Object Detection Algorithm Based on YOLOv3 Model to Detect Occluded Targets

Dawei Zhang
Xiangtan University
School of Computer Science & School of Cyberspace Science, Hunan, China

*Corresponding author e-mail: 1135119686@zhangdawei.xtu.edu.cn

Abstract. Target detection has experienced a rapid development stage from the concept to the practical application, the core technology has been broken through, many difficult problems have been solved, and it has been widely used in various fields. However, the existing target detection models still have some technical problems such as incomplete detection and low accuracy. Therefore, this paper proposes a target detection algorithm based on V3 model to detect occluded targets. This paper makes an in-depth investigation and Research on the existing target detection model, analyzes the shortcomings of the current traditional target detection model through the test results, and the advantages of the network detection model based on YOLOv3 in target detection. Aiming at the main problems existing in the current detection technology, this paper puts forward the optimization and improvement scheme. By reconstructing the whole framework of multi-target tracking algorithm, optimizing the feature fusion algorithm, and improving the NMS algorithm, the scheme greatly improves the accuracy of detecting the occluded objects. The simulation results show that the improved scheme can improve the accuracy by 12.3% compared with the traditional scheme. The analysis shows that the target detection model based on YOLOv3 network can effectively improve the detection accuracy of the model.

Keywords: Target Detection; Neural Networks; Feature Fusion; Check the Accuracy

1. Introduction

Object detection is a basic problem in the field of computer vision. In recent years, great progress has been made in intelligent monitoring, intelligent classification and product detection [1-3]. However, there are still many challenges in the task of target detection under natural conditions, among which the occlusion factor has a great impact on the detection effect. Because occlusion will lead to the loss of local features of the detected target in the image, affecting the integrity of the target features, thus affecting the actual detection accuracy in the detection process [4-5].

In order to achieve high accuracy, the model is widely used in the field of target detection. The convolution neural network is used to extract the features of each part of the image, and the target information is directly regressed to realize the target detection. On the basis of yolo9000 model, the anchor box of fast r-cnn is introduced into YOLOv1 and combined with full convolution network to improve the recognition accuracy and speed of the model [6-8]. Although the above methods improve
the detection accuracy of large and small targets in the image, occlusion still affects the detection accuracy in actual detection. Therefore, in order to better solve the problem of target detection accuracy in occlusion, YOLOv3 is generally considered to be a better solution. In theory, compared with the detection model based on yoo1v, the detection model based on yoo3v network can enhance the ability of target edge detection. The average detection accuracy will be significantly higher than the traditional yoo1v model, which can achieve better detection effect [9-10].

In this paper, the application status of traditional target detection model is investigated. According to the investigation and analysis results, the traditional object detection model based on YOLOv1 network often has some technical problems such as inaccurate detection and low precision when detecting the occluded objects. In view of this problem, an improved optimization design based on YOLOv3 network model is proposed. In this paper, based on the YOLOv3 network model, a series of operations including reconstructing the overall framework of multi-target tracking algorithm, optimizing the feature fusion algorithm, and improving the NMS algorithm are carried out, which greatly improves the accuracy of the detection model in the detection of occluded objects. In order to further verify the actual effect of the optimization model, this paper carries out a simulation test comparing the YOLOv3 network model and the YOLOv1 network model. Through the analysis of the detection results, we can see that the improved model further improves the detection accuracy of the model in the complex background and the target is occluded environment.

2. Target Detection and YOLOv3 Target Detection Algorithm

2.1. Concept of Target Detection

Target detection includes two core subtasks: target localization and target classification. According to the different methods of target location and classification, target detection is divided into two stages: target detection algorithm and target detection algorithm. The difference between them is that the two-stage target detection algorithm carries out the target location and target classification process respectively, while the one-stage target detection algorithm carries out the target location and target classification at the same time. Most of the early target detection methods belong to two-stage target detection algorithm, which can meet the target detection mode of first positioning and then classification. For the general target location problem, in some specific scenes, such as the special moving state of the target or the simple background information of the target, the position of the target can be directly determined, and even the target can be tracked in real time according to the key features. According to different scenes, target localization methods can be divided into statistical feature-based method, contour-based method and visual perception method.

2.2. YOLOv3 Target Detection Algorithm

YOLOv3 target detection algorithm is a real-time target detection algorithm based on regression proposed by Joseph Redmon and Ali Farhadi in 2018. It is a convolutional neural network which can simultaneously predict the position and category of multiple target frames. It is an improvement of YOLOv1 model. YOLOv3 uses the residual neural network as the basic network of feature extraction; on this basis, a convolution layer is added to predict the images of three different scales, so as to obtain higher semantic information. In addition, considering the overlapping of class labels, YOLOv3 uses logistic instead of softmax classifier, and the classification loss is also calculated by binary cross entropy loss. YOLOv3 uses FPN network to detect targets of different sizes on multiple scales. The finer the cell, the finer the target information can be detected. The size of the feature map of each prediction task is as follows:

$$N \times N \times [3 \times (4 + 1 + \text{class_num})]$$

(1)
Where $N$ is the target size, 3 is the number of bounding boxes obtained for each target, 4 is the number of bounding box coordinates, 1 is the predicted value of the target, and $\text{class\_num}$ is the number of categories.

