Face Liveness Detection Based on Parallel CNN

Xin Li\textsuperscript{1, *}, Wei Wu\textsuperscript{1}, Tao Li\textsuperscript{2}, Yang Su\textsuperscript{1} and Lilin Yang\textsuperscript{1}

\textsuperscript{1}School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China
\textsuperscript{2}Biota, Taiyuan Normal University, Jinzhong 030619, China

*Corresponding author e-mail: 15271904775@whut.edu.cn

Abstract. In this paper, we develop an effective framework based on deep learning for face liveness detection. Liveness detection is a great challenge in computer vision. Over the past decade, the interest of people in safety management has increased, and face recognition technology has gradually expanded into the commercial areas. However, general face recognition systems also present risks in terms of security and are vulnerable to spoofing attacks. Therefore, we propose a practical deep learning framework to improve the accuracy of classification. First, we observe the image and label the areas with outline and facial features. Second, we introduce an attention model, which fuses local features and global features to obtain multi-channel features. Third, we use pyramid structures and multiple convolutional neural networks with different depths to classify them in parallel and combine them with multiple results. We evaluated the performance of this method on the CASIA-SURF data set. Experiments show that the proposed frame-trained human face classifier has better performance than the existing classifier.

1. Introduction

Among the information obtained by human apperceive systems, visual information accounts for about 80\% to 85\%. Visual information, images and videos, is becoming more and more important in people's daily life. In the past decade, there has been an increasing interest in security management, using numerous biometric technologies. Biometric technology uses physical or behavioral characteristics for identification. Since 2007, various schemes have been explored, including fingerprints [1], faces [2], and signatures [3]. Today, face recognition technology has gradually spread in many fields, including places with high security, such as border monitoring, and places with low security, such as face payment. However, face recognition systems also have security risks and are vulnerable to spoofing attacks [4]. Therefore, this paper studies the face liveness detection algorithms based on infrared images and depth images. This technology mainly determines whether the input currently received by the system is a living face or a spoofing face.

After studying the existing framework, the traditional method is to find the differences between liveness and non-liveness attacks, and design features based on these differences, then send them to the classifier for making decisions. In other words, that is feature extraction based on texture, frequency, and color, then use support vector machine, linear regression, cosine distance, etc. for distinguishing between living face and spoofing face. Texture may be the most effective evidence for judging spoofing face. Under the assumption that there are certain texture patterns in the spoofing
images that are different from the real images, researchers have proposed a series of algorithms, extracting texture features from facial images to make judgments. In the existing face liveness detections, more than 69% of the research works only use texture descriptors or combine with other descriptors to identify spoofing images. In 1996, [5] proposed the Local Binary Patterns (LBP), which labels every pixel by comparing it with its neighbors, concatenating the results into binary numbers, and describes the texture in a histogram. Until 2015, some people successively studied the use of deep learning for face liveness detection, but because the number of public datasets is too small, the performance has not surpassed the traditional method, and [6] proposed a method to solve multi-frame time series detection by using CNN-LSTM to simulate the traditional method LBP-TOP, but the performance is worrying. In 2018, [7] published an article in CVPR2018, and the performance finally surpassed the traditional method, designing a deep learning framework to predict the pulse statistics and depth map end-to-end.

The above face spoofing detection algorithms extract the global features of the face without focusing on the local features of the face. Therefore, we propose a face liveness detection framework based on deep learning, focusing on the contours and features in depth images and infrared images.

To further illustrate the main idea of our method, a representative example has been shown in Fig.1. This image describes the features in the depth map of the living face and the spoofing face. Obviously, the existing methods almost ignore the difference of contours between the living face and the spoofing face. In contrast, we attach great importance to the information in the saliency area and use parallel network [8] and pyramid structure [9] to fuse local features and global features. And our result obviously acquires better vision performance.

Figure 1. The difference of the features in the depth map. The depth in the left is the fake face, while the right is the live face. In the live face, image have obvious feature about features, while the fake face do not.

First, it labels the attention areas, such as features and contours. Second, train a face liveness detection network associated with a self-attention network [10] on depth image dataset, which combines local features and global features. Third, we use a pyramid structure and multiple convolutional neural networks [11] with different depths to classify them in parallel.

2. Our Framework
This section describes our method, and we first introduce the motivation of this paper. Then, we present three key parts i.e. the attention network for salience detection and the classification of faces on depth images, and using pyramid structure and parallel network in our method to improve the accuracy.

