CNN Face Live Detection Algorithm Based on Binocular Camera

Chunyan Li*, Rui Li and Jianhong Sun
Engineering College, Honghe University, Mengzi 661100, Yunnan, China

*Corresponding author e-mail: lichunyan@uoh.edu.cn

Abstract. In this paper, a convolutional neural network (CNN) detection analysis is performed for live face detection by binocular cameras, and a binocular stereo matching network with fused edge detection is designed and implemented to target the quality of image details and parallax prediction at edges. Experiments show that the random sample pair confusion loss function can effectively improve the accuracy and generalization of the face live detection algorithm; the multi-task training approach can improve the performance of both faces live detection and face recognition; the algorithm shows excellent performance in both faces live detection and face recognition, especially the generalization of face live detection is greatly improved. A pre-trained convolutional neural network is used to extract features, and a content loss function and a domain loss function are designed to measure the feature distance between two images, and a feedforward neural network is trained as an image transformation network to migrate samples to the same domain. Experiments show that the algorithm can reduce the feature differences between the face live detection data of the two domains and can be used to improve the generalization of the face live detection algorithm.

Keywords: Binocular Camera, CNN, Live Face, Detection Algorithm

1. Introduction
With the development of society and the progress of human civilization, social public security and personal information security are increasingly valued and concerned by people. With the continuous development of computer technology and sensor technology, biometric technologies, such as fingerprint recognition, iris recognition, face recognition, etc., use the uniqueness of human characteristics to enter a variety of application scenarios in human life, providing people with a more convenient experience and more secure security [1]. Face recognition technology is a biometric technology based on the information of human facial features to achieve the purpose of identification. Depth estimation based on binocular stereo vision can also be applied in face recognition-based security authentication systems. In recent years, with the continuous development of face detection and recognition technology, face recognition-based authentication systems have been widely used, however, there are still major security risks in the system. To enhance system security, most of the deployed live face detection algorithms are based on a "question-response" model, where the user...
cooperates by performing corresponding actions to provide motion cues to the algorithm. However, the required motion cues for such methods are not very difficult to infer, and the robustness and security of the algorithms need to be improved [2]. Therefore, the depth estimation of binocular stereo vision can play an important role in face anti-counterfeiting applications [3].

Yang et al. designed multi-statistic features, which include reflection features, quality features, and color features, and eventually cascaded these features to form the difference classification features for positive and negative samples, but the test results were average for high-definition RGB print samples or when the image quality loss was not severe [4]. Khan et al. Samarth Bharadwaj [3] et al. proposed the use of multi-frame image faces for live detection, whose main principle is to capture the differences between micro-motions [5]. Kowsalya et al. use multi-frame images then discriminate based on spatial subgraphs of motion energy, so motion-enhanced live detection methods are also widely used [6]. To accomplish the function of authentication by face information, it usually involves designing a face recognition system. From the purpose of the application, it is necessary to achieve various requirements such as the complete design of the whole system process, accuracy of the relevant algorithms to meet the standards, real-time computation time as much as possible, and guaranteed security [7].

Unlike traditional research methods that treat live face detection and face recognition as separate tasks, this paper explores the correlation between the two tasks of live face detection and face recognition, and uses the idea of multi-task learning to design a multi-task convolutional neural network that simultaneously trains live face detection and face recognition, which is capable of handling both tasks of live face detection and face recognition. Then the domain adaptive algorithm for face live detection data is introduced in detail, including the overall design, image transformation network, loss network, and training process, and finally, the impact of the algorithm on the features and face live detection algorithm is analyzed through experiments.

