Proceedings of the 38th Chinese Control Conference
July 27-30, 2019, Guangzhou, China

Target Detection Based on Improved Mask Rcnn in Service Robot

Jie Shi, Yali Zhou,WeiJie Xia Qizhi Zhang

School of Automation, Beijing Information Science & Technology University, Beijing 100192, China
E-mail: stonejack@foxmail.com

Abstract: Target detection is one of the core algorithms in robot applications, and the recognition speed has a significant impact on robot's target capture. In this paper, the scene of long-distance and small targets is used as the test scene, and the purpose is to enhance the speed of detection without reducing the accuracy of detection. Consider that the mask branch and excessive full connection layer in the Mask Rcnn network will take up a lot of network detection time, and the feature map extracted by the convolutional neural network has a high dimension, which will occupy a large amount of computational memory. So, in this paper the Mask Rcnn network is improved: remove the mask branch; introduce Light-Head Rcnn into the Mask Rcnn network, increase R-CNN subnet and RoI warping; adjust the proportion of the Anchor in the RPN network. Finally, the improved model is applied to the tensorflow framework. These methods can save computer memory space and improve detection's speed. In the end, the improved Mask Rcnn network has been verified in a service robot platform with Kinect II. The test results show that compared with Faster Rcnn , the improved Mask Rcnn can have a high accuracy of detection;Compared with the original Mask Rcnn, the improved Mask Rcnn network can greatly improve the speed of the algorithm while ensuring the detection accuracy. The detection time is reduced by more than 2 times, which helps to improve the efficiency of the service robot's target capture task.

Key Words: Mask Rcnn; Rcnn subnet; RoI warping; Remove the mask branch ;

1 Introduction

Among many technologies of robots, image-based target detection and localization are the basis of the robot to see the world, and its related algorithms have always been a research hotspot in the field of computer vision. The target detection model of deep learning has undergone many years of development. Currently, more mainstream frameworks are used: AlexNet [1], GoogleNet [2], VGGNet [3], R-CNN[4],Fast-Rcnn[5],Faster-Rcnn[6],SSD[7],YOLO-v3[8], etc. These frameworks extract features of the target through the Convolution Neural Network (CNN) method. CNN has three characteristics: local perception, weight sharing, and multi-convolution kernel.

At the same time, in the process of CNN training, Back Propagation (BP) is used to adjust the weight between neurons. The learning rule is to use the gradient descent method to minimize the sum of squared errors between the predicted and expected values calculated by the network by adjusting the weights and thresholds.

However, gradient disappearance and gradient explosion problems often occur during the deep network model training. The fundamental reason lies in the fact that after the derivation formula of the chain rule of BP algorithm is derived several times, the updating of the weight will result in the increase of exponential form or the decay of exponential form. In view of the problem of gradient disappearance or gradient explosion, the industry has proposed various methods to suppress gradient explosion and disappear, such as the residual neural network ResNet [9].

In fact, it was the appearance of the residuals network that led to the end of the ImageNet race. Since the residual network was proposed, almost all depth networks are inseparable from the residual structure. The Mask Rcnn [10] network used in this paper also use this structure.

As a representative of the image segmentation direction, Mask Rcnn network follows the network structure of Faster Rcnn algorithm, and uses residual structure and Feature Pyramid Networks (FPN) in CNN [11].. The FPN method generates six feature maps, and several of them are sent to the Region Proposal Network (RPN). The RPN network generates k "Anchors", and the RPN network will output 2k at the end. Target score and 4k coordinates are finally mapped to the feature map generated by the CNN. Then, the branching and the pixel-level segmentation of the Mask branch are respectively obtained by the classification prediction branch, and the mask information of the category, location, and article of the object is obtained.

Aiming at the high complexity of the application scenario of the home service robot, the relatively small target object in the whole scene, and the diversity of the target, in this paper, an improved Mask Rcnn target detection system suitable for the service robot is proposed. The specific improvement method of the system is as follows: Firstly, the method of separating the large convolution in Light-Head Rcnn and the method of removing the fully connected layer are applied to the Mask Rcnn network to improve the detection speed of the model. Separating large convolutions uses a full convolutional network, using a larger convolution kernel to extract feature features while reducing feature map dimensions. Second, deleting the mask pixel classification network, decreasing CPU computation and reducing network complexity; The collected data adjust the Anchor ratio in the RPN network to reduce the amount of extra calculations in the network and shorten the detection time. In the end, this paper combines the improved Mask Rcnn network with the tensorflow framework and applies it to the existing home service robot platform.