2.1. Motivation
According to [12], the depth map of the facial features of the living face and spoofing face are different. The living face has obvious distribution about depth map, while the corresponding spoofing face does not have obvious distribution about depth map. Therefore, face liveness detection algorithm which performs the same processing at the edge and the inside of the input image will cause a lot of resources to be wasted. In order to focus on detecting the edge and facial features of the face, when
extracting features, we adopt a two-branch structure, which extracts features from the edge and facial features separately. In the branch of detecting the facial features of the face, an attention model is added.

The framework of our face liveness detection algorithm is shown in Fig. 2(a). First, we designed attention network which is trained on the visible image dataset to focus on the facial features. Then, the attention network added to the pyramid structure and parallel networks are used to classify depth images, then images with medium score are evaluated by the same algorithm for their infrared images. We train the network, and final network can obtain depth images and infrared images with the living face or the spoofing face as input, and then obtain a higher accuracy rate.

2.2. Attention Network de-spoofing on depth maps
In order to deal with the problem of accurately distinguishing the living face and the spoofing face, we choose to use multiple parallel networks, and add a pyramid structure and an attention model to the parallel networks. Therefore, we first train the designed attention network on the depth map to extract features. This part mainly introduces the proposed attention network architecture, as shown in Eq.1.

\[
F_{\text{attention}} = \sum_i \omega_{i,k} \cdot x_{i,k} + \sum_i \omega_{i,k,j} \cdot x_{i,k,j} + \sum_i \omega_{i,k,j} \cdot x_{i,k,j}
\]  

Figure 2. The structure of our framework (a) and attention block (b). (a) The left image is our framework, the same image as the input of two parallel backbone network, the purple area is attention block, the yellow area is spatial pyramid pooling. The output of framework is our result of prediction of network. (b) The right image is attention block which is used for reassigned feature weights, and make network focus on salient regions in the feature map.

As illustrated in Fig. 2(b), the input of the network is the depth map of the living face and the spoofing face, and the output is the result of classification. Attention domain includes spatial domain, channel domain and mixed domain. Spatial domain ignore the information in the channel domain and treat the image features in each channel equally, limiting the method of spatial domain transformation to the feature extraction stage of the original image; while channel domain is directly perform global average pooling on the information in a channel, and ignore the local information in each channel. Therefore, we combine position space and channel attention mechanism, using spatial domain and channel domain in parallel. First, a feature attention module is introduced into the convolutional neural network, which pays different degrees of attention to each area of the feature map of the original convolutional neural network, reducing the background of the feature map and the interference of negative sample information. This is mainly Feature extraction of facial features. Second, a significant channel self-attention module is introduced for the features after facial features and contour fusion, which distinguish different channels in the feature map and screen out useful channel information to make the features to be more representative.

2.3. Spatial Pyramid Pooling
In a general convolutional neural network, a full connection is usually connected behind the convolutional layer. The number of features in the fully connected layer is fixed, so when the image is
input, the size of the input is fixed-size. But in fact, the size of our input image can't always meet the required size. Therefore, we use spatial pyramid pooling instead of pooling layers in convolutional neural networks. This part mainly introduces spatial pyramid pooling.

As illustrated in Fig. 3(a) [9], the input of the network is the depth map of the living face and the spoofing face, and the output is the result of classification. The entire face to be detected in the convolutional neural network is used to extract features at one time to obtain a feature map, and then find the regions of each candidate frame in the feature map. The spatial pyramid is used as the feature pooling layer of the network, and the salient features of the image are obtained by means of hierarchical pooling features. And then, the bottom-level features and high-level features are cascaded to obtain a vector representation of the multi-scale features of the sample. Finally, a fixed-length feature vector is extracted, and feature vector is input to a classifier for classification.

![Spatial Pyramid Pooling](image)

**Figure 3.** The structure of our SPP and parallel network. (a) The spatial pyramid pooling layer is connected to the last convolutional layer, combining with different scale pooling layer, which is directly using the information of the last convolutional layer. (b) The parallel convolutional neural network is an ensemble framework, which can ensemble many framework of network. We use two frameworks to ensemble.

### 2.4. Parallel Convolutional Neural Network

As showed in Fig. 3(b), face liveness detection algorithms based on convolutional neural networks are mainly classified by texture [13] and spatial relationships of depth maps. Because the similarity between the living face and the spoofing face is relatively high, this similarity is basically reflected in the shape features. Therefore, we extract the features of the face, and then input the features to parallel convolutional neural networks with different scales. In this paper, we mainly extract the spatial features of the depth map, abandon the shape features with high similarity, remove the irrelevant information on the image, reduce the calculation amount of feature extraction, and input to the classifier for classification.

### 3. Experiments

In this section, we compare the results of our method with the state-of-the-art about face liveness detection such as MobilenetV2 [14] and FeatherNetA [14].