2. Design of CNN Detection Algorithm for the Live Face of the Binocular Camera

2.1. Binocular Camera CNN Detection Design

The reconstruction of the 3D structure of the face to be detected by the binocular stereo vision and the stereo structure feature analysis are important clues to perform face live detection. Due to the instability of binocular depth features, i.e., there may be depth estimation noise affecting the feature characterization, the live judgment can be assisted with the help of macrotexture features [8]. Therefore, this chapter investigates the problem of face live detection based on the combination of binocular vision depth features and epistatic features and innovatively proposes two face anti-counterfeiting algorithms. The system extracts binocular depth features for template face alignment from the input binocular image containing the face to be detected: each detected face key point is first dimensionally augmented, and the third dimension of the augmentation is the relative depth of the key point; after that, each face key point is transformed by the proposed template face alignment algorithm to match the corresponding face key point in the template face; after several rounds of "conversion-matching" iterations, the binocular depth feature of template face alignment is obtained, as shown in Figure 1.

For real faces and fake faces, the intra-class differences in the corresponding stereo structures of the two types of faces are largely due to the variety of face poses. Therefore, to truly characterize the difference between the stereo structures of the two types of faces, a standard real face, called "template face", is designed in this paper. Its use is that if the face to be detected is similar to the template face through some transformation, the face is more likely to be judged as a real face. Therefore, the template face can be considered as the standard 3D structure for the face live detection task. Define a template face T, represented by NP standard 3D abstract face key points.
After obtaining the template face and the 3D abstract face key points of the face to be detected, a conversion algorithm needs to be defined to act on the initial abstract face key points and match them with the template face until they are the most similar to the template face. Therefore, the binocular depth feature of the template face alignment will use this important cue to portray the difference in stereoscopic structure between the real and fake faces. The ideal alignment transformation seeks the optimal parameters to obtain the minimum alignment error.

\[
(a, b; X, \Phi) = \int_{-\infty}^{\infty} a^2(\tau) g^2(\tau - t)e^{-j\omega t} \, d\tau
\]

Search for optimal alignment parameters using an iterative optimization algorithm: The optimal transformation parameters are obtained through multiple rounds of iterative corrections, rather than just a single round of computation.

\[
k_m(x_i) = \sum_{i=1}^{N} w_i s_i(x_i) = W^T S(x_i)
\]

**Figure 1.** The flow of binocular depth feature determination algorithm

Due to the limited resolution and Gaussian blur, the recaptured images lose some high-frequency information and features [29, 68, 97]. The recaptured image appearance also changes due to anomalous projections in some regions. Also, printed face images may have defects and videos containing faked faces may have some noise signals. Therefore, macrotexture features are also important clues to distinguish real faces from faked face representations.

\[
D_\varepsilon(x_i(t)) = [s \left( \|x_i(t) - \ell^{-2} r_j \| \right)]
\]
The algorithm requires binocular image pairs containing the face to be detected as input. The binocular depth features of template face alignment are used to characterize the 3D structure of the face to be determined and can distinguish the difference in the stereo structure between real faces and fake faces [9]. The spatial pyramid encoded micro-texture features are used to characterize the local apparent features of the face to be judged, and the spatial pyramid encoding algorithm proposed in this paper enables the SPAT feature subsets to effectively represent the multi-scale apparent information of the face region. Finally, the two feature descriptors are combined and the confidence scores of the corresponding classifiers are fused to obtain the final live face judgment.

When a batch (Batch) of data is input into a neural network, a layer of the network in between is normalized. Specifically, the mean and variance of the features of this batch of data need to be calculated first, and then the variance is subtracted from the features of this layer and then divided by the mean to make it obey a standard normal distribution.

\[
\phi_{m,n} = \frac{1}{\delta^2} \exp(-\frac{(q_{m,n}-z_{m,n})^2}{2\gamma^2}) \ast [e^{i(q_{m,n})} - e^{-\frac{z^2}{2}}]
\]  

Since the actual feature distribution does not necessarily obey the normal distribution, normalizing only the intermediate layer features will affect the ability of the network to extract features. Therefore, a parameter layer needs to be designed after the features conform to the normal distribution so that the network can learn to the new feature distribution adaptively, and the final new feature distribution obtained will be more favorable to the activation function response compared to the feature distribution without batch normalization.

\[
m = P(R|S) = \frac{P(S \cup M)}{P(S)} = \frac{s}{s-h}
\]

With the application of batch normalization in training deep neural networks, its effectiveness and importance have been widely proven. When designing neural networks, batch normalization is usually placed after the convolutional or fully connected layer and before entering the nonlinear activation function layer.