*This work is supported by National Natural Science Foundation (NNSF) of China under Grant 11672044 and 11172047.

2 Related works

2.1 Mask Rcnn

Mask Rcnn is an improved network based on Faster Rcnn. Faster Rcnn mainly inputs the last layer of the feature map extracted by CNN into the RPN network. Although the target feature is more prominent in this feature map, when the target is too small in the original input image. The convolution layer is likely to filter out small targets as noise in the middle of the hidden layer, even if some simple features can be extracted when they are input into the RPN network. Some targets in Anchor are likely to be regarded as background in the later generation of 256-dimensional vectors. Therefore, in order to solve the problem of small targets detection. Mask Rcnn applied the Feature Pyramid Network (FPN). The idea of FPN is to apply the same feature map extracted by CNN to the RPN network, which avoids the filtering of small target features. By increasing the residual structure,the feature map of "top-down" is fused with the original convolution to enhance the feature of the target.

2.2 RPN Network and RoI Align

Mask Rcnn network uses the same RPN network as Faster Rcnn. Anchor utilizes three scales by default. These scales can contain almost all targets in the training process.But training model in some specific scenes will cause unnecessary calculations.

Faster Rcnn network uses ROI pooling method. In the pocess of discretization, the continuous coordinates are rounded up by the position on the corresponding feature layer, which causes the extracted feature image to be misalignment with the original image.The accuracy is affected in the detection process. Instead , ROI Align can locate targets more precisely in space, that is, the quantization process in the ROI pooling is removed, and the decimal position is preserved.

The most important method in ROI Align is the use of bilinear interpolation [12]. ROI Align bilinear interpolation is shown in Fig 1.

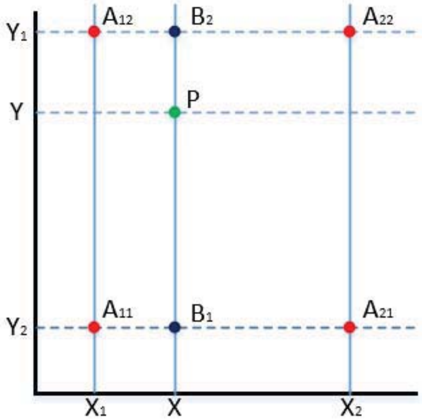

Fig.1 Bilinear interpolation in ROI Align

P is the final interpolation point. A_{11}, A_{12}, A_{21}, A_{22} are known. Linear interpolation is done by A_{11} and A_{21} to obtain B_1. Similarly, B_2 is obtained from A_{12} and A_{22}, as showed in formula (1). . The linear interpolation is done by B_1 and B_2 to obtain the P, as showed in the formula (2).

$$\begin{aligned} f(B_1) &\approx \frac{x_2-x}{x_2-x_1} * f(A_{11}) + \frac{x-x_1}{x_2-x_1} * f(A_{21}) \\ f(B_2) &\approx \frac{x_2-x}{x_2-x_1} * f(A_{12}) + \frac{x-x_1}{x_2-x_1} * f(A_{22}) \end{aligned} \tag{1}$$

$$f(P) \approx \frac{y_2-y}{y_2-y_1} * f(B_1) + \frac{y-y_1}{y_2-y_1} * f(B_2) \tag{2}$$

The bilinear interpolation is (3).

$$\begin{aligned} f(x,y) \approx & \frac{f(A_{11})}{(x_2-x_1)(y_2-y_1)} * (x_2-x)(y_2-y) \\ & + \frac{f(A_{21})}{(x_2-x_1)(y_2-y_1)} * (x-x_1)(y_2-y) \\ & + \frac{f(A_{12})}{(x_2-x_1)(y_2-y_1)} * (x_2-x)(y-y_1) \\ & + \frac{f(A_{22})}{(x_2-x_1)(y_2-y_1)} * (x-x_1)(y-y_1) \end{aligned} \tag{3}$$

2.3 Light-Head Rcnn

Conventional two-stage object detectors usually engaged a heavy head, which has negative influence on the computational speed. "Head" in this paper refers to the structure attached to our backbone base Network. More specifically, there will be two components: R-CNN subnet and ROI warping.

