Research on Multi-Angle Face Detection Method Based on Improved YOLOV2 Algorithm

Huanping Zhang¹, Hanhua Cao¹ *, Zhendan Liu¹, Yuhuai Zhou¹ and Yujuan Wang¹

¹ School of Information Science, Xinhua College Of Sun Yat-Sen University, Guangzhou, Guangdong, 510520, China

*Corresponding author’s e-mail: caohanhua@xhsysu.edu.cn

Abstract. Face detection based on the YOLOV2 algorithm can achieve a higher level of accuracy and show a faster detection speed. Therefore, the YOLOV2 algorithm appears more frequently in the field of target real-time detection. The traditional YOLOV2 algorithm uses the Darknet-19 network. Based on this, this paper builds a new network by merging the residual network (ResNet). In addition, in order to further make up for the defects exposed by the YOLOV2 algorithm in face detection, we change the weight of the loss function and implement an image enhancement scheme on the training set. The YOLOV2 algorithm and the improved YOLOV2 algorithm are compared and tested on the Wider Face dataset, and then we use face data sets from different angles to test the performance of the algorithm. According to the analysis of the experimental results, the improved YOLOV2 algorithm can show higher accuracy and robustness in multi-angle face detection.

1. Introduction

Face detection is mainly realized under the background of artificial intelligence research by using facial features of human faces, whose purpose is to use images or videos as the original carrier to obtain the target, and use tags to mark the valuable faces with size and location information [1]. The face recognition system is the initial application of face detection. The emergence of the Internet of Things (IOT) makes electronic identification and authentication more and more in line with the needs of social development. At present, identity authentication is mainly based on the biological characteristics of the face, which is mainly because the recognition of the face is directly friendly and has a strong identity match. At this stage, popular applications for face recognition include access control systems, criminal investigations, medical analysis, fatigue driving warning [2], video conferencing [3], and 3D dynamic capture technology [4]. The application of face detection technology is very extensive, but its accuracy still needs to be further improved. In recent years, the emergence of deep learning has largely solved the bottleneck in the accuracy of face detection, which provides an opportunity for the evolution of computer vision.

At first, face detection is to recognize faces through comparison and judgment. The comparison objects in this process include the detected image and the given face template image [5]. After the rise of deep learning, convolutional neural networks have become the most common application in face detection, which can achieve higher accuracy and faster detection speed. Since then, in order to obtain better detection performance, people have continued to work hard on the research of deep learning algorithms, and successfully designed R-CNN [6], Fast R-CNN [7], Faster R-CNN [8] and other algorithms, which have improved the development of face detection. Zhang et al. [9] propose a new face...
detection model, which mainly uses a multi-task cascaded convolutional neural network. The highlight of the model is to combine the two originally independent tasks of face detection and alignment to further improve the detection effect. Joseph Redmon et al. [10] use the Darknet-19 network [11] to improve the structure of the backbone network based on the YOLO V1 algorithm. They introduce a convolutional layer with anchor boxes to optimize the accuracy of the algorithm in positioning, and name the new algorithm YOLOV2.

Aiming at the shortcomings of the YOLOV2 algorithm in face detection, this paper optimizes and improves its loss function and image enhancement, and proposes a multi-angle face detection network with better detection results.

2. Multi-Angle Face Detection Network

2.1. The basic structure of YOLOV2
The basic flow of YOLOV2 face detection algorithm is shown in Figure 1. In the YOLOV2 algorithm, the resolution of the pre-training image is 448×448. The basic structure of the Darknet-19 network model is composed of 19 convolutional layers, 5 maximum pooling layers and 1 average pooling layer. As the basic structure, the Darknet-19 network can accurately extract the feature information contained in the image, mainly using the global average pooling layer to achieve accurate prediction. In Darknet-19 network, 1×1 convolution is used between 3×3 convolution, which is mainly used to compress the number of feature map channels. The purpose of this structure is to optimize the parameters of the model and reduce the computational pressure of the model. In addition, each convolutional layer in the Darknet-19 network has BN layer after each convolutional layer. This structure can enable the model to reach a higher speed and reduce the occurrence of model over-fitting.

2.2. Multi-angle face detection network
1) Improvement based on residual network structure
Darknet-19 network can show a strong target detection effect, but it does not have an advantage in small-size face target detection. The detection of small-size faces mainly relies on shallow features. However, in the Darknet-19 network structure, 19 convolutional layers and 4 pooling layers first extract features, which prevents the use of high-resolution shallow features, resulting in insufficient training of shallow features. This paper fully considers the advantages of the residual network (ResNet), and based on the basic skeleton of the Darknet-19 network, the residual component is used to form the residual block. The main improvement process of Darknet-19 network based on ResNet structure improvement:

Residual block 1: The 104×104×128 feature map output by the 6th layer convolution is used as the residual amount, and it is fused with the 26×26×256 feature output by the maximum pooling of the 11th layer.