3. Simulation Test of Main Specifications of YOLOv3 and YOLOv1 Models
The main idea of model training is to use model training data sets on a large scale for fine tuning, and continue to cultivate patterns in new datasets. Yolo coco and VOC, the original authors of training based on the darknet53 model, mean that the self-constructed image data set in 1800 was randomly selected as the training set, and the remaining 650 images were used as the test set. The initial learning rate was 0.001 and the attenuation coefficient was 0.0005. Training YOLOv1 network and YOLOv3 network.

It can be seen from Table 1 that YOLOv3 network is better than YOLOv1 network in two main indexes. Due to the large calibration of feature weight by senet, the influence of important features on classification results is enhanced, and the non-important features are suppressed, which further enhances the feature description ability of the network, and finally improves the recall and accuracy of the network. In terms of algorithm running time, in the gtx1080 video card and cuda9.0 running environment, the computing time of YOLOv3 and YOLOv1 is 10.23s and 11.46s respectively, which are all above 60fps, and the additional computing time brought by network addition is less.

| Network   | Precision | Recall   |
|-----------|-----------|----------|
| YOLOv1    | 80.47%    | 79.97%   |
| YOLOv3    | 86.23%    | 85.44%   |

4. Discussion

4.1. Comparative Analysis on the Actual Detection Effect of Two Kinds of Network Models in Fruit Plantation
In this experiment, we collected four pictures of "cherry", "banana", "grape" and "watermelon". Considering that the detection accuracy of the model will be affected by the integrity of the training data set, we adjust the color, brightness and angle of the collected images, and expand the dataset to improve the integrity of the dataset. Finally, the total number of images was expanded to 800, including 680 and 120 for training and testing. In the experiment, the threshold value of $IOU$ is set to 0.5, that is, if the cross ratio between the predicted frame and the actual frame is greater than or equal to 0.5, the detection result will be taken as an example; if the intersection and combination ratio is less than 0.5, the detection result is considered as a false positive example.

In the experimental results in Figure 1, we can see that the detection accuracy of YOLOv3 model is higher than that of four types of data sets of YOLOv1 model, and the maximum accuracy of cherry data set obtained is improved by 12.3%, indicating that the model is the most sensitive to the characteristics of cherry fruit and is most suitable for cherry target detection. In general, the detection accuracy of YOLOv3 model has been significantly improved for the orchard scene with complex environment.
It can be seen from the experimental results in Figure 2 that the YOLOv3 model improves the detection accuracy of cherry and other fruits in the mutual occlusion scene. The results show that using mutual occlusion markers to describe the occlusion scene can significantly reduce the interference caused by mutual occlusion between fruits, improve the sensitivity of the model to fruit targets, and improve the detection accuracy. In general, the detection accuracy of YOLOv3 model is significantly higher than that of YOLOv1 model in different occlusion scenes, which indicates that the occlusion marking and occlusion compensation mechanism of YOLOv3 model have significant effect on improving the detection accuracy in natural environment.
Figure 2. Comparative analysis of detection effect of two network models in mutual occlusion scene

4.2. Overall Framework of Multi-Target Tracking Algorithm
The whole framework of multi-target tracking algorithm is mainly divided into two modules: detector and tracker. The detector is mainly combined with YOLOv3 for target detection and recognition. Firstly, the video sequence image is preprocessed; secondly, the convolution feature map is obtained by the total convolution neural network, and the input feature map is analyzed and filtered by the detection network; finally, the frame of the optimal target is obtained through the confidence calculation and multi-scale prediction, and the target is classified by the classifier to obtain the coordinates of the center point of the optimal target frame.

The tracker associates and tracks the output data of the detector, and inputs the optimal coordinates of the target center point of this kind of target into Kalman filter to predict the center point at the next moment, that is, trajectory prediction. The frame of the detected object is associated with the inter frame data to determine the number of targets, so as to obtain the center point measurement value of the optimal estimated value target and the estimated value point of the center in the real state. If occlusion results in data association failure, the nearest neighbor optimal tracking algorithm is used to associate the new target with the vanishing target at that time. When the trajectory fluctuates abnormally due to partial occlusion, the trajectory anomaly correction algorithm is used to correct the target frame and trajectory.

4.3. Feature Fusion
In video sequence images, the front and back frames have a great influence on the current video images. In order to make full use of the target information contained in the video sequence, the front and back feature maps of the video image extracted by darknet-53 are effectively fused to improve the accuracy of target detection. When the front frame and the front frame are fused, the original feature map is updated to the fused feature map by using the linear iteration method. Assuming that the
current feature map is $F^*$, the previous frame of the current frame is $F_{n+1}^*$, and the next frame of the current frame is $F_{n+1}^*$, then the fused feature map $F^*$:

$$F^* = \omega * F_{n+1} + F^* + \omega * F_{n-1}^*$$  \hspace{1cm} (2)

Among them,

$$F_{n-1}^* = \omega * F_{n} + F_{n-1}^* + \omega * F_{n-2}^*$$  \hspace{1cm} (3)

And $\omega$ is the correlation factor of adjacent video frames, $\omega \in [0,1]$.