2.4 R-CNN Subnet

The two large fully connected layers used by Faster Rcnn, or all the convolution layers in the fifth stage of Resnet, are used as classifier for the second-stage, which has a good effect on the detection effect. Therefore, Faster Rcnn has high detection accuracy in most datasets, such as COCO dataset. However, when the region proposals box is large, the calculation is also very large. In order to accelerate the Rcnn sub-network, R-FCN first proposes a series of score maps for each region, and the number of these score map channels will become: C × P × P. Where C is the categories.Where P is the size of the next pooling layer. However, as the number of categories increases, the dimension of the score map generated by the R-FCN is still very large, which consumes a lot of time and memory in the whole network.

In terms of detection accuracy, although Faster Rcnn is very good at classifying ROI, it often includes a global average pooling layer at the output of the backbone network. This pooling layer is used to reduce the computational complexity of the first fully connected network.

For the Rcnn subnet, Light-Head Rcnn proposes to use only one simple fully connected layer for the above problem, which makes the network have a good balance between detection accuracy and calculation speed. Fig.2 shows the Faster Rcnn original network.

2.5 RoI warping

The calculation amount and memory consumption of the fully connected layer is also dependent on the number of channels obtained by the ROI processing feature map. Therefore, the ROI warping has to be modified. ROI warping scales the feature map to a specific size before entering the region proposals box into the Rcnn subnet.

Before feeding proposals into the ROI warping, Light-Head Rcnn proposes a fully convolutional network structure that produces small channel feature maps. This ROI warping is based on a thin feature map not only improves accuracy during the training phase, but also saves memory and computation. At the same time, if ROI pooling is implemented to a thin feature map, this can eliminate the global average pooling to improve network detection performance.

That improved architecture of Light-Head Rcnn to Faster Rcnn is Fig.3.

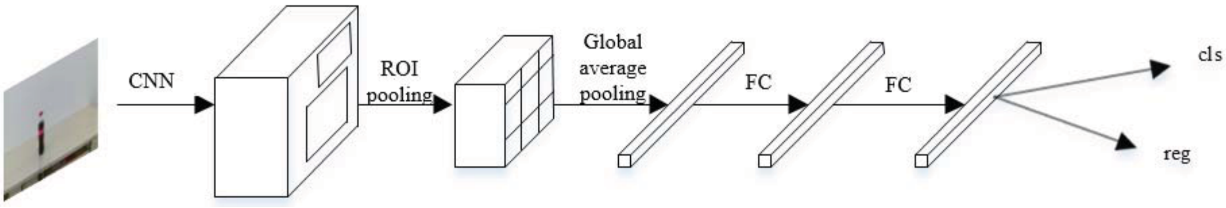

Fig.2 Faster Rcnn original network

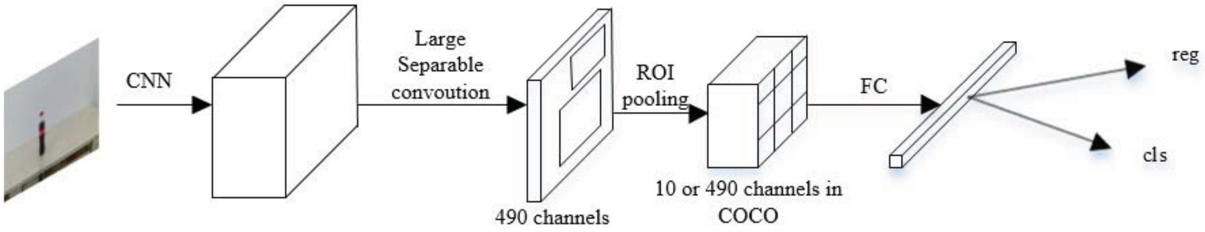

Fig.3 Improved architecture of Light-Head Rcnn to Faster Rcnn

3 Improved Mask Rcnn

3.1 Put light-head rcnn into Mask Rcnn

In this paper, two parallel large convolutions are inserted into the network of Mask Rcnn's basic extraction features. In the feature extractor of Mask Rcnn, Resnet-FPN structure is adopted, in which Resnet adopts 50-layer structure. There are 4 full convolution residual networks in Resnet,and each block contains a different number of residual structures. The first block has 3 residual structures, the second block has 4 residual structures, the third block has 6 residuals,and the fourth block has three residual structures. In the Resnet version of Faster Rcnn, the input of the RPN network is the output of the third residual structure, which is passed into the fourth residual structure through ROI Pooling, and the accuracy is exchanged by reducing the detection speed.

