ISSN: 1813-4890

An Integrated Neural Network Model for Efficient Face Detection in Complex Backgrounds

Jie Huang

Tianjin University of Technology and Education, Electronic Engineering Department, Tianjin 300222, China;

2327240272@qq.com

Abstract

A face detection approach which is named MobileNetV3-MTCNN was proposed to enhance the precision of small sized face detection in complex backgrounds. On the backbone of MTCNN, this approach uses the lower computational MobileNetV3 bottleneck module which is in the MobileNetV3 to replace the convolutional function and abandon the common convolution which occupies computer resources to configure the network's feature extraction function.. Experimental results show that this method can greatly improve the model detection accuracy. Compared with MTCNN, the accuracy is improved by 4.8%, 6.2%, and 7.5% on Easy, Medium, and Hard validation sets, respectively, which can improve the accuracy of the model for small-size and multi-angle face detection in complex backgrounds.

Keywords

Complex background, small-sized faces, face detection, multi-task cascaded convolutional neural network.

1. Introduction

Face detection technology is widely used in attendance systems, unlocking devices, identity verification, surveillance places, and automatic driving., etc.[1] . In the current face detection field, deep neural network architecture is usually used. In 2014 Girshick et al. proposed R-CNN[4]target detection algorithm model successfully applied deep learning to the field of target detection, this target detection algorithm uses a candidate region-based detection method. Ren et al. proposed FasterR-CNN based on FastR-CNN[5], which proposes a dedicated candidate region network, and further optimizes the candidate region generation part on top of FastR-CNN, and in further speed improvements. In addition to the above two-step target detection network models, there are also network models based on single target detection, such as YOLO[6] and SSD[9] . The advantage of this type of approach is the fast detection speed, but the detection of small targets is not effective. These deep neural networks are very resource-consuming to deploy in edge devices, with high requirements on hardware computational power and energy consumption, and are difficult to be applied in real-world scenarios. MTCNN[10] (Multi-task Cascaded Convolutional Networks), as a classical face detection method, is known for its efficient performance and low model complexity, and it is more suitable for applications in edge devices. However, with the increasing complexity of face detection tasks, MTCNN faces a series of challenges, such as decreased detection effect in small sizes, occlusion, multiple angles, and illumination changes. Literature 11 combines MTCNN with VGGNet to improve the network detection accuracy, although the accuracy of the model detection set is improved, the corresponding model computation becomes more. Literature 12 combines MobileNet with MTCNN and proposes Mobile MTCNN[12] scheme to reduce the floating-point

ISSN: 1813-4890

number of operations required by the network, but it also leads to a decrease in model detection accuracy.

Aiming at the above problems, this paper proposes an improved model MobileNetV3 -MTCNN, which reconstructs the feature extraction network by the Bottleneck block in MobileNetV3 with MTCNN as the backbone network. With these improvements, the method proposed in this paper demonstrates better performance than MTCNN in face detection tasks in complex backgrounds, effectively improving the accuracy and robustness of face detection in complex backgrounds. This feature makes the method in this paper has a broad application prospect in densely populated scenarios such as subways and shopping malls.

2. Introduction to the methodology

2.1. Principles of the MTCNN algorithm

In deep learning, multi-task cascade convolutional neural network (MTCNN) is a commonly used face detection model, and the main design concept is to gradually achieve the task of face detection through cascading multiple convolutional neural networks, and the model detection process is shown in Fig. 1.MTCNN consists of three networks respectively P-Net, R-Net, and O-Net, and the input of the network adopts the form of image pyramid, and the original input image is scaled into a series of different sizes and fed into P-Net for processing to generate a large number of candidate frames. The face candidate frames generated by P-Net are cropped on the original image and then fed into R-Net for processing. R-Net network simplifies a large number of candidate frames, removes non-face candidate frames, and corrects the coordinates and confidence level of the face frames generated by P-Net network. Finally, the R-Net results are fed into the O-Net network for processing to output the detection results of the whole model.