4.4. Improvement of NMS Algorithm

NMS algorithm is a common post-processing operation, which is used to delete high cross-high ratio of repeated prediction frames, and output of reserved prediction frames and detection results with the highest score, so as to reduce false positive results and find the best target detection location. The traditional NMS has a great disadvantage. When the cross ratio between the prediction frame and the prediction frame with the highest score is greater than 17, the frame will be deleted directly, which is not conducive to the detection of adjacent objects. The size difference of partially occluded objects is large, and the arrangement is compact and complex. The traditional NMS algorithm is easy to lead to some targets missing detection. Therefore, in order to improve the detection rate of dense occlusion targets, linear weighted NMS algorithm is considered. When the cross ratio between the frame to be processed and the frame with the highest score is greater than the set value of 17, the score of the frame to be processed is linearly weighted. The larger the cross-combination ratio, the more serious the score decrease. By reasonably adjusting the value of 17, we can not only remove the repeated frames, but also detect the adjacent occluded objects better.

The improved NMS formula is as follows:

$$s_i = \begin{cases} s_i & J < N_i; \\ s_i(1 - I) & J \geq N_i. \end{cases}$$  \hspace{1cm} (4)

Where: $s_i$ is the score of prediction box; $N_i$ is the threshold value; $I$ is the overlap ratio between the current detection box and the prediction box with the highest score.

5. Conclusions

In the research of object detection algorithm based on V3 model to detect occluded targets, this paper makes an in-depth investigation on the current mainstream target detection technology. Through the simulation test of the main specifications of the detection model based on the YOLOv1 network and the main specifications of the YOLOv3 network detection model, it can be seen that the traditional YOLOv1 network detection model has obvious technical defects of low accuracy and low recall rate. Therefore, in order to further improve the accuracy of target detection model in the detection of occluded objects in complex background, this paper uses YOLOv3 network as the main body to construct a new target detection technology model. On this basis, a series of operations including reconstructing the overall framework of multi-target tracking algorithm, optimizing feature fusion algorithm, and improving NMS algorithm are carried out, which not only improves the comprehensive performance of detection model, but also greatly improves the accuracy of detection of occluded objects. Through the simulation test results, this paper analyzes that YOLOv3 network technology itself is an upgrade of YOLOv1 network technology, and further optimizes the performance of feature extraction and Analysis on the basis of YOLOv3 network, which can better meet the requirements of object detection in complex environment when objects are blocked.

References
[1] Cheng Junchang, Liu Jiyong, & Liu Xing. (2017). Research on target detection technology of infrared and microwave composite detection based on related information. Foreign Electronic Measurement Technology, 036 (010), 6-9.

[2] Song Dalin, Chang Jun, Zhao Yifei, & Zhang Zexia. (2018). Anti-detection technology of cat eye target based on decentered field lens. Chinese Physics B, 27 (9), 94220-094220.

[3] Liu Yang, Xia Runqiu, & LV Yong. (2019). Effectiveness evaluation system for anti-cat’s eye detection technology using optoelectronic target. Optics and precision engineering, 027 (005), 1033-1038

[4] LIU Lutao, WANG Xiao, & LI Xinyu. (2019). An aerial target detection algorithm based on visual significance. Applied Science and Technology, 046 (001), 88-93.

[5] Ma Tao, & Geng Min. (2017). Simulation of infrared target detection system. Optoelectronic Technology Application, 032 (002), 62-67.

[6] Chun-Ming Tsai, & Chiu-Lan Lee. (2019). The application of deep learning yolov3 to improve the accuracy of detecting and identifying players in badminton competitions. New Knowledge of Chinese Education, 66 (1), 51 -64.

[7] Lu Shilei, Lu Sihua, Li Zhen, Hong Tiansheng, Xue Yueju, & Wu Benlei. (2019). Citrus recognition method based on improved yolov3-lite lightweight neural network. Transactions of the Chinese Society of Agricultural Engineering, 35 (17), 205-214.

[8] Pich, D., & Nakahira, K. (2020). A comparative analysis of the dataset for training underwater fish detector based on yolov3. IEEJ Transactions on Electronics Information and Systems, 140(9), 1091-1095.

[9] Zhi wins Liu. (2020). Application of cascade r-cnn and yolov3 in missile target recognition. Journal of Image and Signal Processing, 09(2), 102-110.

[10] Wang, Y., Jia, K., & Liu, P. (2020). Impolite pedestrian detection by using enhanced yolov3-tiny. Journal on Artificial Intelligence, 2(3), 113-124.