Research on Face Recognition Technology Based on Improved YOLO Deep Convolution Neural Network

Yu Fan, Yiyue Luo*, Xianjun Chen
Haikou University of Economics, Haikou, Hainan 571127, China

*Corresponding author: luoyiyue@126.com

Abstract. With the development of information technology, deep learning has become a research hotspot in the field of computer vision, among which face recognition technology, as one of its applications, has received extensive attention in recent years. A face recognition method based on improved YOLOv3 deep convolution neural network is proposed. The feature pyramid network is used to obtain the four scale features of the target to fuse the shallow features and the deep features, and the influence weight of the loss function is adjusted according to the size of the detected target, so as to enhance the detection effect of small targets and mutually occluded objects, and the Softmax function is used for classification and recognition. Experimental results on mixed data sets show that this method can improve the real-time performance of face detection.

Keywords: Improve YOLOv3; Deep convolution neural network; Face recognition; loss function

1. Introduction

Face recognition belongs to the category of computer vision, which refers to the "intelligent" technology that computers automatically identify people by analyzing and comparing face visual feature information. Because the face contains complex feature information, the early research can only be applied to pictures taken under strict restriction conditions. With the development of feature extraction technology and computer technology, face recognition can be realized under actual unrestricted natural conditions. However, apart from the complexity of face features, the changes of the face itself and its environment, such as expression, occlusion, and the ambient light intensity of the image, can easily affect the recognition performance, which makes the actual recognition result not ideal [1]. Therefore, there are still great difficulties to overcome if face recognition technology is to be widely used in practice.

Face recognition technology uses the core algorithm to calculate and analyze the position, face shape and angle of the acquired face, and then compares it with the existing templates in its own database to identify the authenticity of the user's identity. Face recognition technology mainly includes three steps: face detection, feature extraction and feature matching.

With the continuous enrichment of data resources and the improvement of graphics processor performance, the deep convolution neural network method has been applied in the field of target detection. R-CNN(Regions with CNN features) target detection method is proposed for the first time
in literature [2], which uses deep convolution neural network to extract target features and achieves high target detection accuracy. On this basis, fast r-cnn [3] and faster r-CNN [4] methods have appeared one after another, and the accuracy and speed of target detection have been further improved; However, the above methods still need to generate candidate regions and classify targets in stages when detecting targets, and the detection speed is difficult to meet the real-time requirements. Literature [5] puts forward YOLO(You Only Look Once) target detection method based on the idea of regression prediction, which realizes real-time target detection and greatly improves the detection speed.

In this paper, the improved YOLOv3 convolutional neural network is used to improve the accuracy of face detection by modifying and testing the network structure of YOLOv3.

2. Introduction to YOLOv3

2.1. YOLOv3 thought

YOLO input image is divided into $S \times S$ uniform grids, and each grid consists of $(x, y, w, h)$ and confidence $C(Object)$. In which the coordinate $(x, y)$ represents the position of the center of the detection bounding box relative to the grid, and $(w, h)$ is the width and height of the detection bounding box [6].

![Figure 1 The anchor box correction of YOLOv3](image)

YOLOv3 sets each grid cell to predict 3 anchorboxes, each anchor box to predict 3 bounding boxes, and each bounding box to predict 4 values, which are $(h, w, x, y)$. If the target center in the cell has an offset $(c_x, c_y)$ relative to the upper left corner of the image, as shown in Figure 1 [7], it will be corrected, and the specific formula [8] is:

$$
\begin{align*}
    b_x &= \sigma(t_x) + c_x \\
    b_y &= \sigma(t_y) + c_y \\
    b_w &= p_w e^{t_w} \\
    b_h &= p_h e^{t_h}
\end{align*}
$$

(1)

In which $p_w, p_h$ represents the width and height of the anchor box corresponding to the grid, calculate the position of the boundary box, filter the prediction boxes with low scores by setting appropriate confidence scores, and perform non-maximum suppression processing on the remaining prediction boxes [9] to obtain the final prediction result.

Confidence $Pr(Object)$ indicates whether the object is included, and confidence $C(Object)$ indicates the accuracy of the position when the object is included, which is defined as:
\[ C(Object) = Pr(Object) \times IOU(box, centroid) \] (1)

If the detection grid does not contain the target object, \( C(Object) = 0 \). \( IOU \) (Intersection over Union) is the overlap ratio between predicted value and real value, that is, the ratio of their intersection and union, as shown in the following formula:

\[ IOU(box, centroid) = \frac{\text{area}(box \cap box(centroid))}{\text{area}(box \cup box(centroid))} \] (2)

### 2.2. Network structure
YOLOv3 network structure has no pooling layer. In the process of forward propagation, the pooling layer is replaced by changing the step size of convolution kernel. The feature extraction model uses many \( 3 \times 3 \) and \( 1 \times 1 \) convolution layers, and there are 53 convolution layers in the fully connected layer, so YOLOv3 is usually called Darknet-53. YOLOv3 firstly extracts the features of the input image through the feature extraction network to obtain the feature map, and uses three scales to predict, as shown in Fig. 1, each scale has three limit values, and finally predicts the object by the limit value with the largest intersection ratio (IOU) with the ground truth.