Residual block 2: The 52×52×256 feature map output by the 10th layer convolution is used as the residual amount, and it is fused with the 13×13×512 feature output by the maximum pooling of the 17th layer.

Residual block 3: The 13×13×1024 feature map output by the 18th layer convolution is used as the residual amount, and merge with the 20th layer output in the same dimension.
Residual block 4: same as residual block 3, residual block 4 is composed of 20 layers and 22 layers. The fusion of residual block 1, residual block 2, residual block 3 and residual block 4 can be expressed as:

\[
H_{11}(x) = F_6(x) + F_{11}(x)
\]

(1)

\[
H_{17}(x) = F_{10}(x) + F_{17}(x)
\]

(2)

\[
H_{20}(x) = F_{18}(x) + F_{20}(x)
\]

(3)

\[
H_{22}(x) = F_{20}(x) + F_{22}(x)
\]

(4)

For residual block 1, using 1×1 convolution with a convolution kernel of 1, we convolve the shallow features with a step size of 4, and the obtained 4 feature outputs with 6×26×128 are combined into one feature with 26×26×512, which is the residual amount output from the 6th layer. Finally, the residual amount is fused with the output features of the 11th layer, and an example of the multi-scale fusion process is shown in Figure 2.

In the same way, the output features of the 10th layer are reorganized to obtain a residual quantity with a dimension of 13×13×1024. After fusion, an output feature of 13×13×1536 is obtained. For residual block 3 and residual block 4, there is no need to recombine, they form 13×13×2048 and 13×13×3072 feature outputs respectively after fusion.

In this paper, based on the characteristics of the Darknet-19 network, the networks we use for residual blocks are all over 2 layers. Among them, the residual blocks 1 and 2 can achieve large cross-dimensional fusion, and the residual blocks 3 and 4 can build a residual structure in the deep layer to avoid the occurrence of gradient explosion and gradient disappearance. The Darknet-19 network architecture is shown in Figure 3.

Based on the basic structure of the Darknet-19 network, the improved network is shown in Figure 4. In the improved network, the residual structure adopts route to extract and cache output features. In addition, in order to allow the extracted features to be incorporated into the residual block in a reasonable dimension, the residual structure puts the reorg process of the extracted features in the reorg link.
2) Improvement based on loss function

In the face detection process, the same error has different effects on the size of the face. Therefore, it is necessary to improve the loss function to eliminate the difference in the influence of errors on large and small faces. In order to avoid introducing a new error term and causing the loss function to fail to converge, it is necessary to start with the original variance term for optimization. Therefore, this paper adds the denominator to the variance term to ensure that the influence of the face image errors of different sizes is consistent during size prediction. The resulting loss function expression is:

$$\text{Loss} = \lambda_{\text{coord}} \sum_{k=0}^{S^2} \sum_{n=0}^{B} I_{\text{obj}}^n \left( x_k - x'_k \right)^2 + \left( y_k - y'_k \right)^2 + \lambda_{\text{coord}} \sum_{k=0}^{S^2} \sum_{n=0}^{B} I_{\text{obj}}^n \left( \frac{\sqrt{w_k} - \sqrt{w'_k}}{\sqrt{w_k}} \right)^2 + \left( \frac{\sqrt{h_k} - \sqrt{h'_k}}{\sqrt{h_k}} \right)^2$$

(5)

3) Improvement based on image enhancement

There are different face sizes in the face detection process, which puts forward higher requirements on the detection method. Use 0.5 as the zoom factor to reduce the face images in the dataset to 2 as the zoom factor to enlarge the face image in the dataset. The effect diagram of the data set after zooming is shown in Figure 5. Assuming that there is a point $P(x,y)$ in the training set image, the value of the point P in the channel Q can be expressed as $I(x,y,q)$, and the characteristic value obtained after the horizontal mirroring process should be:

$$I'(x,y,p) = I(w-x-1,y,p)$$

(6)

Among them, w represents the width of the image. The effect of the data set image mirroring is shown in Figure 6.
3. Experiment and Result Analysis

3.1. Network training

1) Anchor box selection

This paper uses statistical ideas to implement clustering operations on the faces in the Wider Face data set, and automatically selects the appropriate size of the anchor box to improve the intersection ratio on the Wider Face data set and make the training more adequate. After clustering, a total of 5 anchor boxes of different sizes are obtained.

