Detection method of robot grasp based on lightweight network

Luyuan Zhang1*, Yan Piao1 and Yuheng Liu1

1School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China

*Corresponding author’s e-mail: 2019100503@mails.cust.edu.cn

Abstract. When the robot arm uses the suction cup to grasp the task, it is faced with an unstructured scene, and it is difficult to accurately calculate the grasping posture of the robot due to the irregular placement of the object and its irregular shape. To solve this problem, a grasping detection method of manipulator based on lightweight convolutional neural network was proposed. Firstly, the MobileNet-YOLOV4 algorithm based on lightweight convolutional neural network was used to detect the target object in the image, and the classification and location information of the target were obtained. Then according to the final detection results of the image threshold segmentation, the anchor point is corrected, and finally the corrected positioning result is obtained. The grasping experiment was carried out on the Probot anno manipulator platform. The experimental results show that, compared with other image processing methods, the proposed method can realize fast detection and location of irregular target objects, and has better robustness for the diversity of object morphology and environment.

1. Introduction

Visual inspection and positioning is a key technique for smart robot servo control. Deep learning with its prominent target detection accuracy and reliability, has become a research hotspot in the field of robotic arm servo crawling in recent years.

With the in-depth study, how to accurately grab the grab posture of irregular goals, it is a technology bottleneck that intelligently crashed. Wang S Y et al. [1] carried out multi-scale filtering analysis on the input image in the framework of Canny algorithm, extracted its precise edge features, and established a template library by producing the detected object model with different rotation angles. Asif et al. [2] from the different levels of the image to predict the grabbing area, overcome the limitations of only the image capture area of the single level predict the image, and the method is superior to the latest method of Cornell grabbing data set. Du Xuedan et al. [3] using convolution based on region faster in deep learning neural network (faster region – based convolutional neural network Faster R - CNN) algorithm for object recognition and localization task, and their training target detection model has good generalization ability and higher accuracy, but the detection speed is slow, and has poor real-time performance.

Based on the above problems, this paper proposes a rapid detection and positioning method for robotic arms based on lightweight convolutional neural networks, and completes the rapid detection and sorting of items commonly used in life on the PROBOT Anno robotic arm platform. The method proposed herein uses the Mobilenetv3 network-based YOLOV4 network to capture the target positioning, the auxiliary algorithm handles a plurality of sites in the scene, acquire the target category and border position; Then, based on the detection results, the images are pre-processed and edge detected, and finally the corrected results are obtained. The methods proposed herein have high identification and positioning accuracy, and can be classified and improve efficiency.
2. The mechanical arm sorting system and algorithm model
Although the sorting objects of different sorting tasks in industrial systems are different, but as long as they obtain the positioning coordinates and category information of the target object, quickly and accurate sorting of complex targets can be achieved. The algorithm flow is divided into the following two parts.

- Use a target detection algorithm based on a lightweight convolutional neural network to detect the target object in the image to obtain its category and predictive frame position information.

- Cut the image based on the test results, pre-cutting the cropped image is pre-processed and edge detection to obtain the calibrated target positioning coordinates.

3. Objective Detection

3.1. Object Detection Scheme
YOLOv4 algorithm is improved on the basis of Yolo, Yolov2, Yolov3. First, first segment the target picture to be detected into different sizes, each grid is responsible for different areas, if the center of the target is falling in a certain center in the grid, the target is detected by the grid. In view of model performance and industrial application requirements, this paper uses YOLOv4 algorithm to achieve visual inspection in mechanical arm grab.

3.2. YOLOv4 convolutional neural network
The network structure of the YOLOV4 algorithm mainly includes a backbone feature extraction network (CSPDarknet 53), a spatial pyramidal structure (SPP), path aggregation network (PANet), when the input picture is 416 × 416 network structure diagram, as shown in the Fig.1.

Fig.1 YOLOv4 network structure

In the part of characteristic pyramid structure, YOLOV4 adopts SPP structure and PANet structure. The SPP structure is the maximum pooling of the result of the characteristic layer P5 after the third convolution. After PANet structure, different feature layers are fully fused, which can effectively improve the feature extraction ability of defects. Finally, YOLOv4Head using PANet three characteristics of the layer after processing to predict the results.
3.3. MobileNetv3-YOLOv4 lightweight neural network

Although the YOLOV4 detection model has excellent performance, it uses the amount of CSPDarkNet backbone network parameters, and the calculation parameter is large in the process of feature extraction, and it takes a long time. Since the robot should have a high real-time performance in grasping detection to quickly respond quickly, it is imperative to improve the model to reduce the amount of parameters. In order to improve the detection speed, this article adjusts the base network to MobileNet-V3. MobileNet-V3 is a lightweight deep convolutional neural network that improves and optimizes on the basis of MobileNet-V2, which can better meet the requirements of the mechanical arm on the detection speed in the grab mission.

3.3.1. Depth separable convolution

Googler et al. proposed MobileNet-v3, a lightweight model for embedded mobile devices, as shown in the Fig. 2. The core idea of its use is deep separable convolution blocks. There are two versions of MobileNet-V3, MobileNet-V3 Small and MobileNet-V3 Large, in order to make the real-time grasp detection of the robotic arm higher, the small version is used as the backbone network. MobileNet-V3 combines the deep separable volume of MobileNet-V1 [4] and the inverse residual structure of MobileNet-V2[5] to increase the attention mechanism.