First, the separated large convolution is connected to the convolutional neural network, where the separated large convolution structure is shown in Fig.4. Establishing the value of k to 7, in the whole network structure, all convolution kernel sizes are 1×7 and 7×1, so a thin feature map can be obtained. In this way, the operation efficiency and detection accuracy of the network can be improved, and the memory can be saved. The value of Cmid is set to different value according to different network. For the Xception[13] network, the value of Cmid is 64; for the Resnet series network, the value of Cmid is 256. The value of Cout is set to 10×P×P, P is the pooled size in the ROI pooling layer. In the Faster Rcnn, the size of the pooling layer is 7, so the value of P is 7.In Mask Rcnn,the pool size is also 7. This method is different from the C × P × P in the R-FCN, which can reduce the amount of calculation. Benefiting from the large receptive field brought by the large convolution kernel, the feature maps extracted by the convolutional neural network passes through the structure, and the characteristics obtained after the pooling are more obvious. The more obvious the feature is, the better the detection speed of the network is improved,and the accuracy of targeting.

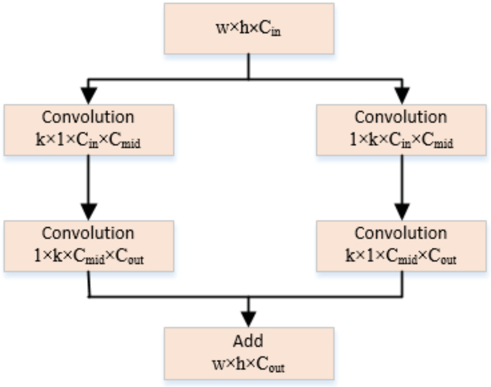
$w\times h\times C_{in}$
Convolution $k\times 1\times C_{in}\times C_{mid}$
Convolution $1\times k\times C_{in}\times C_{mid}$
Convolution $1\times k\times C_{mid}\times C_{out}$
Convolution $k\times 1\times C_{mid}\times C_{out}$
Add $w\times h\times C_{out}$

Fig.4 Large separable convolution architecture

Second, the Rcnn subnet will replace the fully connected layer at the output of the Mask Rcnn network. The precise method is to remove all the fully connected layers from the output of ROI Align to classification and regression. For the feature extraction network based on Resnet, that is, the global average pooling layer. Then a simple fully connected layer is added in the same position, and the dimension is set to 2048 dimension, and remove the dropout method, so that the neurons are no longer inhibited randomly during the back propagation process.Finally , Finally, the network is connected to two fully connected layers of coordinate regression and target classification.

3.2 RPN network adjustment

In Light-Head Rcnn, the RPN network is modified to {32×32, 64×64, 128×128, 256×256, 512×512} for the target of more shapes, and the aspect ratio of Anchor is still { 1:2, 1:1, 2:1}. However, considering the application scenario of the model, the target object has a single shape, and the small target detection is the research direction. Therefore, this article does not modify the original Anchor scaling scale, but will adjust the aspect ratio of the Anchor. Mask Rcnn's official model uses 80 kinds of datasets for training. The Anchor ratio can completely include each type of target. However, for the scenario used throughout this paper, that is, the detection and localization of the targets in service robot supermarket, the Anchor ratio needs to be adjusted. Fig.5 and Fig.6 shows the ratio before and after the modification of Anchor.

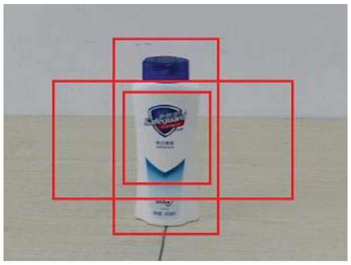

Fig.5 Anchor’s scale: 1:1,1:2,2:1

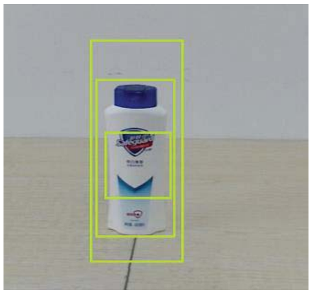

Fig.6 Anchor’s scale:1:1,1:2,1:3

It can be seen from Fig.5 that Anchor's landscape with a ratio of 2:1 does a lot of unnecessary calculations. For the data used in this paper, the width to height ratio of Anchor is changed from {1:1, 1:2, 2:1} to { 1:1, 1:2, 1:3}, that is, remove the horizontal proportion of the Anchor box, increase the vertical Anchor box, and make the model focus on the calculation of these three ratios, which can reduce the excessive calculation amount, save the training memory.

