Accurate Target Positioning and Detection Algorithm Integrating Deep Learning and Prior Knowledge

Ye Liu[1], **Haisheng Song**[1,2,*]

[1]Northwest Normal University, 967 Anning East Road, Anning District, Lanzhou City, Gansu Province, 730070, China

[2]653526491@qq.com
*corresponding author

Abstract. Target targeting and detection is one of the core tasks in the field of computer vision, which has great significance for autonomous driving, intelligent surveillance, medical image analysis and other fields. In this paper, we propose an accurate target positioning and detection algorithm integrating deep learning and prior knowledge, aiming to combine the high efficiency of Faster R-CNN algorithm and the accuracy of prior knowledge to improve the accuracy and efficiency of target detection. Experimental results show that the proposed method achieves excellent performance on several datasets.

Keywords: Deep learning, prior knowledge, fusion, target detection, localization.

1. Introduction

Object detection and localization is an important topic in the field of computer vision, with the goal of identifying target objects in images or videos and determining their positions. With the rapid development of deep learning technology, especially the widespread application of convolutional neural networks (CNN), the performance of object detection and localization has been significantly improved. However, complex backgrounds and variable targets still challenge the accuracy and robustness of object detection and localization. The fast R-CNN algorithm is one of the classic algorithms in the field of object detection. Efficient object detection and localization have been achieved through the combination of Regional Recommendation Network (RPN) and Convolutional Neural Network (CNN). However, the Faster R-CNN algorithm still faces difficulties in detecting small, occluded, and complex background targets. To address these issues, this paper proposes an accurate target localization and detection algorithm that combines deep learning and prior knowledge, aiming to improve the accuracy and efficiency of the Faster R-CNN algorithm.

2. FasterR-CNN algorithm

In 2014, the R-CNN algorithm proposed by Girshick et al. first applied the convolutional neural network in the field of target recognition, and its effect is far beyond the traditional algorithm[1]. R-CNN uses Selective Search or Edge Boxes to generate candidate regions, uses CNN to extract features, and then uses SVM for classification. Commonused classification models include AlexNet, VGG, Google Net, ResNet, etc. The detection accuracy of R-CNN on PASCALVOC 2007 is nearly twice that of DPM, opening the first example of deep learning target recognition. However, it needs to extract all the characteristics of candidate regions and has large computational amount. The subsequent proposed addition of Fast R-CNN to the pyramid pooling layer reduces the computational amount, but it still takes time to find candidate boxes. Faster R-CNN improves on the RPN network acquisition of candidate regions and uses the Softmax multi-task classifier to achieve excellent results in the field of target detection.

The specific flow chart is shown in Figure 1.

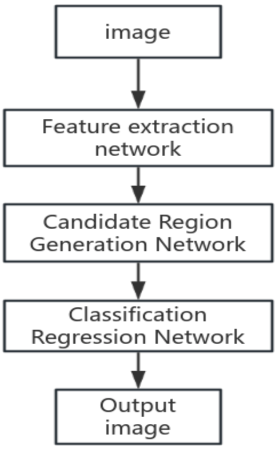


image
Feature extraction network
Candidate Region Generation Network
Classification Regression Network
Output image

Figure 1. Faster R-CNN flow chart

The main steps in the figure are explained as follows:

1) Feature extraction network

Feature extraction network is a convolutional neural network that can be replaced according to the actual needs. The amount of training data affects the final performance and effect. The most commonly used training networks include ImageNet, ZF-Net and so on. The ReLU number is the more common number of activations, defined as follows:

$$f\left(x\right)=\max\left(0,x\right)=\begin{cases}0, & x \quad 0\\ x, & x>0\end{cases}$$

The ReLU function, because the gradient value is constant to 1, avoids the disappearance of the gradient and increases the convergence rate.