### 2.2. Experimental Design

At present, since there is no publicly available dataset for live face detection based on binocular stereo images, this paper proposes a self-built binocular live face dataset. The dataset consists of binocular image pairs, which are acquired by two fixed and stereo-calibrated ordinary webcams [10]. The resolution of both webcams is 640 × 480. The pupil distance of the binocular camera is set to 12 cm. As shown in Table 1, a total of 15 targets are invited to participate in the acquisition, and a total of 60 fake face images are produced to simulate a total of three fake face attacks. During the acquisition of real face binocular images, the targets were asked to raise their heads, lower their heads, rotate their heads and faces, and sit in different positions at different distances from the camera. When acquiring a fake face binocular image, the target holding the fake face image was asked to move the fake face image horizontally, vertically, forward, backward, and rotate it in the depth direction at different lighting conditions and different distances. Also, the printed dummy face images were bent inward or outward to different degrees to simulate the three-dimensional structure and facial movement of the human face. After the sample-rich acquisition, a total of 12,000 binocular image pairs were collected from the live binocular face dataset.

| Table 1. Composition of self-built binocular in vivo dataset |
|-------------------------|---------|---------|---------|---------|---------|
| Dataset      | Genuine | Photo   | IPAD    | Cell phone | Computer |
| Image Paris  | 4000    | 5000    | 4000    | 1500       | 2000     |
| Subjects     | 200     | 250     | 200     | 150        | 180      |
Since the combination of binocular depth features for template face alignment and micro-texture features for spatial pyramid coding proposed in this paper and the combination of SSD for deep learning detection network and micro-texture features for spatial pyramid coding are based on images for in vivo determination, each frame of the video is used in the experiments in this section, and this operation is called "video unframing" [11].

The multi-task learning approach is a good way to exploit the correlation between live face detection and face recognition, and the mutual constraint between face recognition and live face detection tasks can avoid model overfitting to a certain extent, which is also useful for improving the generalization of the live face detection algorithm. Therefore, in this section, a multi-task convolutional neural network is constructed using the idea of multi-task learning to train both face detection and face recognition tasks using live face detection data to improve the robustness and generalization of the features extracted by the model.

3. Organization of the Text

When testing against a single dataset, this section only requires testing against Anti-USTIC using the three underlying test networks again and comparing the results with those in the previous sections, as shown in Figure 2.

![Figure 2. Comparison results of the three basic test networks](image)

The analysis of the incorrectly identified samples further shows that the classification network is unable to learn the distinguishing features between live and non-live samples, and the problem becomes more serious as the conditions of the filming equipment, the filming environment, and the changes of the characters' movements change.

To test the face recognition performance, Siaw, the latest face live detection database with the largest number of subjects, was used as the training database. 90 subjects' data were used as the training data and 75 subjects' data were used as the test data according to the evaluation protocol of Siaw. Meanwhile, to test the effect of multi-task training on face recognition, two algorithm models were trained on the Siaw database, one was obtained by training the face recognition network branch using only Regi-Loss in a single task, and the other was obtained by combining the face live detection network branch with multi-task training, and three loss functions were used in the training, Anti During training and testing, real face pictures and fake face pictures of the same subject are considered as one class in face recognition, as shown in Figure 3.
From the results tested on Siaw, the face recognition accuracy can reach 92.6% when combined with face live detection for multi-task training, while the accuracy of single-task training is 83.6%, and the ROC curve of face recognition for multi-task training is also above the ROC curve of single-task training. The test results show that the ML-MTL-CNN algorithm in this chapter has a high accuracy rate of face recognition on the database on Siaw, and the multi-task training method combined with face live detection can effectively improve the accuracy rate of face recognition.