### 3. Face recognition method based on improved YOLOv3 convolutional neural network

#### 3.1. YOLOv3 target detection
YOLOv3 target detection method divides the input image into \( S \times S \) grids, and each grid uses convolutional neural network to predict the positions of \( k \) bounding boxes, which contain the conditional probability \( Pr(Class|Object) \) of the target belonging to \( C \) types and the confidence \( Pr(Class|Object) \) of the target (IOU, which represents the intersection ratio between the predicted box and the real box). The category confidence of each prediction frame is:

\[ Pr(Class|Object) \times Pr(Object) \times IOU_{\text{truth}} = \] (3)

It characterizes the coincidence degree between the real frame of the target and the predicted frame of the target and the probability that the target belongs to each category. YOLOv3 convolutional neural network predicts that the dimension of output layer is \( S \times S \times (K \times (5 + C)) \).

YOLOv3 convolution neural network adopts convolution - downsampling method in the middle layer of feature extraction, the gradient back propagation will become smaller and smaller layer by layer, and finally "gradient disappears" will appear. YOLOv3 focuses on fusing the global features of different layers, and does not make full use of multi-scale local features, which limits the accuracy of YOLOv3 target detection.

#### 3.2. Feature extraction network
Inspired by the jumping connection mechanism and residual block structure of ResNet network structure, this paper improves the original YOLOv3Darknet-53 structure and constructs Darknet-50 structure. The transition layer is \( 1 \times 1 \) and \( 3 \times 3 \) convolution layers which are used alternately. Using \( 1 \times 1 \) convolution layer is helpful to smooth the extracted features and avoid losing more feature information during downsampling.

The left side of Darknet-50 network structure designed in this paper is Darknet-53 structure of original YOLOv3. The improved feature extraction network Darknet-50 pays more attention to the information extraction of shallow features, and removes redundant convolution layer. The additional \( 1 \times 1 \) convolution layer in the transition layer can effectively avoid losing more feature information in the process of downsampling, and further enhance the discrimination ability of target features.

At the same time, YOLOv3 draws on the idea of FPN. As shown in Figure 2, it uses three feature
layers with different scales to predict the target to be detected, and outputs three feature maps with scales of 13 × 13, 26 × 26 and 52 × 52, which are sampled on the last two feature maps output by the network and fused with the feature maps with corresponding scales in the early stage of the network to make prediction.

![Figure 2 Multi-scale fusion of FPN in YOLOv3](image)

Compared with YOLOv2 and YOLO, YOLOv3 improves the detection accuracy on the basis of losing a certain speed, but still meets the real-time requirements. The detection method based on multi-scale fusion has good robustness for the detection of targets with different sizes in the image, but the shallow information is not fully utilized by using only three-scale features, which leads to the low-level features with detailed information not being fully utilized, and the position information of many small targets is lost, which is not conducive to the detection of small targets.

3.3. Design of Improved YOLOv3 Network Structure

With the deepening of network layers, although ResNet module can alleviate the phenomenon of gradient explosion, so that the accuracy will not decrease. However, this is based on the premise that the network structure can be more complex and there is no high requirement for memory occupation. In order to minimize the complexity of the network structure, reduce the memory occupation of the network model and ensure higher accuracy, Dense block module will perform better than ResNet module.

In order to compress the model and improve the reuse rate of feature information, the network structure needs to be adjusted. Considering that the feature graphs at 32×32 and 16×16 scale contain more representation information, but the representation learning ability at 8×8 scale is limited, the DenseNet module at YOLO v3 scale is replaced by the Dense block module adapted to its dimension, and its updated network is shown in Figure 3.

![Figure 3 Improve YOLOv3 network structure](image)

By constructing such a densely connected network structure for scale 2 and scale 3, the characterization information learned from different dimensions can be utilized and summarized to the extreme, which provides an effective guarantee for the next accurate prediction.
3.4. Model training and parameter setting
In model training, the official website pre-training weight file is used to initialize the network weight parameters, which accelerates the convergence of the model. Training parameters are set according to experience, and the values of training parameters are shown in Table 1. In order to reduce the pressure on the graphics card and find the best balance between memory efficiency and memory capacity, 74 samples are used as a batch of samples during training, and the parameters of each batch of samples are updated once, and each batch of samples is divided into six times and sent to the trainer in several times.

