IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 5, 2019, accepted August 6, 2019, date of publication August 12, 2019, date of current version August 22, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2934563

Face Detection Method Based on Cascaded Convolutional Networks

RONG QI[1], RUI-SHENG JIA[1,2], QI-CHAO MAO[1], HONG-MEI SUN[1,2], AND LING-QUN ZUO[1]

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
[2]Shandong Province Key Laboratory of Wisdom Mine Information Technology, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding authors: Rui-Sheng Jia (jrs716@163.com) and Hong-Mei Sun (shm0221@163.com)

This work was supported in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2018MEE008, and in part by the Key Research and Development Program of Shandong Province, China, under Grant 2017GSF20115.

ABSTRACT Deep learning achieves substantial improvements in face detection. However, the existing methods need to input fixed-size images for image processing and most methods use a single network for feature extraction, which makes the model generalization ability weak. In response to the above problems, our framework leverages a cascaded architecture with three stages of deep convolutional networks to improve detection performance. The network can predict face in a coarse-to-fine manner. We replace the standard convolution with a combination of separable convolution and residual structure in the network. Extensive experiments on the challenging FDDB and WIDER FACE benchmarks demonstrate that our method achieves competitive accuracy to the state-of-the-art techniques while keeps real-time performance.

INDEX TERMS Face detection, cascade convolutional neural networks, depthwise separable convolution, residual structure.

I. INTRODUCTION

Face detection is a basic research problem in the field of computer vision and pattern recognition, as well as a fundamental step of face-related research, such as face verification [1], [2], face recognition [3] and face tracking [4]. The purpose of face detection is to detect human face from video images and provide a basis for subsequent research on face recognition. After decades of development and research, face detection has become a research hotspot in the field of video images, more and more attention has been paid by researchers.

The most classic method of face detection is the VJ face detection method proposed by Viola and Jones in 2001 [5], which uses simple Haar features and cascade AdaBoost classifier for face detection to achieve the efficient and real-time performance of face detection. Since then many scholars have used more features to improve detection accuracy, such as Local Binary Pattern (LBP) [6], Scale-Invariant Feature Transform (SIFT) [7], Histogram of Oriented Gradient (HOG) [8]. However, the performance of this kind of face detector will significantly decrease with the change of face visual diversity in practical applications. In addition to cascade structure, Felzenszwalb proposed HOG based Deformable Part Model (DPM) detection method in 2008 [9], using SVM as a classifier, which can achieve remarkable performance when using a small number of incompletely labeled samples. However, the calculation is too complicated, mostly relying on artificially designed features, lack of stability. ACF [10] uses aggregate channel features for multi-view face detection and achieved great progress in the field of non-depth learning.

In recent years, the face detection method based on convolutional neural network (CNN) has made a breakthrough and become the mainstream of face detection method, which has been applied in various fields of life [11], [12]. Zhang *et al.* proposed to use deep convolutional neural network for face alignment [13]. Yang *et al.* conducted face attribute recognition through deep convolutional neural network [14]. However, the research on face detection in recent years has mainly focused on the uncontrollable part of the face area, such as exaggerated expressions, posture changes, facial occlusion. In the face of so many problems, it is difficult to generate good generalization ability by only relying on a single structure model for detection, which makes the model less robust in practical application. In order to overcome this shortcoming, a series of improved deep learning method has emerged in recent years. Cascade CNN [15] effectively solves the above problems by using a cascade structure, which can

The associate editor coordinating the review of this article and approving it for publication was Lei Wei.

capture various complicated and variable situations in the face region during the process of training a large number of samples. The Faceness network proposed by Yang *et al.* [16] performs face detection by sharing multiple local networks. The MTCNN proposed by Zhang *et al.* [17] uses multi-level network cascading and multi-task training for face detection. HyperFace [18] completed multiple face detection tasks using iterative region selection, key point-based NMS post-processing methods and multi-task learning. Conv3D [19] integrates CNN with 3D face model in an end-to-end multi-task learning framework. UnitBox [20] proposed a new loss function, which treats the four regression values as a whole, which not only improves the accuracy but also accelerates the convergence. FaceBoxes [21] use RDCL, MSCL and anchor densification strategy to achieve real-time face detection in CPU. ICC-CNN [22] uses different layers of the same CNN for cascade.

In practical applications, most of them detect faces in video and require real-time stable detection of faces with large angle changes and large occlusion areas. Although cascading convolutional neural networks have excellent performance in the field of face detection, with the improvement of people's requirements for detection accuracy, the layers of the convolutional neural networks become deeper, the number of parameters increases sharply, training networks and Running the network is extremely time consuming. In this paper, we propose to combine the separable convolution in Mobilenet [23] and the residual structure in Resnet [24] into a separable residual module instead of the standard convolution in the cascade network, our method achieves competitive accuracy while keeps real-time performance.