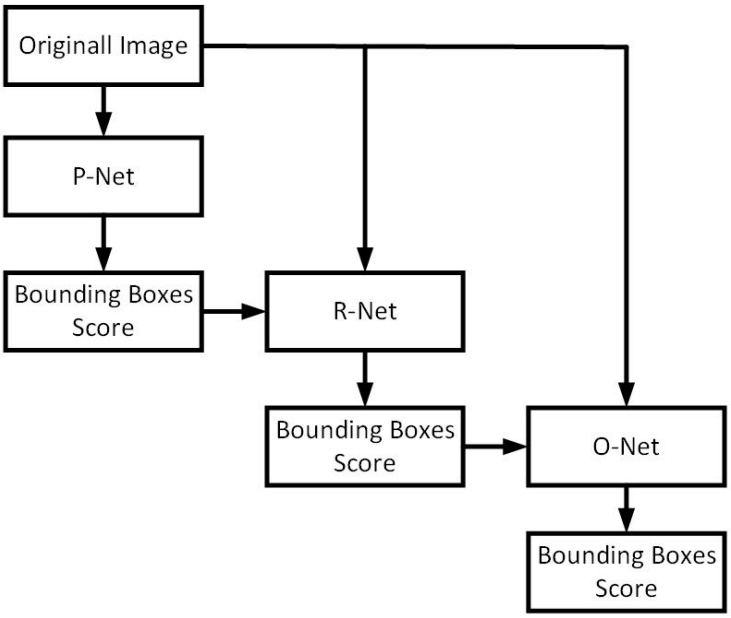


Fig. 1 MTCNN detection process

2.2. MobileNetV3 network

MobileNetV3 is a new lightweight neural network architecture that is better for efficient computation on edge devices and generates more feature maps through cheap operations.MobileNetV3 is constructed by two different step size of Bottleneck modules as shown in Fig. 2. In Bottleneck module the first 1x1 pointwise convolutions are used to increase the number of channels, which is beneficial for the network to perform feature extraction. The

ISSN: 1813-4890

middle 3x3 Depthwise Convolution performs feature extraction and the second 1x1 pointwise convolutions decreases the number of channels to reduce the model complexity.

