, pp.* 1–6. 2016.

[3] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, "Detecting Silicone Mask based Presentation Attack via Deep Dictionary Learning," *IEEE Trans. Info. Forensics and Security*, pp. 1713 - 1723 vol. 12, July 2017,

[4] J. Yang, Z. Lei, and S. Z. Li. Learn convolutional neural network for face anti-spoofing. *CoRR*, abs/1408.5601, 2014.

[5] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," *Proc. IJCB 2011*, Washington DC, pp. 1-7, 2011.

[6] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.W. Cheung, " Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Visual Comm. & Image Repr.*, pp. 451–460, vol. 38, 2016.

[7] Z. Xu, S. Li, and W. Deng. "Learning temporal features using lstm-cnn architecture for face anti-spoofing," *Proc. 3rd ACPR*, pp. 141–145, 2015.

[8] Z. Boulkenafet, J. Komulainen, A. Hadid, "Face Spoofing Detection Using Colour Texture Analysis," *IEEE Trans. Info. Forensics and Security*, pp. 1818-1830, Aug. 2016.

[9] D. Wen, H. Han, A. K. Jain, "Face Spoof Detection With Image Distortion Analysis," *IEEE Trans. Info. Forensics and Security*, pp. 746-761, 2015.

[10] Atoum Y, Liu Y, Jourabloo A, et al. Face Anti-Spoofing Using Patch and Depth-Based CNNs, *Proc. Intl. Joint Conf. Biometrics*. IJCB 2017. Denver, 2017.

[11] ISO. "ISO/IEC CD 30107-3, *Information technology -- Biometric presentation attack detection -- Part 3: Testing and Reporting*

[12] W. Bao, H. Li, N. Li, W. Jiang, "A liveness detection method for face recognition based on optical flow field," *Proc. Intel. Conf. Image Analysis and Signal Processing,* Taizhou, China, pp. 233–236, 2009.

[13] S. Bharadwaj, T.I. Dhamecha, M. Vatsa, R. Singh, "Computationally efficient face spoofing detection with motion magnification," *Proc. CVPR 2013 Biometrics Workshop*, pp. 105–110. Portland, USA, 2013,

[14] T.F. Pereira, J. Komulainen, A. Anjos, J.M.D. Martino, A. Hadid, M. Pietikainen, S. Marcel, "Face liveness detection using dynamic texture," *EURASIP J. Image Video Process*. 1 2014.

[15] J. Komulainen, A. Hadid, M. Pietikainen, A. Anjos, S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," *Proc. ICB 2013*, Madrid, 2013.

[16] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation pplied to handwritten zip code recognition," *Neural Computation*, pp. 541-551, 1(4), 1989.

[17] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *Intl. J. Computer Vision,* pp. 211-252, 2015.

[18] A. Krizhevsky,S. Ilya and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*. NIPS 2012, pp. 1097-1105, 2012

[19] R. Ramachandra, K. R. Bylappa, S. Venkatesh, C. Christoph, "Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images," *Proc. CVPR Biometrics Workshop,* July 2017.

[20] R. Ranjan, V. M. Patel, R. Challepa, "HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 1, Dec. 2017.

[21] Face++: Leading Face Recognition on Cloud. www.faceplusplus.com/ 2017

[22] N. Erdogmus and S. Marcel, "Spoofing face recognition with 3D masks," *IEEE Trans. Info. Forensics and Security*,, vol. 9, no. 7, pp. 1084–1097, 2014.

[23] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process*., vol. 24, no. 12, pp. 4726–4740, 2015.

[24] T. A. Siddiqui, S. Bharadwaj, T. Dhamecha, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha, "Face anti-spoofing with multifeature videolet aggregation," *Proc. Intl. Conf. Pattern Recognition*, ICPR 2016, Cancun, Mexico, Dec. 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, " Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proc. ICCV 2015*, pp. 1026-1034, 2015.

[26] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki, and A. T. S. Ho, "Detection of face spoofing using visual dynamics," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 762–777, Apr. 2015..

[27] R. Ramachandra, C. Busch, "Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey," *ACM Comp. Surveys.* vol. 50, Apr. 2017.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *Proc. 22nd ACM Int. Conf. Multimedia*, pp. 675-678, Nov. 2014.

[29] A. Agrawal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, A. Noore, "Face presentation attack with latex masks in multispectral videos," *Proc. CVPR Biometrics Workshop 2017*, Hawaii, July 2017.