2) Learning rate adjustment

Table 1 shows the decline of the loss function under different initial learning rates. According to the experience of the YOLOV2 algorithm on the data set, when the Loss function drops below 0.1, the requirements of this paper can be met. Therefore, this paper selects 0.0005 as the initial learning rate for learning.

| Initial learning rate | 0.00001 | 0.0001 | 0.0005 | 0.001 |
|-----------------------|---------|--------|--------|-------|
| The decline of the Loss | Down to 0.18 | Down to 0.18 | Keep falling | Local optimum |

3.2. Detection performance analysis

In order to verify the practical application effect of the improved algorithm proposed in this paper on face detection, we conducted a comparative experiment on the traditional YOLOV2 algorithm and the improved algorithm. Table 2 lists the AP values and single slice test time of the two comparative models on the Wider Face dataset.

| Models               | AP       | Single slice test time |
|----------------------|----------|------------------------|
| YOLOV2               | 84.7%    | 0.243s                 |
| Improved YOLOV2      | 89.8%    | 0.206s                 |

Compared with YOLOV2, the improved YOLOV2 can increase the detection accuracy by 5.1%, which shows that the improvement of ResNet and loss function enhances the network's effect.

Figure 7. Comparison of the results of face detection

Figure 8. Comparison of multi-angle face detection results
In order to compare with the detection results of YOLOV2 more intuitively, we filter some images in the WIDER FACE dataset with certain standards. We select representative images to verify the actual effect of the face detection algorithm. The comparison result is shown in Figure 7. In addition, in order to verify the effect of the improved algorithm on multi-angle face detection, we selected three face images with different angles as the test objects. The detection results are shown in Figure 8.

It can be seen from Figure 7 and Figure 8 that YOLOV2 and the improved YOLOV2 can basically detect human faces. But when detecting faces from different angles, it is obvious that the improved YOLOV2 shows better detection effect and robustness than the traditional YOLOV2. Therefore, the improved YOLOV2 can significantly improve the performance of the face detection model.

4. Conclusion
Focusing on the contradictions and problems encountered in the development of face detection at this stage, this paper proposes a multi-angle face detection method based on deep learning to meet the needs of the development of society for high-precision face detection. We integrate the concept of residual network into the construction process of the network based on the basic characteristics of YOLOV2, which fundamentally improves the training effect of the network model and the robustness of face detection. In addition, in view of the shortcomings of the traditional YOLOV2 algorithm in small face detection, the loss function is further optimized to achieve multi-angle face detection. Taking the Wider Face dataset as the training and testing carrier, the results of the comparative experiment fully prove that the improved network model has superior performance in face detection.

Acknowledgments
This work was financially supported by reforming on "graphic design" course under model of "Internet + OTO Cloud Classroom" (NO.2019J034) and PHP website development technology (No.2019JC007), which are Teaching quality and teaching reform projects of Xinhua College of Sun Yat-Sen University Guangzhou.

References
[1] Soula, A., Tbarki, K., Ksantini, R., et al. (2020) A Novel Incremental Kernel Nonparametric SVM Model (IKN-SVM) for Data Classification: An Application to Face Detection. J. Engineering Applications of Artificial Intelligence. 89: 1-10.
[2] Xiao, Z., Hu, Z., Geng, L., et al. (2019) Fatigue Driving Recognition Network: Fatigue Driving Recognition Via Convolutional Neural Network and Long Short-term Memory Units. J. Intelligent Transport Systems.
[3] Muttaqin, R., Nopendri, Fuada, S., et al. (2020) Attendance System using Machine Learning-based Face Detection for Meeting Room Application. J. International Journal of Advanced Computer Science and Applications. 11: 286-293.
[4] Jerome, C. L. (2005) Surprise as a Facial Expression on Animation-Examples from Ghibli Studio Animated Films. J. Xuanzang Journal of Information Communication. 2: 51-80.
[5] Maggu, M. S. (2021) Image-Based Face Detection and Recognition. J. International Journal for Modern Trends in Science and Technology. 6: 466-470.
[6] Vukovic T, Petrovic R, Pavlovic M, et al. (2019) Thermal Image Degradation Influence on R-CNN Face Detection Performance. C. 27th Telecommunications Forum (TELFOR).
[7] Wang, K., Dong, Y., Bai, H., et al. (2017) Use Fast R-CNN and Cascade Structure for Face Detection. C. Visual Communications & Image Processing. IEEE.
[8] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. J. IEEE Trans. PAMI. 39: 1137-1149.
[9] Chang, S. Y., Li, Y. C., Sun, P. Y., et al. (2019) A Video Face Detection and Recognition Method based on MTCNN. J. Journal of Xuchang University. 38: 149-152.
[10] Redmon, J., Farhadi, A. (2017) YOLO9000: better, faster, stronger. C. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE. 6517-6525.
[11] Akbarian, S., Delfi, G., Zhu, K., et al. (2019) Automated Non-contact Detection of Head and Body Positions during Sleep. J. IEEE Access. 99: 1-1.