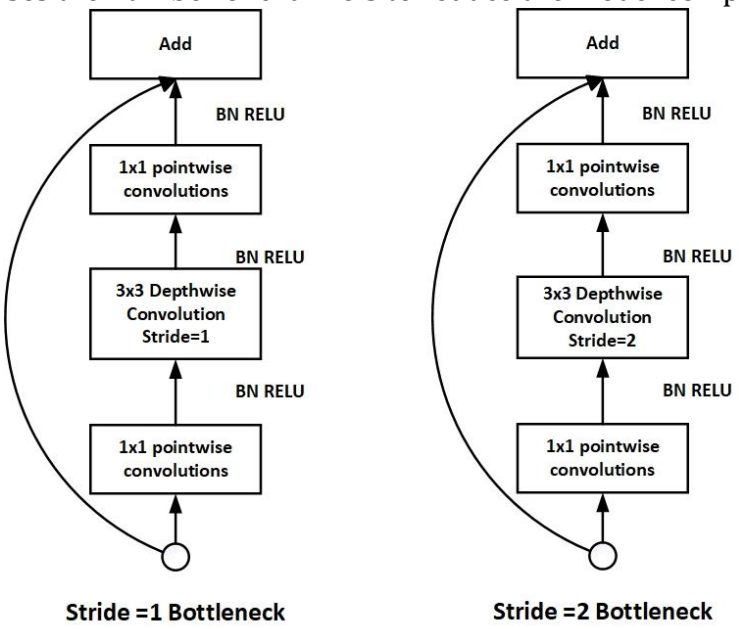


Fig 2 MobileNetV3 Bottleneck Block

2.3. Feature extraction network improvement based on MobileNetV3

The MobileNetV3-MTCNN model proposed in this paper takes advantage of the MobileNetV3 Bottleneck module to reconstruct the three networks, P-Net, R-Net, and O-Net, using the MobileNetV3 bottleneck block based on the MTCNN network architecture. In order to better feature extraction of the image data, the first convolutional layer of the three networks P-Net, R-Net, O-Net uses the traditional convolutional layer, and the subsequent convolutional layers use the MobileNetV3 bottleneck block to replace the network, and the network structures of P-Net, R-Net, O-Net are shown in Tables 1-3. The input image pixel size of the P-Net is 12 × 12 × 3 after a 3 × 3 common The pixel size of P-Net input image is 12×12×3 after 3×3 ordinary convolutional layer and maximum pooling layer to 6×6×16, and further feature extraction is performed by MobileNetV3 bottleneck block with step size 1 and 2, and the height and width of the image is changed to 1×1 by 3×3 convolution, and finally the face confidence and bounding box are calculated by using two point-by-point convolution with 1x1, because the P-Net network does not contain a fully connected layer, so the network can deal with different pixel sizes. Because the P-Net network does not contain a fully connected layer, the network can handle images with different pixel sizes, and the images are fed into the network after a series of scaling operations to generate face candidate frames of different sizes.The fixed pixel size of the input image to the R-Net is 24 × 24 × 3, and the pixel size of the image is changed to 12 × 12 × 28 after the feature extraction by the ordinary convolutional layer of 3 × 3 and the maximal pooling layer, which is obtained by a MobileNetV3 bottleneck of one step of 1 and two steps of 2, and then the pixel size is changed to 12 × 12 × 28. MobileNetV3 bottleneck block for further feature extraction the image pixel size becomes 3×3×64, after 1x1 point-by-point convolution, 3×3 maximum pooling layer, 1x1 point-by-point convolution the image pixel size becomes 1×1×128, and finally the face confidence and bounding box offset are generated through the two full joins, which achieves to streamline a large number of face generated by the P-Net candidate frames for streamlining.The fixed pixel size of the O-Net input image is 48×48×3, after 3×3 ordinary convolutional layer and maximum pooling layer the image pixel size becomes 24×24×32, and further feature extraction is carried out by one MobileNetV3 bottleneck block with a step size of 1 and three steps with a step size of 2 the image pixel size becomes 3×3×256. After 1x1 point-by-point convolution, 3×3 maximum pooling layer, 1x1

ISSN: 1813-4890

point-by-point convolution the image pixel size becomes 1×1×256 and finally the final result of the model is generated by two full connections. The improved model detection process is shown in Fig. 3, the picture is processed by P-Net to generate a large number of face candidate frames of different sizes, the faces in the candidate frames are cut and fed into the R-Net network, face screening is performed to remove redundant candidate frames, and candidate frames that are judged as the presence of a face are cut and fed into the O-Net for the final determination of the network.

Table 1 P-Net network structure

Input	Operator	Exp size	#Out	SE	NL	s
12x12x3	Conv2d 3x3	-	16	-	HS	2
6x6x16	Bneck, 3x3	16	16	-	RE	1
6x6x16	Bneck, 3x3	96	64	-	RE	2
3x3x64	Conv2d 3x3	-	32	-	RE	1
1x1x32	Conv2d 1x1 NBN	-	1	-	-	1
1x1x32	Conv2d 1x1 NBN	-	4	-	-	1

SE denotes whether there is a Squeeze-And-Excite in that block. NL denotes the type of nonlinearity used. Here, HS denotes h-swish and RE denotes ReLU. NBN denotes no batch normalization. s denotes stride.

Table 2 R-Net network structure

Input	Operator	Exp size	#Out	SE	NL	s
24x24x3	Conv2d 3x3	-	28	-	HS	2
12x12x28	Bneck, 3x3	28	28	-	RE	1
12x12x28	Bneck, 3x3	72	48	-	RE	2
6x6x48	Bneck, 3x3	240	64	✓	RE	2
3x3x64	Conv2d 1x1	-	960	-	HS	1
3x3x960	Pool 3x3	-	-	-	-	-
1x1x960	Conv2d 1x1 NBN	-	128	-	HS	1
128	Linear	-	1	-	-	-
128	Linear	-	4	-	-	-

Table 3 O-Net network structure

Input	Operator	Exp size	#Out	SE	NL	s
48x48x3	Conv2d 3x3	-	32	-	HS	2
24x24x32	Bneck, 3x3	32	32	-	RE	1
24x24x32	Bneck, 3x3	128	64	-	RE	2
12x12x64	Bneck,5x5	256	128	✓	RE	2
6x6x128	Bneck, 3x3	240	256	-	HS	2
3x3x256	Conv2d 1x1	-	960	-	HS	1
3x3x960	Pool 3x3	-	-	-	-	-
1x1x960	Conv2d 1x1 NBN	-	256	-	HS	1
256	Linear	-	1	-	-	-
256	Linear	-	4	-	-	-