3.3 Remove Mask branch

Remove the mask prediction branch. For the target location task, the pixel-level target classification is removed, and the network is concentrated on the feature map extracted by the CNN for positioning and classification tasks, which can save the FCN calculation task and save a lot of training and prediction time. At the same time, the mask branch and the softmax classification prediction branch are parallel to each other, so removing the Mask branch will not affect the training and prediction of other layers of the network.

The improved Mask Rcnn network structure is shown in Fig.7.

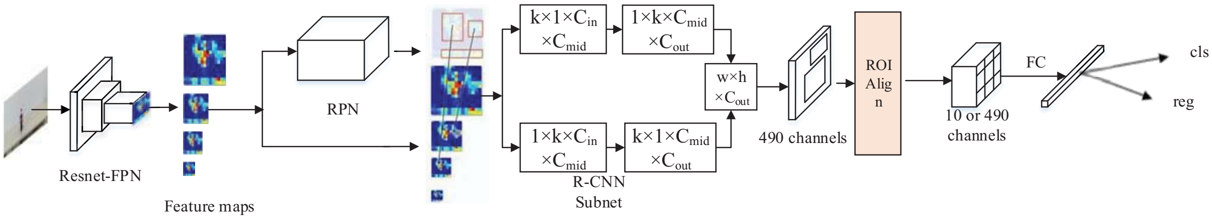
Resnet-FPN
Feature maps
RPN
$k\times1\times C_{in}$ $\times C_{mid}$
$1\times k\times C_{mid}$ $\times C_{out}$
$1\times k\times C_{in}$ $\times C_{mid}$
$k\times1\times C_{mid}$ $\times C_{out}$
R-CNN Subnet
$w\times h$ $\times C_{out}$
490 channels
ROI Alig n
10 or 490 channels
FC
cls
reg

Fig.7 The architecture of improved Mask Rcnn

4 Experiments

4.1 Experiment Platform

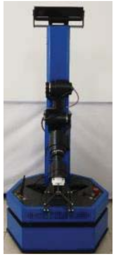

Fig.8 Sun@Home service robot

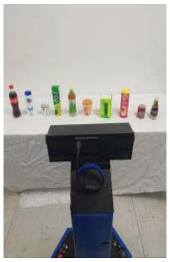

Fig.9 The actual scene of the robot collecting data

Fig.8 shows the our experimental robot -Sun@Home ,which consists of an omnidirectional wheel chassis, a lifting and lowering mechanism (lifting range of 0~500 mm), a laser radar with a detection range of 270°, a three-degree-of-freedom manipulator and a sensor required for this experiment—Microsoft Kinect II camera.

Images acquired by the Kinect II camera have higher resolution and color recognition, which allows the algorithm to extract key information from the target more accurately. Currently, Kinect II can capture the highest image resolution of 1920×1080.

4.2 Experimental scene and Data set

1) All data categories

coffe,cola,safeguard,toothpaste,milk,napkin,chips, Wangwang, youlemei,safeguard. Ten categories in total.

2) Number of dataset

Each category collects about 108 image data, a total of 1080 images. In this experiment, in order to verify the performance of the algorithm, only one kind of target appears in each picture in the data used, and the picture of the multi-category target is not added. Use the annotation tool to mark out 1080 JSON files, extract the coordinates of the targets.At the same time,use the same number of xml files to create VOC datasets and finally complete the dataset.

3) Training platform

CPU: Inter Core i7-6700K 4000GHz×8; GPU: Geforce GTX 1080; system environment: Ubuntu14.04 64-bit, CUDA version 9.0, CUDA Deep Neural Network (CUDNN) version 7.5; hard disk information: Solid state drive 256GB. To compare the experimental results, faster Rcnn also uses the same server platform for training.