2) Candidate region generation network

The detected images are coarse to output the rectangular candidate regions with multiple scales and aspect ratios. Four correction parameters are output for each reference rectangle candidate region boxtx,ty,tw,thAfter the correction, the final candidate area box can be obtained, and the reference rectangular box formula is given as follows:

$$x=w_\alpha t_x+x_\alpha$$
$$y=h_\alpha t_y+y_\alpha$$
$$w=w_\alpha \exp(t_w)$$
$$h=h_\alpha \exp(t_h)$$

3) Categorical regression network

The feature map of the network output and the candidate region of the network output are the sender, and the output candidate region corresponds to the confidence and correction parameters of each category. However, the algorithm still lacks in real-time performance,

3. Target identification and localization based on prior knowledge

3.1. Positioning experiment

The coordinates of the target in the left camera are converted into the coordinates in the world coordinate system for the positioning experiment. Select the position of the camera and the table, and do the positioning experiment at the 4 positions as shown in Figure 4.18 (the origin of the coordinate system is set at a certain point on the inland surface of the laboratory). Table 1 presents the position and pose of the tables calculated by the method of this paper.

Table 1. Table of localization experiments

pattern	Measured Coordinates (x, y, z) (m)	Solution coordinates (x, y, z) (m)	Coordinate error (△ x, △ y, △ z) (m)	attitude error (°)
Pattern 1	(2.51,0.90,0.98)	(2.53,0.91,0.95)	(0.02,0.01,0.03)	2.6
Pattern 2	(2.76,0.90,1.90)	(2.78,0.90,1.89)	(0.01,0.0,0.01)	2.4
Pattern 3	(2.31,0.90,1.28)	(2.31,0.91,1.26)	(0.0,0.01,0.02)	1.2
Pattern 4	(2.29,0.90,1.08)	(2.27,0.90,1.07)	(0.02,0.0,0.01)	2.0

3.2. Optimization of target recognition based on prior knowledge

3.2.1. Low-threshold detection strategy

In the Faster R-CNN framework, the probability threshold of target detection plays a crucial role. This threshold is calculated by the network output layer by softmax function and represents the confidence of the existence of the target. Raising this threshold means that only high-confidence targets are detected, while lowering the threshold eases the detection criteria, allowing more possible targets to be considered[2-3]. For example, when the threshold was set to 0.8, the white water cup was not identified because the color was similar to the background and the characteristics were not significant. To improve the recall rate of detection, we adopted a strategy of lowering the threshold and relaxing the detection conditions to capture more potential target regions. Considering the non-linear distribution of probability values of softmax output, the threshold was adjusted directly using the linear output before softmax, so that the target score extends from 0 to 1 to a broader linear value domain, thus successfully identifying the white cup with a lower score.

3.2.2. Non-maximum suppression technique

After reducing the detection threshold, although more candidate detection boxes appeared in the image, a large amount of redundancy was also introduced, where many highly overlapping detection boxes actually pointed to the same target. Non maximum suppression (NMS) technology is used to simplify detection results and improve computational efficiency [4]. The basic idea of non maximum suppression is to select the local optimal value (i.e. maximum value) among many candidates and eliminate the remaining non optimal terms. In object detection scenarios, this means selecting the detection box with the highest confidence from many detection boxes as the final detection box representing the target, while deleting other highly overlapping low confidence detection boxes. NMS is widely used in various engineering practices due to its high efficiency, and its execution steps are as follows:

(1) All detection boxes are ranked in descending order of confidence score.

(2) Choose whether the detection box with the highest confidence is the current best (maximum).

(3) Check the remaining detection boxes one by one to calculate their overlap rate (IoU) with the current best detection box. If IoU is less than the preset threshold, it is considered an independent target and retained; If IoU is greater than or equal to the threshold, it is considered redundant and deleted.

3.2.3. Target screening based on 3 D spatial constraints

In order to determine the position of the large target, the target between two pasted patterns is selected instead of directly using the disparity method of the large target at the center of the box. This is because when the relative angle between the large target and the camera is large, the distance and coordinates of the midpoint X2 and X1 can only represent the position of the point in the target, as shown in Figure 2. The position information of point A is obtained based on the center point.