| Parameter name                  | Parameter value |
|---------------------------------|-----------------|
| Batch sample size               | 74              |
| Batch splitting                 | 6               |
| Iterate                         | 6000 times      |
| Learning rate attenuation step  | 4500 times -5000 times |
| Learning rate attenuation factor| 0.2,0.2         |
| Momentum                        | 0.8             |
| Weight attenuation              | 0.0006          |
| Non-maximum threshold           | 0.22            |

In training, the network model is optimized by using random gradient descent (SGD) algorithm. The learning rate is adjusted according to the number of iterations, and the initial learning rate is 0.001. When the model is iterated to 8000 and 9000 times, the learning rate decays 10 times. At the beginning of training, burn_in parameter is set to stabilize the model, and burn_in is set to 1000. When the update times are less than 1000, the learning rate strategy changes from small to large, and when the update times are greater than 1000, the set learning rate strategy changes from large to small.

![Figure 4 Model loss curve](image)

The final model was trained 6 000 times, which took 6 hours. A total of 45000 images were used in the training (4 500 images were randomly selected and reused). In the training process, the dynamic training process is observed by drawing the model loss curve, and the corresponding loss value change curve is shown in Figure 4. It can be seen that the loss value of the model decreases rapidly in the previous iteration, and the model fits quickly; After iterative training for 2 000 times, the loss value decreased slowly. When the iteration reaches 6 000 times, the loss value converges to 0. 002 6, and the training ends.

4. Experimental results and analysis

4.1. Experimental environment and data
In this paper, pycharm3.6+anaconda3+tensorflow experimental platform architecture is adopted, and the experimental data is manually collected and labeled data face_data. TensorFlow is an open source software library for numerical calculation using data flow graph [10], which is mainly used for
machine learning and deep neural network research.

4.2. Algorithm realization
The specific implementation steps of the improved YOLO deep convolution neural network face recognition algorithm are as follows.

Step 1: Enter variables. The face image detected by the improved YOLO deep convolution neural network is transformed into single-channel image data through gray processing, and its image dimension is set to 64×64×1, so the dimension of the input layer is N×64×64×1.

Step 2: data partition. Image data is divided into training set and test set. The parameters of YOLO deep convolution neural network are improved by training set, and the accuracy of parameters is verified by test set.

Step 3: Convolution. The first convolution layer convolves the input layer (64×64×1) with 32 3×3×1 convolution cores, then uses ReLU activation function to perform nonlinear mapping to enhance the fitting ability of the model, and then makes maximum pooling for each region. The second convolution layer convolves the feature map generated by the first convolution layer with 64 3×3×32 convolution kernels. The third convolution layer convolves the feature map of the second convolution layer with 64 3×3×64 convolution cores.

The generated convolution result takes the first convolution layer as an example, as shown in fig. 5.

![Figure 5: The first layer convolution result](image)

Output of full connection layer is realized by Softmax loss function [16]. It maps the features generated by convolution process to probability distribution, which is a commonly used multi-classification loss function in neural network learning, and can be defined as:

$$L_s = -\frac{1}{N} \sum_{i=1}^{N} \log P_i$$  \hspace{1cm} (4)

Specifically, it can be expressed as:

$$L_s = -\frac{1}{N} \sum_{i=1}^{N} \ln \frac{e^{y_i^Tw_j+b_j}}{\sum_{j=1}^{C} e^{y_i^Tw_j+b_j}}$$ \hspace{1cm} (5)

$N$ represents the number of training samples, $P_i$ represents the posterior probability that the features of the $i$-th sample output by the fully connected layer are correctly classified, $W_j$ represents the $j$-th column of the fully connected layer weight parameter $W$, $b$ represents the offset term, and $C$ represents the number of categories.

Step 4: Training process of face recognition. Every time, $N$ samples are randomly selected from the training set for training until the accuracy of prediction reaches a certain threshold and converges.

4.3. Optimal model selection
The weights are output once every 100 iterations during training, and a total of 50 weight models are obtained. The best model is screened out by calculating the m AP value of each weight, and the results are shown in Figure 5. It can be seen that when the number of iterations reaches about 2000, the m AP
value tends to be stable, and then the highest m AP value is selected as the best weight model.

![Figure 6](image)

**Figure 6** Variation of mean value of average precision with iteration times

Select the best confidence threshold for the trained model to ensure the best performance of the model. The experiment screened out the best threshold model by calculating the accuracy, recall rate, F1 value and intersection ratio under different confidence thresholds, and the results are shown in Figure 7.

![Figure 7](image)

**Figure 7** Confidence threshold

20 groups of data were calculated by calculating the value once every span of 0.05 in the threshold interval of (0,1). Priority of comparison: accuracy > recall > IOU. After the threshold reaches 0.65, the accuracy gradually tends to be stable, and the optimal range of threshold is about 0.65 ~ 1.0. In this range, the best recall rate is 0.88, the corresponding confidence threshold is 0.66, and the IOU value is about 0.76. Therefore, the best threshold is 0.66.