ISSN: 1813-4890

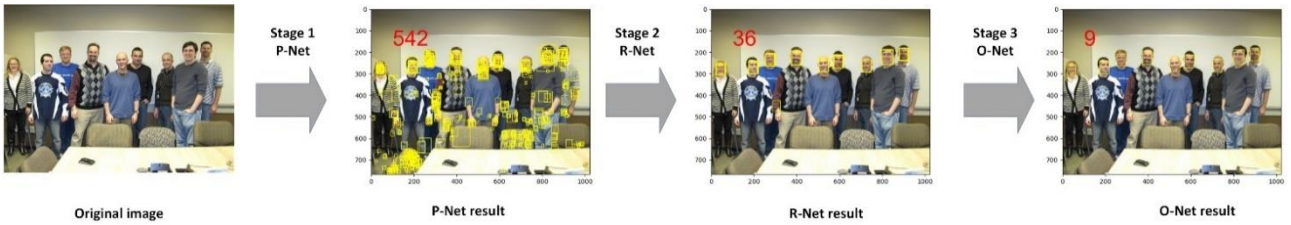


Fig. 3 MobileNetV3 -MTCNN model detection flow

3. Analysis of experimental results

3.1. Experimental Parameter Settings

The hardware configuration used for the experimental tests in this paper is Intel(R) Core(TM) i5-13500HX ,NVIDIA RTX 4050, running memory 16GB, training framework Pytorch 1.10.2, Cuda 12.0, programming language version Python 3.6, and the operating system is Windows 11.The public dataset of WIDER_FACE is selected for the experiments, in which there are 12793 training images containing face data in different scenes. WIDER_FACE public dataset for the experiment, in which there are 12,793 training images, including face data in different scenes, and some parameters in the experiment are shown in Table 4.

Table 4 Partial experimental parameters

parameter name	Parameter description	parameter value
Batchsize	Number of samples in a single training session	[128,640,640]
Learningrate	Initial network learning rate	[0.01,0.005,0.05]
Lr_factor	Learning rate decay factor	[0.6,0.1,0.1]
Thresh	Per-layer network confidence thresholds	[0.6,0.7,0.7]
Min_face	Minimum Face Size	12

3.2. evaluation parameters

Neural networks measure the strength of the model usually comparing the model's AP (Average Precision) value, Precision, Recall, model parameters, etc. The AP value is the average of the precision rates at different levels of recall, which provides a score to summarise the model's performance at the classification threshold. The precision rate is the number of faces correctly detected by the model divided by the total number of faces detected by the model. A high precision rate means fewer false positives (i.e., incorrectly recognising non-faces as faces). Recall measures the proportion of faces recognised by the model to the total number of actual faces, which is the number of correctly detected faces divided by the total number of actual faces. A high recall means fewer false negatives (i.e., missed faces). Model parameters refer to the learnable elements that make up a neural network model, such as weights and biases, and the number of model parameters usually affects the complexity and computational requirements of the model.

In this paper, the model is evaluated using the number of model parameters, Precision (P), Recall (R), and Accuracy (AP) as shown in equations (1)-(3).

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 P(R)dR \tag{3}$$

ISSN: 1813-4890

In the above equation: TP stands for positive samples that are correctly classified, FP stands for negative samples that are misclassified, and FN stands for positive samples that are misclassified; AP stands for average accuracy of face detection.

3.3. Comparison of experimental results between the new method and other algorithms

In this paper, Original-MTCNN and MobileNetV3-MTCNN networks proposed in this paper were trained using the same dataset and training parameters respectively. Tests and evaluations were performed using the validation set on the WIDER_FACE dataset, which contains 3200 images. In order to better evaluate the detection effect of the model, the confidence level of the three networks, P-Net, R-Net, and O-Net, is set to 0.7, 0.8, and 0.8 respectively. the precision rate, recall rate and other indexes of the different models for face detection are tested for the three different types of face images, namely, EASY, MEDIUM, and HARD. The various evaluation indexes of the models are shown in Table 5-Table 7.