This paper is divided into five chapters. The first chapter introduces the development status and related background of face detection. The second chapter briefly introduces the related technologies involved in this paper. The third chapter gives a detailed introduction to the separable residual modules and networks we designed. The fourth chapter verifies the superiority of the network we designed through multiple sets of experiments. The fifth chapter summarizes the research content of this paper and looks forward to the next research.

II. RELATED WORK

Depthwise separable convolution is a way of miniaturizing the network model, its essence is to decompose standard convolution into two steps. The first step is the channel-by-channel convolution, one convolution filter corresponds to one channel and one channel is extracted by only one convolution filter. The second step is Pointwise, using 1×1 convolution filters to concatenate the feature maps obtained in the first step to maintain the integrity of the features. This structure can achieve cross-channel information integration while reducing the number of output channels and keep the performance of the method while reducing the amount of computation.

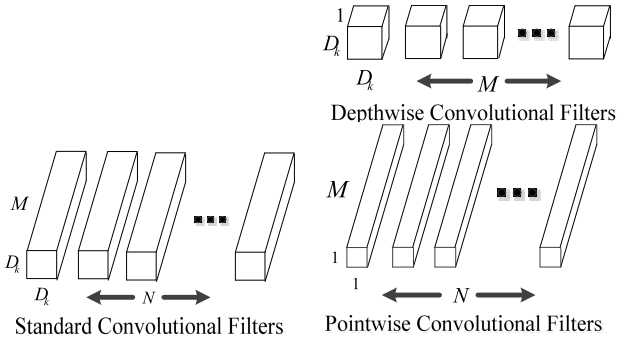


FIGURE 1. Standard convolution and depth separable convolution.

The ratio of the amount of computation of depthwise separable convolution and standard convolution is:

$$\frac{D_k \times D_k \times M \times 1 + 1 \times 1 \times M \times N}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k^2} \quad (1)$$

where D_k is the convolution filter size, M is the input channel and N is the output channel. According to formula (1), the amount of computation required to use a depthwise separable convolution instead of a standard convolution in the network is greatly reduced.

In theory, the increase of network layer can improve the performance of neural network, but in the actual test, the accuracy decreases with the deepening of network layer. This is mainly due to the gradient degradation problem with the deepening of network layers. The introduction of the residual structure solves this problem very well. Compared with the traditional neural network, the residual structure adds a shortcut connection to make the network easier to train and converge.

For the deep network, Resnet optimizes the residual structure by replacing the two convolution layers of 3∗3 with the convolution layer of 1∗1+3∗3+1∗1. The purpose of the first 1∗1 convolutional layer is to reduce the number of input channels to 1/4 of the original number. The purpose of the second 1∗1 convolutional layer is to restore the number of output channels to the original dimension. In this way, the accuracy is maintained and the computation is reduced. The structure is shown in FIGURE 2.

III. FACE DETECTION FRAMEWORK

A. SEPARABLE RESIDUAL MODULE

In order to maintain the advantages of the residual structure accuracy and reduce the computation amount, a new residual module called the separable residual module is designed based on the separable convolution and residual structure. The separable residual module directly adds the shortcut connection to the separable convolution model. The input channel is first reduced to 1/4 of the original input channel through convolution of 1∗1, then extracts features through the 3∗3 channel-by-channel convolution. Finally, the convolution of 1∗1 connects the features and restores them to the original number of channels.

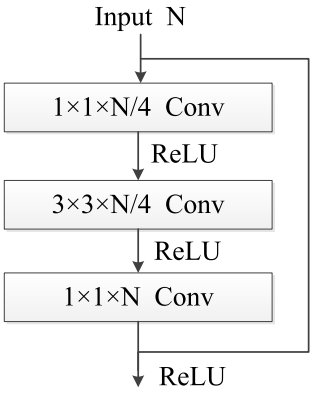


FIGURE 2. Residual structure.

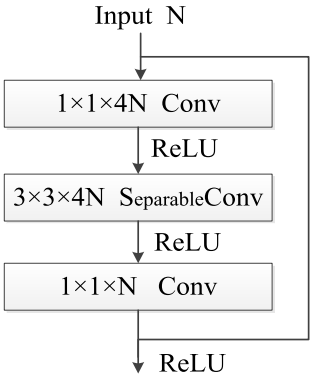


FIGURE 3. Separable residual structure.

The essence of separable convolution is to extract features from each channel and then connect the features. If the residual structure and the separable convolution are directly combined, the input channel has become the original a quarter when performing channel-by-channel convolution, the feature extracted in this way will also be 1/4 of the original feature, which will reduce the detection accuracy of the network. We changed the effect of the first 1∗1 convolution from shrinking the channel to 1/4 to expanding the channel to 4 times. The separable residual module ensures that the network can still maintain a low amount of computation when the number of layers is deep. The separable residual structure is shown in FIGURE 3.