4.4. Comparison of detection effects of different algorithms

Through the comparative experiment [11], the detection results of the four methods are shown in Table 2. We can intuitively see the advantages of this method in real-time face detection.

| Test method   | Recognition speed /ms | Accuracy/% |
|---------------|-----------------------|------------|
| Faster R-CNN  | 216.35                | 88.35      |
| Improved YOLOv3 | 63.54                | 86.74      |
| Literature [10] | 71.16                | 83.01      |
| YOLOv3        | 85.69                 | 80.22      |

At the same time, the recall rate curves of the four methods are obtained [12], as shown in Figure 8. It can be seen that the improved YOLOv3 has strong adaptability.
Figure 8 Recall rate curve

It can be seen from Figure 8 that Faster R-CNN and the improved YOLOv3 in this article have higher face detection accuracy than the other two methods. Faster R-CNN needs to perform RPN network feature extraction during detection, so the detection speed is faster than this article. The method is slow. Therefore, the improved YOLOv3 in this paper is superior to other detection methods in face detection. It has good real-time performance while maintaining high detection accuracy.

5. Conclusion

In this paper, a face recognition algorithm is constructed based on TensorFlow, which is a deep learning framework. Firstly, YOLOv3 network is improved to increase the model input resolution and improve the model recognition ability. Then DenseNet is embedded in the original YOLOv3 transport layer with lower resolution to enhance the function expansion and promote feature reuse and fusion. At the same time, the convergence of the model is improved by using the ReLU excitation function, and after many experiments, a higher accuracy is finally obtained under different iteration times. In addition, in addition to maintaining the high accuracy of face detection, the detection speed of this method is faster, and the processing time of each frame is 63.54ms, so it has a good balance between speed and accuracy. These characteristics make our method suitable for real-time face detection.

Acknowledgements

This work is supported by Hainan Provincial Natural Science Foundation of China (NO:619QN249); the Scientific Research Projects of Higher Education Institutions in Hainan Province(NO:Hnky2019-74); the Scientific Research Projects of Higher Education Institutions in Hainan Province (NO:Hnky2018ZD-11).

References

[1] Chowdary V S, Teja G, Mounesh D, et al. Sign Board Recognition Based on Convolutional Neural Network Using Yolo-3. Journal of Computational and Theoretical Nanoscience, vol. 17, no. 8, pp. 3478-3483, 2020.
[2] Khan M Z, Harous S, SalEeT-Ul-Hassan S, et al. Deep Unified Model For Face Recognition based on Convolution Neural Network and Edge Computing. IEEE Access, no. 99, pp. 1-1, 2019.
[3] Li Y, Li M, Qi J, et al. Detection of typical obstacles in orchards based on deep convolutional neural network. Computers and Electronics in Agriculture, vol. 181, no. 8, pp. 105932, 2021.
[4] Alotaibi A, Mahmood A. Deep face liveness detection based on nonlinear diffusion using convolution neural network. Signal Image & Video Processing, vol. 11, no. 4, pp. 713-720, 2017.
[5] Wu D, Hu H, Li Y. Roboust face recognition based on significance difference local directional pattern and deep convolution neural network. Guangdianzi Jiguang/Journal of Optoelectronics Laser, vol. 28, no. 8, pp. 902-909, 2017.
[6] Dazhao, FAN, Yang, et al. Satellite Image Matching Method Based on Deep Convolutional Neural...
Network. Journal of Geodesy and Geoinformation Science, 2019, v.2 no. 02, pp. 92-102.
[7] Guangxin, Lou, Hongzhen, et al. Face Image Recognition Based on Convolutional Neural Network. China communications, vol. 17, no. 02, pp. 125-132, 2020.
[8] Lou G, H Shi. Face image recognition based on convolutional neural network. China Communications, vol. 17, no. 2, pp. 117-124, 2020.
[9] Gan J, K Jiang, Tan H, et al. Facial Beauty Prediction Based on Lighted Deep Convolution Neural Network with Feature Extraction Strengthened. Chinese Journal of Electronics, vol. 29, no. 2, pp. 312-321, 2020.
[10] Ye S, Bohush R P, Chen H, et al. Person Tracking and Reidentification for Multicamera Indoor Video Surveillance Systems. Pattern Recognition and Image Analysis, vol. 30, no. 4, pp. 827-837, 2020.
[11] Yu J, Zhang W. Face Mask Wearing Detection Algorithm Based on Improved YOLO-v4. Sensors, vol. 21, no. 9, pp. 3263, 2021.
[12] Huang J, Yuan Z, Zhou X. A Learning Framework for Target Detection and Human Face Recognition in Real Time. International journal of technology and human interaction, vol. 15, no. 3, pp. 63-76, 2019.