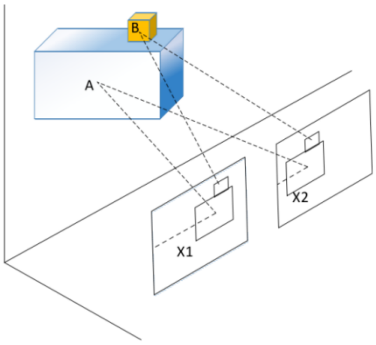


Figure 2. Find the target center location

However, when the volume of the target to be detected is small, the change in object posture has little effect on the position, and the positioning error is only on the centimeter level, as shown in Figure 2. In object B, the position of the center point of the detection box can be approximated as the position of object B.

4. Comprehensive experiment

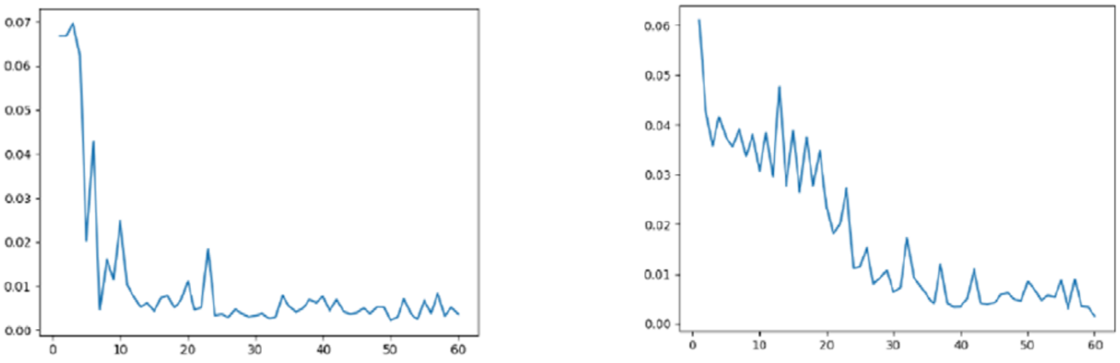
4.1. For experiments comparing ResNet-S and ResNet-50

Network performance experiments by using the laboratory as a scenario. A total of 10 kinds of objects, including water cup, mobile phone, bowl, monitor, mouse, keyboard, trash can, box, backpack, were selected for the experiment.

Table 2. Table of experiments comparing ResNet-50 and RseNet-S

	ResNet-50	RseNet-S
cup	96.53	97.67
cellphone	92.50	87.50
bowl	98.75	98.50
indicator	97.75	98.75
bottle	96.67	98.58
knapsack	98.75	98.75
trash can	97.91	98.20
mouse	94.32	95.57
fingerboard	95.73	97.44
box	98.75	97.46
Average accuracy	96.77	96.74

As can be seen from the data in Table 2, the average recognition accuracy of RseNet-S is basically the same as that of ResNet-50, and the average performance of ResNet-50 is slightly higher than that of RseNet-S, and the stability is stronger, which proves that ResNet-50 is more versatile. However, when classifying targets with more similar features, the recognition accuracy of RseNet-S is slightly higher (such as distinguishing different types of bottles, water cups, etc.), which is mainly due to the network performance gain of the gradient features of HOG. Using RseNet-S in the accurate classification network of this paper has a better effect.