B. FACE DETECTION NETWORK

We proposed the convolutional neural network is a face detection network based on regression theory. The network replaces the standard convolution in MTCNN with the separable residual module and deletes the task of facial landmarks position. Therefore, only two tasks of face classification and bounding box regression are carried out in the detection stage. The detection accuracy of the network is improved by using the cascading convolutional neural network and the method of expanding the channel in the separable residual module. The depthwise separable convolution is used to reduce the amount of network computation to maintain a fast detection speed. The network structure is shown in FIGURE 4.

The purpose of the first stage network is to obtain the region proposal and confidence. First, the fully convolution network [25] is used to extract the features of the input image, FCN can be trained with fixed-size images and tested with images of arbitrary size. Then the face region proposal of the input image is obtained through reverse calculation.

$$x_1 = \frac{(stride * x)}{scale} \quad x_2 = \frac{(stride * x) + cellsize}{scale}$$
$$y_1 = \frac{(stride * y)}{scale} \quad y_2 = \frac{(stride * y) + cellsize}{scale} \tag{2}$$

where (x, y) is the coordinate position of the pixel in the feature map, (x_i, y_i) is the corresponding region proposal coordinate, stride is the step size of the pooled layer, cellsize is the size of the region proposal and scale is the scaling ratio of the current input image and the original image.

After a series of region proposals generated by the first-stage network, the bounding box regression is performed [26]. Finally, the remaining face region proposal is subjected to Non-Maximum Suppression (NMS) to merge the highly overlapping face region proposal, as shown in TABEL I:

TABLE 1. Non-maximum suppression.

Algorithm: Non-Maximum Suppression
Input: region proposal B, confidence s, threshold T;
Output: filtered region proposal B', confidence s;
1. $B' = \{\}$
2. While $B \neq \varnothing$ do
3. $s_k \leftarrow \arg\max s$
4. $B' \leftarrow B' \bigcup B_k; B \leftarrow B - B_k$
5. For B_i in B do
6. If $I(B_k, B_i) \geq T$
7. $B \leftarrow B - B_i; s \leftarrow s - s_i$
8. End
9. End
10. End
11. Return B', s

The second stage network aims to further screen the region proposal of misjudged face region. First, the face region proposal obtained in the first stage network is input and then the wrong face region proposal is filtered by the bounding box regression and the Non-Maximum Suppression. Compared with the first stage network, the second stage network adds a layer of fully connected layer to achieve better filtering effect, so as to obtain more accurate face region proposal. The third stage network adds a layer of convolution layer more than the second stage network, so the result of the screening will be more refined. The third-stage network output final face detection results.

The loss function of this paper is divided into two parts, which are for face classification, bounding box regression. Face classification is a binary classification problem.