#### 3.1. Evaluation Metrics

In order to evaluate performance, the following generally used metrics will be introduced [14]: Average Classification Error Rate (ACER). ACER is treated as the evaluation metric. We also use traditional evaluation metrics, such as EER and AUC to evaluate performance. EER is Equal Error Rate, while AUC is Area under Curve. Otherwise, the other evaluation metrics [15] are used such as $\text{TPR} @ \text{FPR}=10\text{E}-2$.

#### 3.2. Datasets

CASIA-SURF [15], a multi-form (RGB, Depth, IR) dataset containing a large number of people of different ages, contains 1000 different targets and 21,000 videos, using depth and IR in our experiment.
3.3. Result Analysis

Table 1. Three Scheme comparing.

| Name          | ACER | EER  | AUC   | TPR@FPR=10E-2 |
|---------------|------|------|-------|---------------|
| MobileLiteNet | 0.032| 0.028| 0.9962| 0.941         |
| FeatherNetA   | 0.026| 0.025| 0.9969| 0.952         |
| Our method    | 0.024| 0.024| 0.9970| 0.958         |
| GT            | 0    | 0    | 1     | 1             |

Our method is compared with MobileNetV2 and FeatherNetA, which acquire better performance in most evaluation metrics, shown in Table 1. The models are trained with CASIA-SURF training set. As shown in Table 1, compared with existed methods, our method has lower ACER, EER and higher TPR, AUC, and all experiments are the result of six iterations. In order to complete our experiments, we use GTX 970, i5-4440 CPU, and memory is 19.5 GB, while we use pytorch, deep-learning framework.

4. Conclusion

In this paper, we observed the difference between the living face and the spoofing face in the depth map. Therefore, we proposed a practical framework for high-efficiency classification of face liveness detection. The proposed attention network can assign focus for facial features. Spatial pyramid pooling and parallel network can increase the performance of the proposed method under limited training data. The experiments demonstrate that our method for face liveness detection performs better than existed methods. The proposed method achieves higher accuracy than others. In the future, we can train our method using RGB and update our method based on de-spoofing algorithm [16].

Acknowledgments

This work was supported by my teacher, Wei Wu, and my classmates.

References

[1] H. Hasan, S. Abdul-Kareem, Fingerprint image enhancement and recognition algorithms: a survey, Neural Comput. Appl. 23 (2013) 1605–1610.
[2] L. Dora, S. Agrawal, R. Panda, A. Abraham, An evolutionary single Gabor kernel based filter approach to face recognition, Eng. Appl. Artif. Intell. 62 (2017) 286–301.
[3] A. Sammorino, S. Yazid, A survey for handwritten signature verification, in: 2012: pp. 54–57.
[4] P. Meadowcroft, Card fraud – will PCI-DSS have the desired impact?, Card Technol. Today. 20 (2008) 10–11.
[5] T. Ojala, M. Pietikäinen, D. Harwood, A comparative study of texture measures with classification based on featured distributions, Pattern Recognit. 29 (1996) 51–59.
[6] Z. Xu, L. Shan, W. Deng, Learning temporal features using LSTM-CNN architecture for face anti-spoofing, in: 2015 3rd IAPR Asian Conf. Pattern Recognit. ACPR, 2015.
[7] Y. Liu, A. Jourabloo, X. Liu, Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision, (2018).
[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, (2014).
[9] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 1904–16.
[10] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-Excitation Networks, (2017).
[11] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Adv. Neural Inf. Process. Syst. 25 (2012).
[12] Y. Atoum, Y. Liu, A. Jourabloo, X. Liu, Face Anti-Spoofing Using Patch and Depth-Based CNNs, in: Int. Jt. Conf. Biom. IJCB 2017, 2017.
[13] T. de Freitas Pereira, A. Anjos, J.M. De Martino, S. Marcel, LBP – TOP Based Countermeasure
against Face Spoofing Attacks, in: J.-I. Park, J. Kim (Eds.), Comput. Vis. - ACCV 2012 Workshop, Springer Berlin Heidelberg, 2013: pp. 121–132.

[14] P. Zhang, F. Zou, Z. Wu, N. Dai, S. Mark, M. Fu, J. Zhao, K. Li, FeatherNets: Convolutional Neural Networks as Light as Feather for Face Anti-spoofing, ArXiv190409290 Cs. (2019).

[15] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, S.Z. Li, A Dataset and Benchmark for Large-scale Multi-modal Face Anti-spoofing, (2018).

[16] A. Jourabloo, Y. Liu, X. Liu, Face De-Spoofing: Anti-Spoofing via Noise Modeling, 2018.