![Fig2. MobileNetv3 network structure](image)

In the YOLOV4 network structure, the backbone extraction network CSPDarknet53 mainly carries out preliminary feature extraction, while SPP and PANet mainly carry out enhanced feature extraction to extract better features. At this point, MobileNet's core idea of deep separable convolution can be used to replace CSPDarknet53 in YOLOv4 for feature extraction, and three initial effective feature layers of the same size can be enhanced for feature extraction. Combine MobileNet-V3 with PANet. Taking 416x416 image input as an example, this paper combines the feature map of the 39th layer with the feature map of the last bottleneck layer, and applies two up-sampling. The fused feature graph adopts 1x1 convolution to enhance the dimension of the feature graph. Then, the 46th layer is up-sampled and fused with the 11th layer feature map. Third convolution is carried out to obtain 52x52 small target detection feature map. The structure and parameters of the MobileNet-YOLOV4 backbone network are shown in Fig. 3.
4. Experiment

4.1. Experimental Platform
This article is based on the depth learning framework Pytorch and the robot software platform ROS, using Python development. The deep learning hardware platform is a memory with Intelcorei7-8700K processor, 64G memory, NVIDIA GTX10808G, and software environment is Ubuntu16.04, CUDA10.0 and OpenCv3.0. The robot arm is a domestic six-degree-of-freedom robot arm, Anno mechanical arm, which is crawled with starther suction cup. The mechanical arm body and the controller are shown in the Fig.4.
4.2. Experimental Data Sets and Parameters

In this experiment, objects in Cornell data set were used as target detection training samples and re-labeled into VOC format for training. The dataset consists of 240 and 855 images from different samples. Flip all the images left and right, and then flip them up and down to quadruple the original size. Before training, the image of the Cornell data set is randomly divided, training set, verification set, and test set ratios are 8: 1: 1.

The parameters of model training are set as follows: batch processing parameter batch_size is 24, initial learning rate init_learning_rate is 0.0043, learning rate decay strategy is exponential decay, decay_factor is 0.95, and the maximum number of iterations is 20,000.

4.3. Model Evaluation

In target detection, the average accuracy ($AP$) is calculated through accuracy ($P$) and recall rate ($R$), and the average $AP$ value ($mAP$) is used as the comprehensive evaluation index of the detection accuracy of the algorithm. $AP$ is used to evaluate the accuracy performance of the model in a single detection category, and the larger $mAP$ is, the higher the overall detection accuracy is. Formula (1) is the calculation formula for accuracy, Formula (2) is the calculation formula for recall rate, and Formula (3) is the calculation formula for MAP.

$$ P = \frac{TP}{TP + FP} $$

$$ R = \frac{TP}{TP + FN} $$

$$ mAP (class) = \frac{\sum AP}{N(\text{class})} $$

Among them, $TP$ represents the number of correctly detected targets, $FP$ represents the number of wrongly detected targets, $FN$ represents the number of missed detected targets, $AP$ represents the sum of $AP$ values of all categories, and $N(\text{class})$ Represents the total number of categories.

The detection speed is also one of the important indicators to judge the performance of the algorithm. In this paper, the average detection time of a single image is used as the evaluation index of the detection speed.

$$ Time = \frac{TotalTime}{NumFigure} $$

In the formula, TotalTime is the total time of detection, $NumFigure$ is the number of pictures detected.

5. Experimental Results and Analysis

5.1. Detection Accuracy Contrast Verification

The objects detected were common living objects in the Cornell dataset, including bananas, scissors, and ping-pong rackets. In the same experimental environment, experiments were conducted based on the MobileNet-YOLOv4 method and YOLOv4 method, and the detection effect was shown in the figure. Experimental results show that the proposed method has high accuracy in target recognition and location. However, the accuracy of the recognition result and the target position prediction box is low. The green point is the center point of the prediction box, and the purple point is the location grab point.
The detection performance index quantification of the above two methods is shown in Table 1. The results show that the presentation is promised to ensure higher inspection accuracy.

| class   | Average time/s | Detect correct number | Average precision | Average time/s | Detect correct number | Average precision |
|---------|----------------|------------------------|-------------------|----------------|------------------------|-------------------|
| Racket  | 0.078          | 285/300                | 95%               | 2.734          | 300/300                | 100%              |
| Shears  | 0.080          | 281/300                | 93.7%             | 2.721          | 300/300                | 100%              |
| Banana  | 0.082          | 290/300                | 96.7%             | 2.738          | 299/300                | 99.7%             |

6. Summary
This paper proposes a rapid detection method based on a lightweight convolutional neural network, and proves the effectiveness of the method through the probotanno robotic arm platform. In the next work, this method will further integrate into a three-dimensional visual algorithm, enhance the complexity of scenes and detection objects, and improve the processing capabilities of complex scenes.

Reference
[1] WANG S Y, LIN H, GAI R L, et al. An application of vision technology on intelligent sorting system by delta robot[C]/ IEEE International Conference on E-health Networking. Piscataway: IEEE Press, 2017.
[2] ASIF U, TANG J, HARRER S. Densely supervised grasp detector (DSGD) [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8085-8093.
[3] DU X D, CAI Y H, LU T, et al. A robotic grasping method based on deep learning[J]. Robot, 2017,
39(6): 820-828, 837.

[4] Howard A.G, Zhu M, Chen B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[Z]. [2020-07-03], https://arxiv.org/abs/1704.04861.

[5] Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted residuals and linear bottlenecks[Z]. [2020-07-05], https://arxiv.org/abs/1801.04381.

[6] Bochkovskiy A, Wang C Y, Liao H. YOLOv4: Optimal speed and accuracy of object detection[Z]. 2020, [2020-07-03], https://arxiv.org/abs/2004.10934.