Table 5 Easy validation set face type test results

modelling	AP (%)	Recall rate (%)	Precision rate (%)	Number of participants (M)
Original-MTCNN	64	66.8	89.2	1.88
MobileNetV3-MTCNN	68.8	74.2	86.2	3.77

Table 6 Medium validation set face type test results

modelling	AP (%)	Recall rate (%)	Precision rate (%)	Number of participants (M)
Original-MTCNN	61.6	63.8	93.1	1.88
MobileNetV3-MTCNN	67.8	70.9	91.6	3.77

Table 7 Hard validation set face type test results

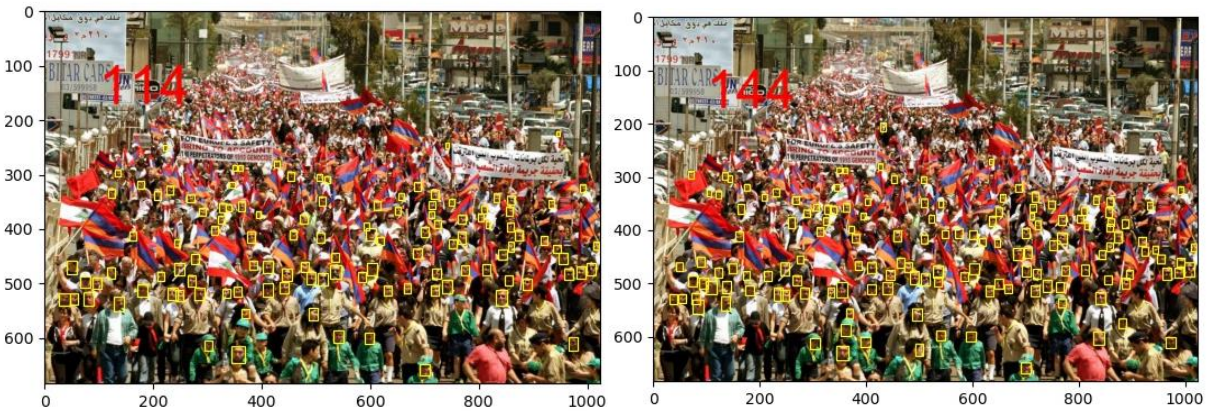
modelling	AP (%)	Recall rate (%)	Precision rate (%)	Number of participants (M)
Original-MTCNN	40.9	41.8	96.8	1.88
MobileNetV3-MTCNN	48.4	49.7	94.8	3.77

As shown in Tables 5-Table 7, the MobileNetV3-MTCNN model has a significant improvement in the two metrics of AP value and recall rate compared to the MTCNN model. Specifically, improves the AP value by 4.8% and the recall by 7.4% in the face type of easy; improves the AP value by 6.2% and the recall by 7.1% in the face type of medium; and improves the AP value by 7.5% and the recall by 7.9% in the face type of hard . The proposed model in this paper has a very large improvement in the accuracy and comprehensiveness of face detection compared to the original model although the model parameters have increased by 1.89M.

In order to see the detection effect of the model more intuitively, images of small-size single-angle faces as well as small-size multi-angle faces are selected for model testing. Fig.4 and Fig. 5 compare the detection results of the two models under the same picture, respectively. Figure 4 shows the detection results under the small size single angle face, Figure 4(a) shows that the original MTCNN detects a total of 114 faces for the small size single type images, and Figure 4(b) shows that MobileNetV3-MTCNN detects a total of 144 faces for the small size single type images, and the model improved in this paper detects 30 more faces than the original model,