Fig.9 shows the actual scene of the robot collecting data. and the distance between the robot and targets in the test is 1.5 times the distance when the data is collected.So the experiment scene is long-distance small targets detection.

4.3 Experimental results and analysis

In this paper, the improved Mask Rcnn is compared with the original Mask Rcnn and Faster Rcnn network, and the contrast experiment is carried out in the real environment in the context of serving the robot application scenario.

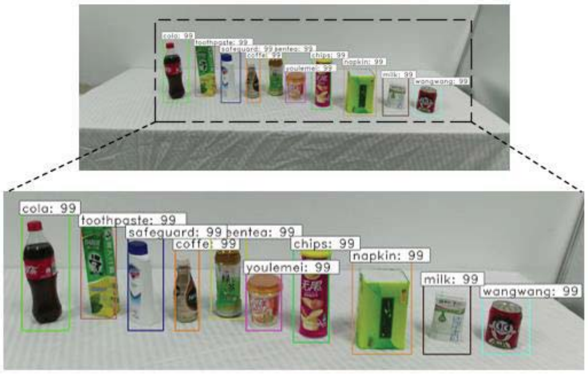

Fig.10 Improved Mask Rcnn experimental result

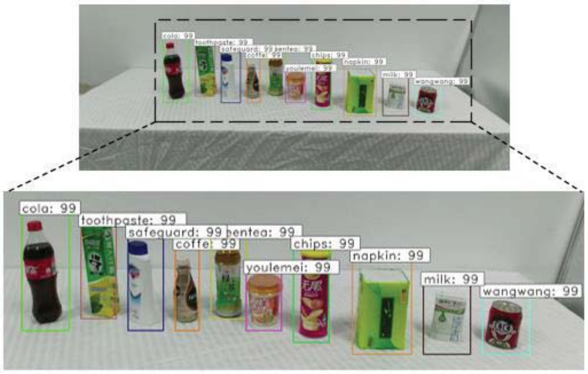

Fig.11 Original Mask Rcnn experimental result

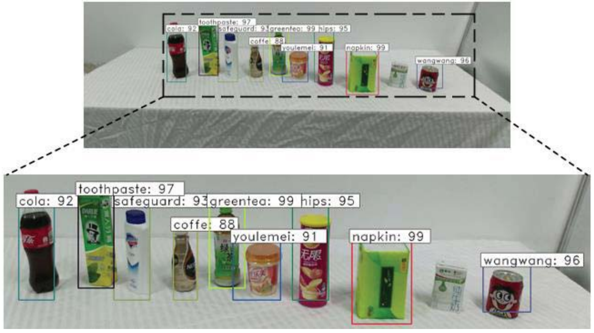

Fig.12 Faster Rcnn experimental result

This experiment tests all the targets in the datasets,Fig.10 shows the improved Mask Rcnn experimental result.Fig.11 shows the original Mask Rcnn experimental result.Fig.12 shows the Faster Rcnn experimential result.For the sake of clearity, the following images are the resules of zoomed in,in Fig.10,Fig.11 and Fig.12.

It can be seen from the experimental results that improved Mask Rcnn network ,original Mask Rcnn network and Faster Rcnn can locate and correctly classify targets from the datasets. At the same time, the two Mask Rcnn networks largely scale the region proposal box on the target size.However,Faster Rcnn has one missed:milk,the reason of which is that the milk is small in size and its color is similar to color of background when the detection distance is far away, so CNN mistakenly recongnizes milk as background when extracting target features.On the contrary,two Mask Rcnn networks' test results are better than Faster Rcnn.And through many experiments, the detection accuracy of improved Mask Rcnn network is the same as that of the original Mask Rcnn network; However, the detection speed of improved Mask Rcnn is speed up. The improved Mask Rcnn network is much faster than the original Mask Rcnn network. At the same time, after many experiments, it can be found that when the number of test targets in the experimental scene is less than 15, the service robot runs the original Mask Rcnn network, and the detection speed is about 8 seconds. When using the improved Mask Rcnn network, the service is utilized. Robot detects the same number of targets in the same scene at a speed of about 3.5 seconds. The detection speed of improved Mask Rcnn is at least twice as fast as the speed of original Mask Rcnn.