This dataset also provides some testing criteria including 1:1 face verification, so this dataset is widely used for the evaluation of face recognition algorithms and many people verify on this dataset to get a quick idea of the performance of their models. Here the model after multi-task training on the Siaw database is tested on the DEEP face recognition database by taking a face verification test. Many face recognition algorithms have been tested on DEEP, and to make a reasonable comparison, a comparison with human eye recognition, DEEP has been chosen, as shown in Figure 4.

The accuracy of the CNN algorithm on Siaw combined with face live detection for multi-task training of the network model on the DEEP database is 97.2%, which is close to the human eye recognition rate of 97.5%, indicating that the ML-MTL-CNN algorithm in this chapter has a good performance for face recognition. It illustrates that the ML-MTL-CNN algorithm can achieve relatively effective face recognition performance with small sample training. At present, many excellent face recognition algorithms represented by DEEP can achieve face recognition accuracy of more than 99.5% on the DEEP database, compared with this algorithm, there is still a gap in face recognition accuracy of the algorithm in this chapter.
4. Conclusion
The depth corresponding to each pixel position is regressed by a binocular camera model. Before the emergence of binocular stereo matching algorithms based on deep learning, a stable and mature algorithmic process existed for stereo matching tasks that consisted of four steps, however, hand-designed features and regularization functions greatly limited the performance of traditional stereo matching algorithms. The best generalization performance among current face lives detection algorithms. We analyze the basic principle of weight quantization and propose an improvement scheme for the analysis and testing of weight trivialization networks to reduce the hardware requirements and arithmetic power consumption for algorithm model deployment. An innovative deep learning-based detection network is applied to the face live detection task based on the proposed decision-level cascade algorithm, which combines micro-texture features encoded by spatial pyramids for end-to-end face anti-counterfeit determination. Experiments are conducted on three mainstream public live face detection datasets, as well as a self-built live face detection dataset based on binocular cameras, and the experimental results show that the proposed algorithm achieves the current leading level of face anti-counterfeiting results.

Acknowledgments
This work was financially supported by the Scientific Research Foundation of Yunnan Education Department (Research on Face Anti-spoofing based on feature fusion, No. 2021J0550).

References
[1] Jiang B, Yang J, Lv Z, et al. Wearable vision assistance system based on binocular sensors for visually impaired users[J]. IEEE Internet of Things Journal, 2018, 6(2): 1375-1383.
[2] Bhandari V, Rai K, Shrivastava R. IOT Based Wearing Visual Assistance System Based on Binocular Sensors[J]. International Journal, 2020, 8(3): 260-263.
[3] Ni Y, Sun B. A remote free-head pupillometry based on deep learning and binocular system[J]. IEEE Sensors Journal, 2018, 19(6): 2362-2369.
[4] Yang J, Wang C, Jiang B, et al. Visual perception enabled industry intelligence: state of the art, challenges and prospects[J]. IEEE Transactions on Industrial Informatics, 2020, 17(3): 2204-2219.
[5] Khan M A, Paul P, Rashid M, et al. An AI-Based Visual Aid With Integrated Reading Assistant for the Completely Blind[J]. IEEE Transactions on Human-Machine Systems, 2020, 50(6): 507-517.
[6] Kowsalya S, Periasamy P S. Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization[J]. Multimedia Tools and Applications, 2019, 78(17): 25043-25061.
[7] Jiang Q, Zhou W, Chai X, et al. A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9784-9796.
[8] Ruan H, Zou D, Wang W, et al. Online Live Working Safety Monitoring and Early Warning Based on Spatial Cross - Border Prevention[J]. IEEJ Transactions on Electrical and Electronic Engineering, 2020, 15(6): 881-893.
[9] Yang J, Sim K, Gao X, et al. A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route[J]. IEEE Transactions on Image Processing, 2018, 28(3): 1314-1328.
[10] Beddiar D R, Nini B, Sabokrou M, et al. Vision-based human activity recognition: a survey[J]. Multimedia Tools and Applications, 2020, 79(41): 30509-30555.
[11] Chen J, Xu W, Xu H, et al. Fast vehicle detection using a disparity projection method[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(9): 2801-2813.