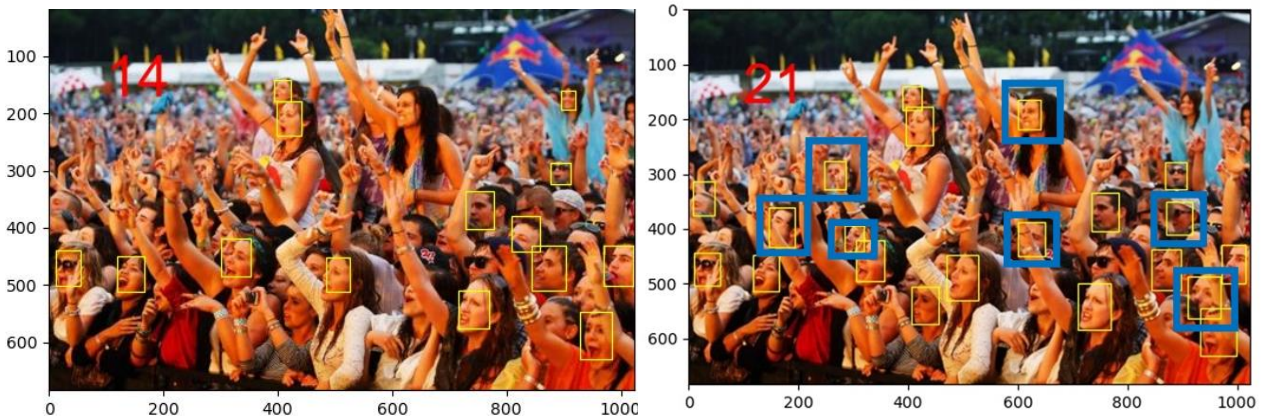
ISSN: 1813-4890

and it can be see that the model proposed in this paper can detect more small-size single-angle faces. Figure 5 shows the detection results of the two models in small-size multi-angle faces, Figure 5(a) shows that the original MTCNN model detected a total of 14 faces, Figure 5(b) shows that MobileNetV3-MTCNN detected 21 faces more than the original model detected 7 faces, in which the blue rectangular box of the faces in the figure is the face detected more than the original model. In summary in complex environments the model proposed in this paper can detect more small-size single-angle and small-size multi-angle faces compared to other models, which reflects the better detection results after the algorithmic improvements in this paper.



(a) Original detection results (b) MobileNetV3-MTCNN detection results

Fig. 4 Small size single angle face model detection results



(a) Original detection results (b) MobileNetV3-MTCNN detection results

Fig. 5 Small size multi-angle model inspection results

3.4. model ablation

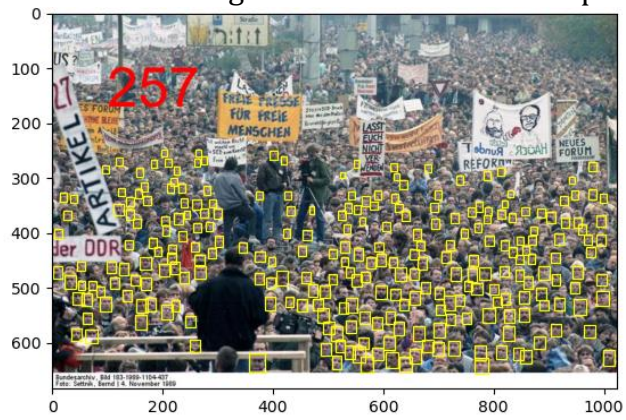
In order to verify the effectiveness of this paper's method on the performance improvement of MTCNN, ablation experiments are designed for the face categories with the validation set of HARD in both training and validation sets are the same, and the results are shown in the following table. Since MTCNN is a multilevel cascade structure, the three networks P-Net, R-Net, and O-Net are improved sequentially.

Table 8 Ablation experiments

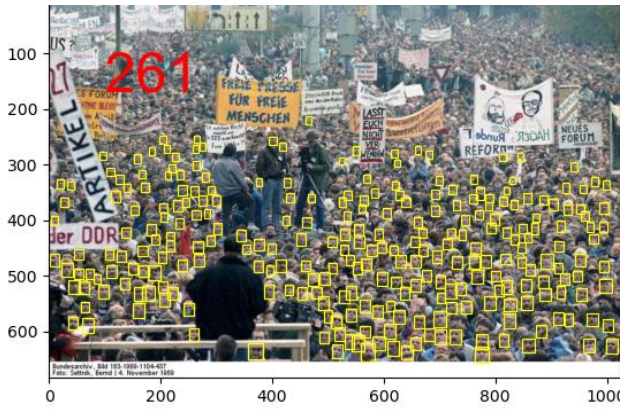
base network	P-Net	R-Net	O-Net	AP (%)	Recall rate (%)	Precision rate (%)
MobileNetV3 - MTCNN				40.9	41.8	96.8
			√	42.6	43.6	95.3
		√	√	47.7	49	94.9
	√	√	√	48.4	49.7	94.8