Based on the above experimental results, the reason why the improved Mask Rcnn detection speed is improved is that the network in Light-Head Rcnn is applied to Mask Rcnn, where the separation large convolution reduces the dimension of the feature map. At the same time, the network output is used a single 2048-dimensional fully connected layer replaces the two fully connected layers that take up a lot of computations,that saving a lot of training memory and running time while maintaining the same precision, so there is a faster increase in speed. This improved Mask Rcnn network plays a crucial role in the specific target grabbing of the robot. When the robot is far away from the targets,the speed of detection determines the capture efficiency of the target.

Finally, combined with the above experimental results and the results of many experiments in other test scenarios, the parameters of the improved Mask Rcnn network,the original Mask Rcnn network and the Faster Rcnn network in the long-distance home service robot experiment are summarized in Table 1.

Table 1: Comparison of improved Mask Rcnn , original Mask Rcnn and Faster Rcnn experimental parameters

	Original Faster Rcnn	Original Mask Rcnn	Improved Mask Rcnn
mAP	70%	85%	87%
Training	11h	13h	12h
Test	3.2s	8s	3.5s
False detection rate	20%	10%	9%

Missed Detection rate	10%	5%	4%

Table 1 shows that the improved Mask Rcnn is close to the training time of the original Mask Rcnn network,but which is slower than Faster Rcnn. The improved Mask Rcnn network is superior to the original Mask Rcnn network in terms of detection accuracy, false detection rate and missed detection rate. For the detection speed of the model, the improved Mask Rcnn network takes about 3.5 seconds, which is much faster than the original Mask Rcnn network but slower than Faster Rcnn. The increase in detection speed will save the service robot more time to complete other more complex tasks. This improvement plays an important role in the future specific target capture process and the service robot to complete other tasks.

5 Conclusion

Although both the improved Mask Rcnn and the original Mask Rcnn are based on Faster Rcnn, the experimental results is shown that both of them are superior to Faster Rcnn in location and recognition in real environment. At the same time, the complexity of network structure makes the detection time of the original Mask Rcnn slower than Faster Rcnn. However, the improved model of Mask Rcnn cannot meet the real-time requirements, but the detection speed of the improved Mask Rcnn network has been significantly improved, it is state of the art.

Finally, aiming at the problems of detection speed and precision, the team will continue to study in depth, which will provide a more optimized technical scheme for the future service robot to be able to detect targets in more detection scenarios.

References

[1] LV H M, ZHAO D, CHI X B, Deep Learning for Early Diagnosis of Alzhemer's Disease Based on Intensive AlexNet[J], *Computer Science*,2017,44(S1):50-60.

[2] Pengjie Tang,Hanli Wang,Sam Kwong. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition[J]. *Neurocomputing*,2017,225.

[3] QI Y C, ZHAO Z B, DU LQ,el. A Classification Method of Aerial Targets Based on VGGNet and Label Distribution Learning[J], *Electric Power Construction*, 2018,(2):109-115.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proc of ImageNet Large-Scale Visual Recognition Challenge Workshop. [S.l.]ICCV Press, 2013: 10-15.

[5] Girshick R. Fast R-CNN [C]//Proc of IEEE International Conference on Computer Vision . ICCV Press, 2015: 10-15.

[6] SHI J,ZHOU Y L , ZHANG Q Z. Item Recongnition Based on Faster R-CNN in Service Robot[J] , *Application Research of Computers*, 2019(11):1-9.

[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy,el. SSD: Single Shot MultiBox Detector[Z], *Computer Vision and Pattern Recognition(CVPR)*,2015.

[8] Joseph Redmon, Ali Farhadi. YOLOv3: An Incremental Improvement[Z], *Computer Vision and Pattern Recognition(CVPR)*,2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition[Z], *Computer Vision and Pattern Recognition(CVPR)*,2015.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick.Mask R-CNN[Z], *Computer Vision and Pattern Recognition(CVPR)*,2017.

[11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. Feature Pyramid Networks for Object Detection[Z], *Computer Vision and Pattern Recognition(CVPR)*,2016.

[12] WANG S, YANG K J.An Image Scaling Algorithm Based on Bilinear Interpolation with VC++[J], *Techniques of Automation and Applications*, 2008(07):44-45+35.

[13] François Chollet. Xception: Deep Learning with Depthwise Separable Convolutions[J], *Computer Vision and Pattern Recognition(CVPR)*,2016.