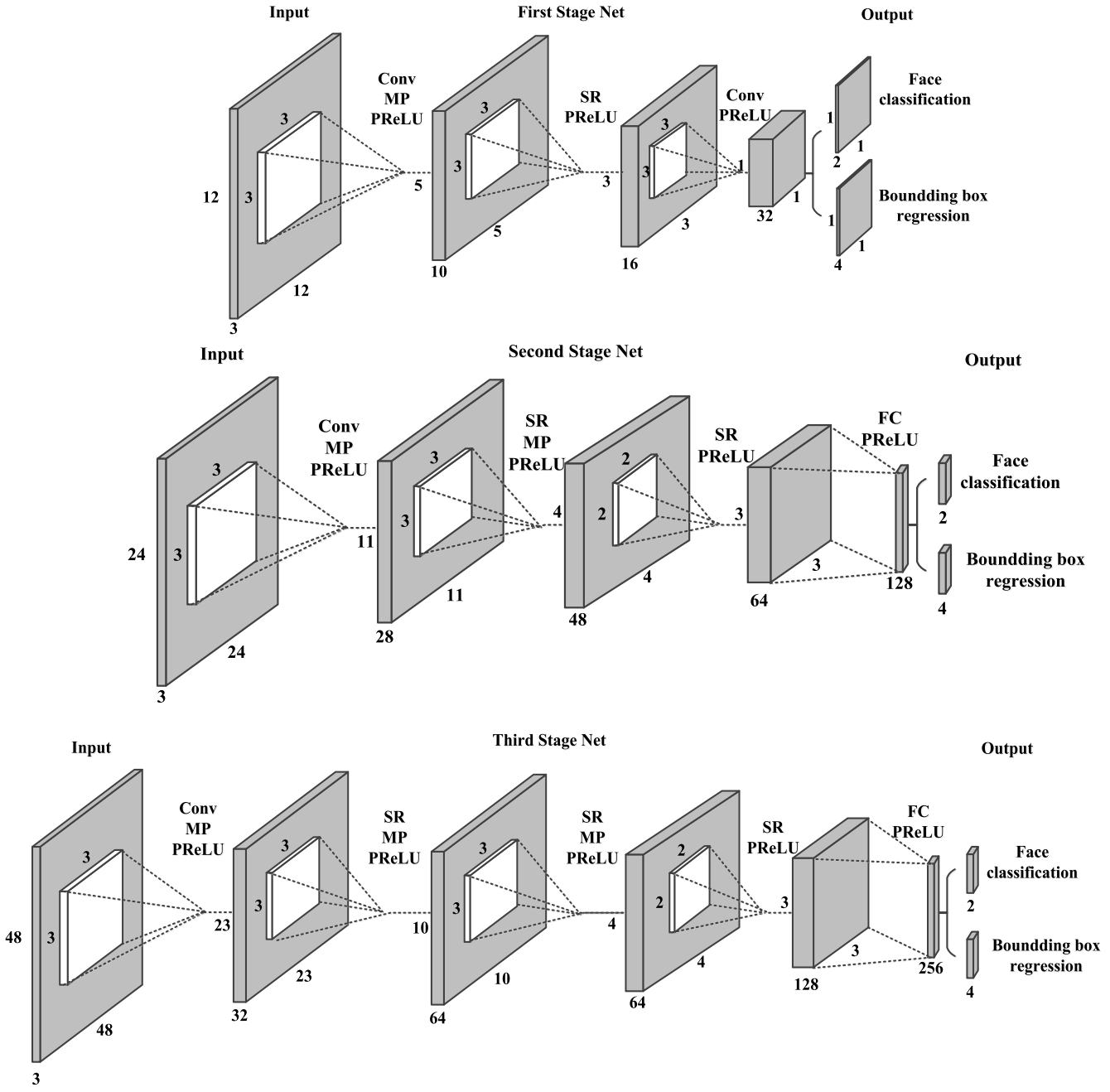


FIGURE 4. The architectures of First Stage Net, Second Stage Net and Third Stage Net, where "MP" means max pooling, "FC" means fully connected, "Conv" means convolution and "SR" means separable residual convolution. The step size in convolution, separable residual convolution and pooling is 1,1 and 2, All convolutional layers and fully connected layers are followed by PReLU layer except the output layers.

We adopt the cross-entropy loss:

$$L_{cls} = -(y_i^{\text{det}} \log(p_i) + (1 - y_i^{\text{det}})(1 - \log(p_i))) \quad (3)$$

where p_i is obtained through the network, which means that sample x_i is the probability of a face. The notation $y_i^{\text{det}}(y_i^{\text{det}} \in \{0, 1\})$ denotes the ground truth label.

For bounding box regression, we use the Euclidean distance loss. Predicting the deviation of each candidate window $\hat{y}_i^{box}$ from its nearest ground truth(y_i^{box}), each bounding box contains four elements: upper-left coordinate, height and width.

$$L_{reg} = ||\hat{y}_i^{box} - y_i^{box}||_2^2 \quad (4)$$

Because the model applies different tasks in each stage of convolutional neural network, different types of training images need to be entered. In this case, the loss function for the above two tasks may not be used at the same time. So the overall learning target can use a loss:

$$Loss = L_{cls} + \alpha L_{reg}, \alpha = 0.5 \quad (5)$$

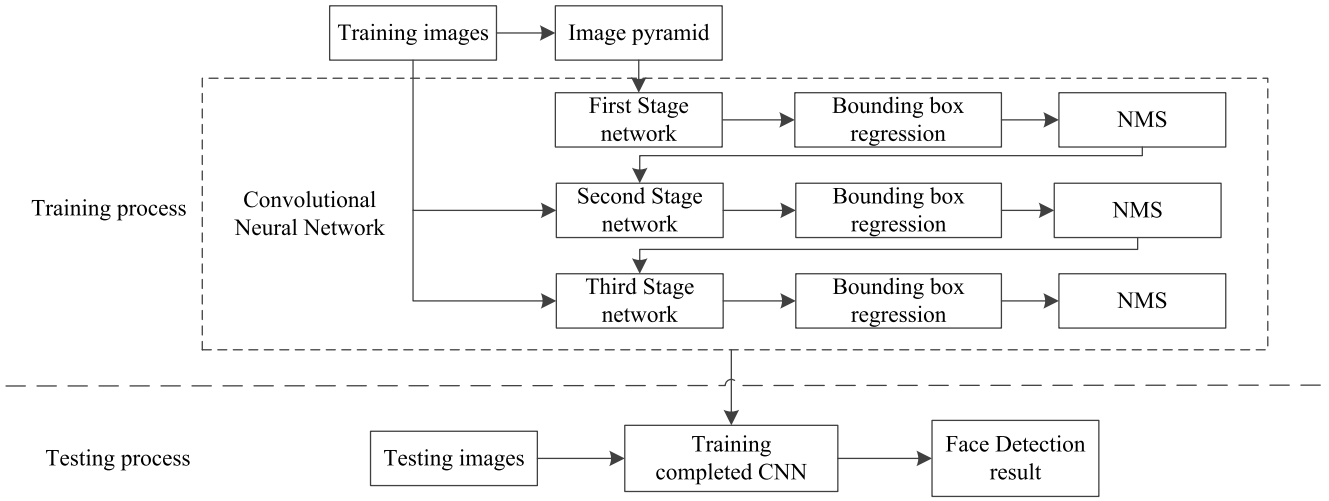


FIGURE 5. Face detection process.

During the training, in order to achieve better detection effect of the method, in each mini-batch, we sorted all samples according to the value of loss function, set the first 70% of samples as effective samples and calculated the gradient only with the effective samples. It reduces the time required for training and improves the network detection effect. Neglected samples have smaller loss function values, indicating that the network has been able to make better predictions for these samples, so these samples have little meaning to the network during the current training process.

C. FACE DETECTION PROCESS

The overall pipeline of our approach consists of two processes: the training process and the testing process. The training process inputs the training image, we initially resize it to different scales to build an image pyramid [27], which is the input of the convolutional neural network and iteratively trains to obtain the final convolutional neural network. The testing process is to input the testing images into the convolutional neural network completed by training to obtain the face detection results. The face detection process is shown in FIGURE 5:

Default parameters: T_1, T_2, T_3 are the face region proposal confidence thresholds for each stage of network. $B = \{B_1, B_2, \cdots B_i\}$ is the output face region proposal set, where B_i is the coordinate vector of the face candidate window, including the upper left coordinate (x_1, y_1) and the lower right coordinate (x_2, y_2). Our specific method steps are as follows:

Step1: Input image $I_0(x, y)$, perform image pyramid [23] preprocessing to obtain set I=$\{I_1, I_2 \cdots I_n\}$.

Step2: The set I is input into the first-stage network to generate face region proposal and the b_i in the face region proposal with a higher confidence than the threshold T_1 is outputted to form the set B.

Step3: The face region proposal in the set B is subjected to bounding box regression and the highly overlapping face region proposal is filtered by the NMS method.

Step4: Map the filtered face region proposal to the original image I_0 and then resize to a 24×24 size window as the input to the second stage network.

Step5: Output b_i in the face region proposal with a confidence higher than the threshold T_2 and update the set B.

Step6: The face region proposal in the updated set B is subjected to bounding box regression and the highly overlapping face region proposal is filtered by the NMS method.

Step7: The filtered face region proposal is mapped to the original image I_0 and then resize to a 48∗48 window as the input to the third stage network.

Step8: Filtering the face region proposal with b_i confidence lower than the threshold T_3 and performing the bounding box regression on the retained face region proposal and removing the highly overlapping face region proposal by using the NMS method, the network output final face detection results.

To much clearer about the effect of the proposed framework, we show some examples for the progressive process of face detection in FIGURE 6.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL DATA AND ENVIRONMENT

Dataset adopted for model training is WIDER FACE dataset [28]. WIDER FACE dataset contains 393, 703 annotated faces with large variations in scale, pose and occlusion in total 32, 203 images. For each of the 60 event classes, 40%, 10%, 50% images of the database are randomly selected as training, validation and testing sets. The network uses three different kinds of data in training process: (1) Negatives:

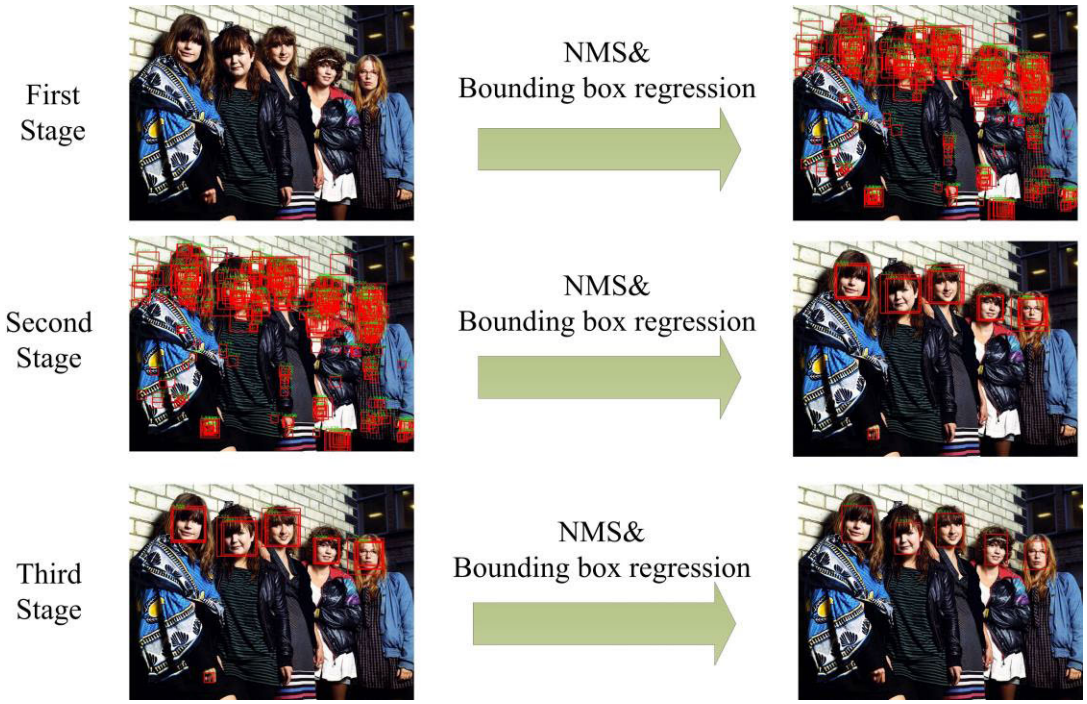


FIGURE 6. some examples for the progressive process of face detection.

images whose the Intersection-over-Union (IoU) with any ground truth face was less than 0.3; (2) Positives: image with IoU greater than 0.65 compared with ground truth face; (3) Part faces: IoU between 0.4 and 0.65 to a ground truth face. As different annotation styles will lead to facial differences, data with IoU between 0.3∼0.4 are discarded. Negatives and positives are used for face classification tasks, positives and part faces are used for bounding box regression.

Experimental test dataset used the WIDER FACE dataset [28] and the FDDB dataset [29]. The WIDER FACE was classified into Easy, Medium and Hard levels according to the degree of Angle and occlusion. FDDB dataset is one of the most authoritative face detection and evaluation platforms in the world, with 2,845 images and a total of 5,171 faces as test sets. These images are taken from natural human faces, which have great diversity in attitude, expression, illumination, sharpness, resolution and shielding degree and are close to the real application scenes.

The experimental software environment is the operating system Ubuntu 16.04, The deep learning framework is Tensorflow. The experimental hardware environment is Intel Core i7 8700 processor GPU for NVIDIA GTX 1080.

B. ANALYSIS ON BOUNDING BOX REGRESSION

According to our experimental results, the binary classification is easier to converge than other nonlinear regressions, which leads to over-fitting in model training. Bounding box regression can be used as a regularization factor for face classification and overfitting can be avoided by inhibiting the convergence rate of face classification. In order to better understand bounding box regression, we retrained a model without bounding box regression to compare with the original model. The results are shown in TABLE 2. The gap between training and testing narrows with bounding box regression, which means that the generalization ability of the model is enhanced, the window after regression is closer to ground truth and the confidence of the window is also got an improvement.

TABLE 2. Comparison of bounding box regression.

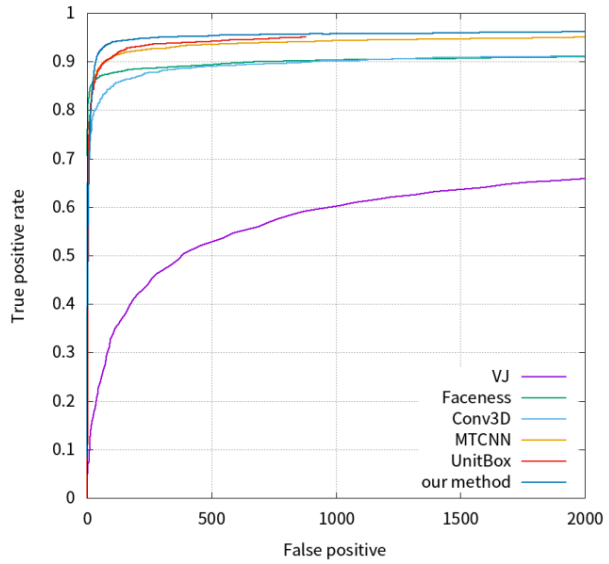
Bounding box regression	Train Accuracy	Test Accuracy
×	97.9%	95.3%
√	97.1%	96.1%

TABLE 3. AUC comparison with other state-of-the-art methods.