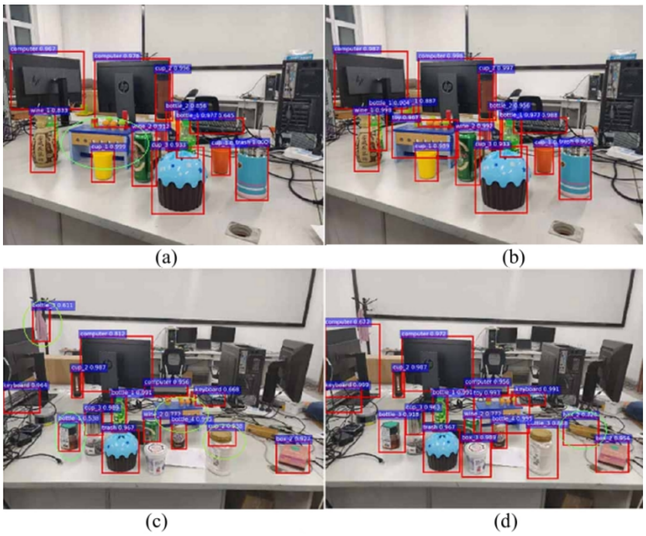
(A) Change in ResNet-50 network loss

(B) Change in the ResNet-S network loss

Figure 3. Network loss drop plot

4.2. Accurate classification network experiment

Firstly, the exact classification experiment was directly conducted with Faster-Rcnn, and then the low threshold detection was combined with ResNet-S binary classifier experiment[5]. The test results are shown in Figure 4 (d), in which the left figure is the original method and the right figure is the improvement method. The probability values in the improved method are the probability output of the dichotomous network, keeping only detection boxes with probability values above 0.5.



(a)

(b)

(c)

(d)

Figure 4. Experimental diagram of the network for accurate classification

Figure 4 shows that the omission rate is lower than that in the left figure, and the average probability score is higher, but occasional errors, such as the binocular camera is misdetected as a box, but less frequently. Misdetection between the same categories rarely occurs in the method of this paper. In terms of detection time, due to the use of multiple binary classification networks, the detection time was improved compared with Faster-Rcnn, and the detection time consumption of detecting 20 categories increased by about 20%. Using 0.5 as the probability threshold, the accuracy and recall of the two methods on 50 pictures are only considered as correct when the detection results accurately match the target category, and the results are shown in Figure 3. Accurate classification methods combined with prior knowledge performed better in the laboratory environment.

Table 3. Comparison of conventional Faster-Rcnn and methods in this paper

	Conventional Faster-Rcnn	The method of this paper
precision	85.25	92.83
recall	73.63	89.47

5. Conclusion

By utilizing the principle of binocular vision object localization and the Faster Rcnn object detection algorithm and binocular localization, the position and posture of the target in the scene are solved. Based on the target position and prior knowledge, the reasonable space where the target may appear is defined, the score threshold of the target is reduced, and more suspicious detection boxes are found in the image. Finally, on the basis of spatial constraints, reasonable detection boxes outside the space are eliminated.

References

[1] Cheknane M , Bendouma T , Boudouh S S .Advancing fire detection: two-stage deep learning with hybrid feature extraction using faster R-CNN approach[J].Signal, Image and Video Processing, 2024, 18(6-7):5503-5510.DOI:10.1007/s11760-024-03250-w.

[2] Zhu H .Research on defect detection of improved target detection algorithm on the image surface of 5G communication ring[J].International journal of modeling, simulation and scientific computing, 2023.

[3] Das N , Swarna R N , Hossain M S .Deep learning-based circular disk type radar target detection in complex environment[J].Physical Communication, 2023, 58(Jun.):102014.1-102014.12. DOI:10.1016/j.phycom.2023.102014.

[4] Fan Y , Tian S , Sheng Q , et al.A coarse-to-fine vehicle detection in large SAR scenes based on GL-CFAR and PRID R-CNN[J].International journal of remote sensing, 2023.DOI:10.1080/01431161.2023.2203341.

[5] Du L , Sun Y , Chen S , et al.A Novel Object Detection Model Based on Faster R-CNN for Spodoptera frugiperda According to Feeding Trace of Corn Leaves[J].Agriculture, 2022, 12. DOI:10.3390/agriculture12020248.