Manipulator Grasping Based on Object Detection

Xin Shu[1,2], Chang Liu[1, a] and Tong Li[1, b]

[1]National Institute of Standards and Tec State Key Laboratory of Transducer, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China;

[2] University of Chinese Academy of Sciences, Beijing 100049, China.

[a]tuengineer@qq.com, [b]tli@mail.ie.ac.cn

Abstract. To make sure that manipulator can perform well on novel environment, a new grasping approach based on object detection is proposed. A pose estimation network is used as usual to predict the grasping pose of the object while an object detection network is added before it as the input information of the pose estimation. This combination of object detection and pose estimation improves the grasping accuracy by 28% and shows grasping robustness to objects which are not seen by manipulator before.

Keywords: Grasping; Neural networks; Robot vision systems.

1. Introduction

Grasping operation of manipulator is becoming a more and more popular research direction nowadays. Classical manipulator grasping is based on manual teaching. Manipulator can grasp the object after being taught to a fix position. But it will face some challenges when it is placed to a novel environment. Therefore, with the development of artificial intelligent, manipulator need to learn by itself. Grasping methods based on template matching and deep learning are proposed. For template matching methods [1][2], they firstly build 3D point cloud and then match with template in the template library. These methods do not perform well on object which is self-occlusion or mutual-occlusion and require a large set of template library. As for learning based method, convolutional neural networks [3][4] and autoencoders [5] are used to predict grasping pose on a RGB image. These methods rely on texture of objects and have trouble faced texture-less objects.

Therefore, a grasping method based on object detection and pose estimation is proposed in this paper, which shows great robustness to both texture-less objects and novel objects which do not appear in template library. The main contribution of this paper is that we combine the object detection network and pose estimation network to improve the grasping accuracy.

2. Design of Grasping Network

2.1 Object Detection Network

Object detection outputs the bounding box and class of the detected object. Our detector is based on two-stage, it firstly uses region proposal network to propose bounding boxes and then regress the bounding box and classify the object in the box. Resnet-101 is a kind of art-of-state detection framework and based on shortcut connection which allows a large network depth of 101 convolutional layers. We use Resnet-101 to extract the feature map of the whole RGB image. Region proposal network works on the feature map and outputs bounding boxes faster compared with selective search. The ROIPool layer is added to obtain the fix-sized global feature maps from the whole feature map with different sizes of bounding boxes. These global feature maps are sent to the following network to regress and classify. This regression and classification network can share their weights with region proposal network, which accelerates the training and inference method of object detection. Besides, smooth L2 function of regression and SoftMax function of classification is designed as the multi task loss function of object detection, which can converge well on the method of training.

2.2 Pose Estimation Network

Pose estimation network includes five convolutional layers and two fully-connected layers. We sample a fix number of patches from the whole image and sample patches can be seen in Figure 1. Fix-sized patches are sent to the five convolutional layers to extract the image feature map. Two fully-connected layers are designed to predict the grasping attitude and grasping position. The estimation of grasping attitude can be regarded the as a classification problem instead of a regression problem as we set the step of the attitude of 10° and this performs a better result than regression. SoftMax function is used as the loss function of classification. Grasping position is defined in a two-dimensional pixel and we just regress it. When it comes to inference, we select the attitude and position with the highest soccer from the whole patches and poses as the estimated pose.