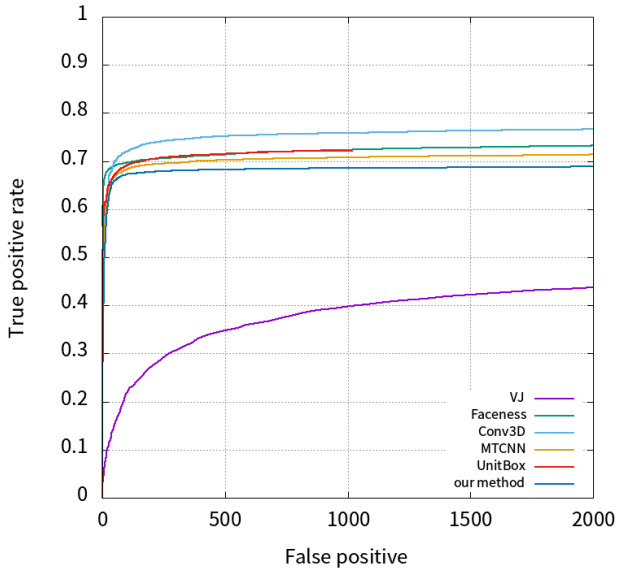
Method	AUC
ViolaJones	0.559
Faceness	0.899
Conv3D	0.912
MTCNN	0.938
Proposed method	0.947

C. COMPARSION ON BENCHMARKS

To evaluate the performance of our method, we compare our method against the state-of-the-art methods in FDDB. The evaluation indicators include: recall rate is used to evaluate the proportion of the detected face to the total face of the sample mark; false positive is the number of errors in the detected face. These two indicators are expressed by the ROC



(a)Discontinous ROC curves



(b)Continous ROC curves

FIGURE 7. Comparison of different face detectors on FDDB dataset. We show the ROC curves on the (a) Discontinous ROC curve and (b) Continous ROC curve.

(Receiver Operating Characteristic) curve. The results are shown in FIGURE. 7(a) and FIGURE.7(b).

The ROC curve detection results show that the traditional face detection method VJ recall rate is only 66.6%, the detection method based on deep learning has been greatly improved. Our method achieves state-of-the-art performance in terms of both the discrete ROC curve and continuous ROC curve. Our discrete ROC curve is superior to the MTCNN. We also obtain the best true positive rate of the discrete ROC curve at 2000 false positives (96.1%). In addition, the possible influencing factor is that our method is not very effective in detecting the side face. It is easy to miss the side face and the face blocked by the object in the detection. FIGURE 8. shows some qualitative results on the FDDB.

The ROC curve does not clearly indicate which method is better, so another indicator AUC is used to illustrate the pros and cons of the method. AUC represents the area proportion under the ROC curve and the value is between 0 and 1. The higher the AUC value is, the better the method performance will be.

Then test on the WIDER FACE dataset, WIDER FACE is a more challenging benchmark than FDDB in face detection. The test results are shown in TABLE 4. It is very encouraging to see that our model consistently achieves the competitive performance across the three subsets. It has higher robustness for faces with large occlusion and Angle change, which is basically consistent with the evaluation results in the FDDB dataset. FIGURE 9. shows more examples to demonstrate the effects of our method on handling faces with various variations. The experimental results show that the proposed method has good robustness in the real environment.

D. RUNTIME EFFICIENCY

To verify the real-time performance of our method, we compared with face detection methods such as Faceness, ICC-CNN, MTCNN and FaceBoxes. Experimental results are all based on the FDDB dataset. The detection time is the average detection time of all the pictures in the FDDB database. As can be seen from TABLE 5, the detection speed of our method reaches 28FPS, which is slightly lower than MTCNN, such computation speed is quite fast among the state-of-the-art, meeting the real-time requirements of video detection.

Then compare the network model we proposed with the parameters of MTCNN. It can be seen from the data in TABLE 6 that the number of parameters of our model is slightly higher than that of MTCNN, but the detection

TABLE 4. Detection performance comparison on wider face.

Method	Detection average accuracy		
	Easy set	Medium set	Hard set
ACF	0.695	0.588	0.290
Faceness	0.716	0.604	0.315
Cascade CNN	0.711	0.636	0.401
MTCNN	0.849	0.823	0.602
Proposed method	0.869	0.847	0.664

TABLE 5. Speed comparison with other state-of-the-art methods.

Method	Detection time / second	Speed / fps
Faceness	0.119	9
ICC-CNN	0.063	16
MTCNN	0.034	30
FaceBoxes	0.028	36
Proposed method	0.036	28

TABLE 6. Comparison of network structure parameters.

Method	P-Net/k	R-Net/k	O-Net/k
MTCNN	6.83	100.66	388.5
Proposed method	6.79	108.28	398.9

FIGURE 8. Face detection examples in the FDDB dataset.