ISSN: 1813-4890

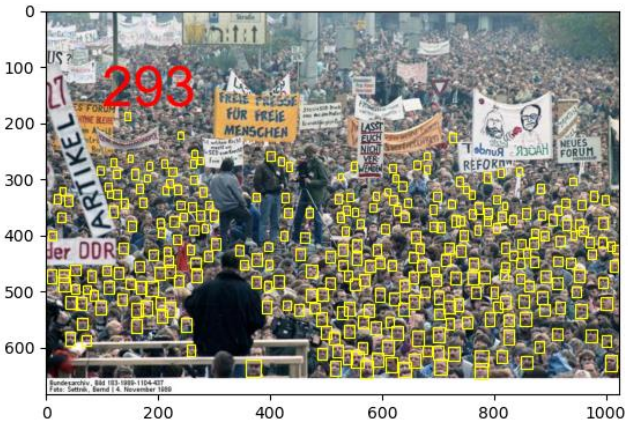
According to the results in Table 8, modifying the three network structure models of P-Net, R-Net, and O-Net sequentially, the detection accuracy is gradually improving, and its detection results are as follows. The detection results of the original network shown in Fig. 6(a), the total number of detected faces is 257; the detection results after modifying the O-Net network shown in Fig. 6(b), the total number of detected faces is 261; the detection results after modifying the R-Net, O-Net network shown in Fig. 6(c), the total number of detected faces is 293; the detection results after modifying the P-Net, R -Net, O-Net network, the total number of detected faces is 308. With the modification of the network structure the model detection ability rises gradually. In summary, the MobileNetV3-MTCNN model proposed in this paper has a better detection effect for detecting small-sized faces in complex environments.



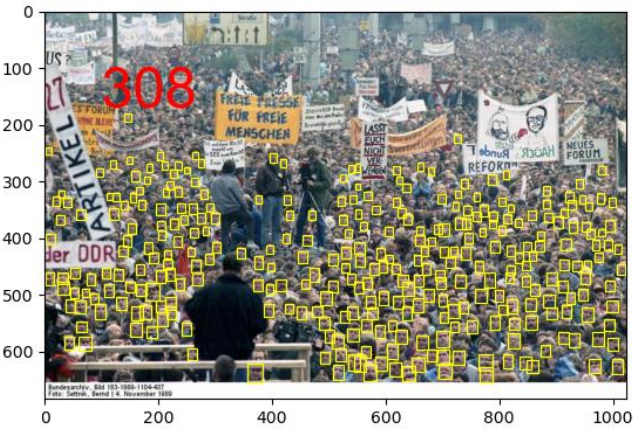
(a) Original network detection results



(b) Improved O-Net network detection results



(c) Improved R, O-Net network detection results



(d) Improved P, R, O-Net network detection results

Fig. 6 Model ablation experiment detection results

4. Conclusion

In order to provide better detection of small-size multi-angle faces in complex backgrounds, and at the same time, it can be more easily achieved by deploying the model on edge devices, this paper proposes a new neural network model (MobileNetV3-MTCNN), which is based on MTCNN, and reconstructs the feature extraction network by using the MobileNetV3 bottleneck module. The comparison experiments with the original MTCNN and the model ablation experiments prove that the proposed model can better achieve the detection of small-sized faces in complex scenes. After the above improvement, good detection results can be achieved in scenes with high human traffic.

ISSN: 1813-4890

References

[1] ROWLEY H A, BALUJA S, KANADE T: Neural network-based face detection. IEEE Transactions on pattern analysis and machine intelligence, (1998) 20(1), p.23-38.

[2] V IOLA P, JONES M J. :Robust real-time face detection. International journal of computer vision, (2004) 57, p.137-154.

[3] ZHAO Z.: Application of improved CNN-based face detection technology in public administration. Journal of Computational Methods in Sciences and Engineering,(2023) 23(4), p.1985-1997.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al.: Rich feature hierarchies for accurate object detection and semantic segmentation //Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) , p.580-587.

[5] REN S, HE K, GIRSHICK R, et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2017) 39(6),p.1137-1149.

[6] H REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection, Proceedings of the IEEE conference on computer vision and pattern recognition. (2016), p.779-788.

[7] R EDMOM J, F ARHADI A. YOLO9000: better, faster, stronger, Proceedings of the IEEE conference on computer vision and pattern recognition. (2017), p.7263 -7271.

[8] REDMOM J, FARHADI A. Yolov3: An incremental improvement. arXiv preprint (2018),p. 1804. 02767

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, Amsterdam, NETHERLANDS, (2016),p.21-37.

[10] ZHANG K, ZHANG Z, LI Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters, (2016),23(10),p.1499-1503.

[11] KU H, DONG W. Face recognition based on mtcnn and convolutional neural network[J]. Frontiers in Signal Processing, (2020)4(1),p. 37-42.

[12] Chen Zhengsheng. :Research on mask wearing detection method based on deep learning: (Huazhong Normal University,2021).