FIGURE 9. Face detection examples in the WIDER FACE dataset.

accuracy is greatly improved. It can be seen from the comparative experiment that the reasonable use of depthwise separable convolution can effectively reduce the computational burden.

V. CONCLUSION

In order to solve the problems of weak generalization ability of single structure and large network parameters, we propose a cascade convolutional neural network based on separable residual convolution with superior performance on both speed and accuracy. The network uses a cascade structure to generate face candidate windows, multi-layer network screening can effectively improve the accuracy of candidate window positioning. The expansion of input channel for residual structure improves the accuracy of face detection, the separable convolution enables network to achieve real-time speed. The experimental results show that the proposed method has some advantages compared with other methods, which can improve the accuracy of face detection and ensure real-time performance.

REFERENCES

[1] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *Proc. IEEE Int. Conf. Autom. Face Gesture*, Jun. 2017, pp. 1–8.
[2] A. Majumdar, R. Singh, and M. Vatsa, "Face verification via class sparsity based supervised encoding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1273–1280, Jun. 2017.
[3] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.
[4] M. H. Khan, J. McDonagh, and G. Tzimiropoulos, "Synergy between face alignment and tracking via discriminative global consensus optimization," in *Proc. IEEE ICCV*, Oct. 2017, pp. 3811–3819.
[5] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[6] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.
[9] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[10] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep./Oct. 2014, pp. 1–8.
[11] Y. Ren, Y. Sun, X. Jing, Z. Cui, and Z. Shi, "Adaptive makeup transfer via bat algorithm," *Mathematics*, vol. 7, no. 3, p. 273, 2019. doi: 10.3390/math7030273.
[12] Z. Cui, L. Du, P. Wang, X. Cai, and W. Zhang, "Malicious code detection based on CNNs and multi-objective algorithm," *J. Parallel Distrib. Comput.*, vol. 129, pp. 50–58, Jul. 2019. doi: 10.1016/j.jpdc.2019.03.010.
[13] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.
[14] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "From facial parts responses to face detection: A deep learning approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3676–3684.
[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5325–5334.
[16] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-net: Face detection through deep facial part responses," 2017, *arXiv:1701.08393*. [Online]. Available: https://arxiv.org/abs/1701.08393
[17] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.
[18] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 121–135, Jan. 2019.
[19] Y. Li, B. Sun, T. Wu, Y. Wang, and W. Gao, "Face detection with end-to-end integration of a convnet and a 3D model," in *Proc. ECCV*, 2016, pp. 420–436. [Online]. Available: https://arxiv.org/abs/1606.00850
[20] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520.
[21] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "FaceBoxes: A CPU real-time face detector with high accuracy," in *Proc. IJCB*, Oct. 2017, pp. 1–9.
[22] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual CNN," in *Proc. ICCV*, Oct. 2017, pp. 3171–3179.
[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: https://arxiv.org/abs/1512.03385
[25] Y. Bai, W. Ma, Y. Li, L. Cao, W. Guo, and L. Yang, "Multi-scale fully convolutional network for fast face detection," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*. Surrey, U.K.: BMVA Press, Sep. 2016, pp. 1–12.
[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, Jul. 2017, pp. 936–944.
[28] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. CVPR*, Jun. 2016, pp. 5525–5533.
[29] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," Univ. Massachusetts Amherst, Tech. Rep. UM-CS-2010-009, 2010.

RONG QI was born in Shandong, China, in 1994. He received the B.S. degree from the University of Jinan, China, in 2017. He is currently pursuing the M.S. degree with the Shandong University of Science and Technology. His research interests include image processing and deep learning.

RUI-SHENG JIA is currently a Full Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He has more than 30 first-author publications and has more than 25 coauthor publications. His research interests include artificial intelligence, big data processing, information fusion, and microseismic monitoring and inversion.

QI-CHAO MAO was born in Shandong, China, in 1995. He received the B.S. degree from the Shandong University of Technology, China, in 2017. He is currently pursuing the M.S. degree with the Shandong University of Science and Technology. His research interests include image processing and deep learning.

HONG-MEI SUN received the B.S. and M.S. degrees in computer science from the Shandong University of Science and Technology, China, in 1995 and 2005, respectively, where she is currently a Lecturer with the College of Computer Science and Engineering. She is also the Leader of the Key Research and Development Projects of Shandong Province, China. She has four first-author publications and has five coauthor publications. Her research interests include microseismic monitoring technology and software engineering.

LING-QUN ZUO was born in Shandong, China, in 1991. He received the B.S. degree from Dezhou University, China, in 2017. He is currently pursuing the M.S. degree with the Shandong University of Science and Technology. His research interests include image processing and deep learning.