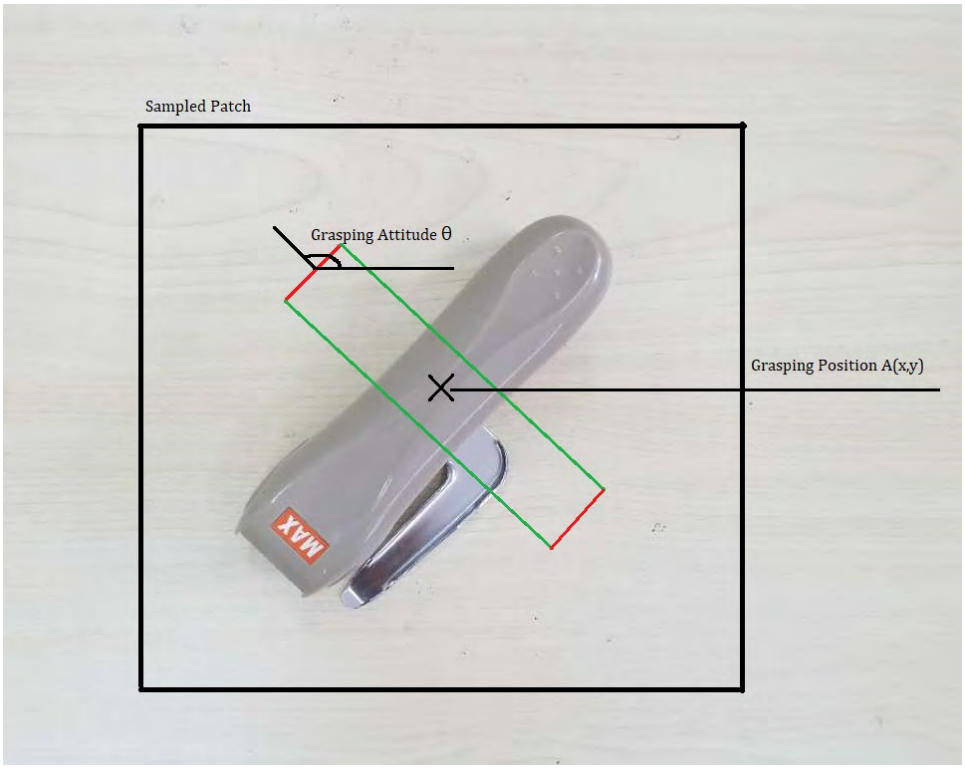

Fig. 1 Sampled Patch

In this paper, we combine object detection network and pose estimation network in the following way. We sample patches in the range of bounding box detected instead of the whole image, this way can improve the estimation accuracy and remove the obvious error in estimation.

3. Experiments

We use Kinect to extract the RGB image of grasping object. Detection network is trained and validated with COCO_2017, while pose estimation network is trained with the grasping dataset of Carnegie Mellon University and is validated with the dataset of Cornell University. This validation dataset is novel to manipulator since it has not appeared in training dataset.

3.1 Results of Object Detection

We train the object detection network about 500000 iterations and the network learning rate is set to 0.001 at beginning. We decrease it by 0.1 at the 400000 iteration and 450000 iteration since the learning rate is relatively large. we also use COCO_2017 dataset for validation, which includes 5000 images. The result in Table 1 shows that the network trained as the above mentioned can easily meet the requirement of the following step.

Table 1. Object Detection Accuracy

Method	Dataset	AP, IOU:			AP, Area:		
		0.5:0.95	0.5	0.75	S	M	L
Our Network	Val 2017	36.0	62.4	37.7	22.4	46.6	61.6

3.2 Results of Pose Estimation

We propose a new network combination approach to improve the grasping accuracy. The object detection network can output the bounding box of object and we sample patches in the range of bounding box, which can improve the grasping accuracy in the similar calculation time. We train the pose estimation network and validate it with the dataset of Cornell University.

Table 2. Pose estimation Accuracy

Approaches	Position success (%)	Position error avg.± std(mm)	Attitude-success (%)	Attitude error avg.± std(deg)	Pose Success (%)
Pose Estimation	42.36	12.77±6.36	50.06	21.97±7.25	23.22
Object Detection+ Pose Estimation	88.30	11.18±6.12	59.57	19.15±6.49	51.06

As shown in Table 2, two groups of evaluations are validated. We categorize the pose failure as exceeding errors of more than 30mms in position and 20 degrees in attitude compared with the ground truth. The first one is that we use the pose estimated only and it comes out to a low pose accuracy. While the second one combines the pose estimation network with the object detection network, which improves the pose accuracy by around 28% and shows more robust to novel objects which have not appeared in training dataset.

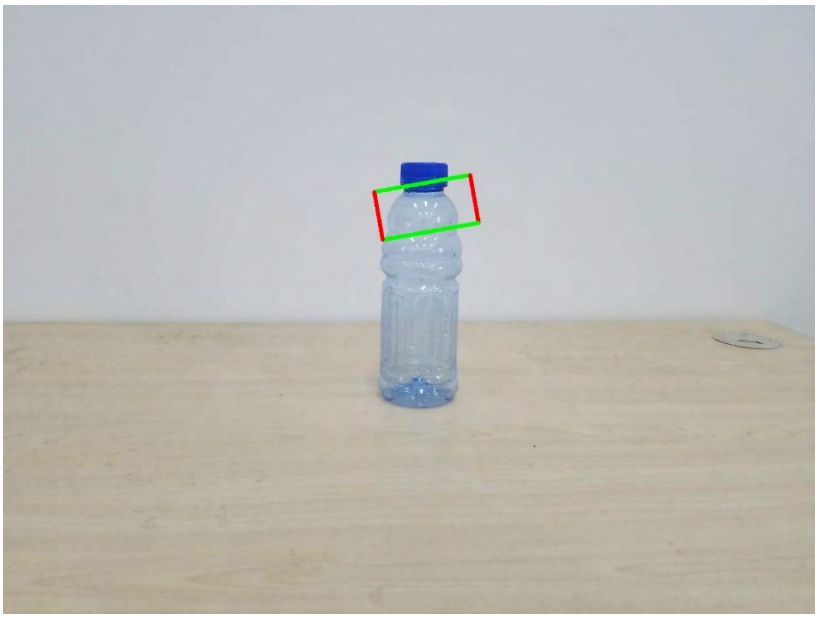

Fig 2. Result of Combined Network

4. Conclusion

In this paper, we propose a new manipulator grasping method based on object detection. A detection network and a grasping pose estimation network both based on deep learning are combined into a stable network. With the help of this network, manipulator can detect the position of object and

calculate the pose of grasping. Compared to the original grasping network, the accuracy of grasping can be increased by 28%.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61802363).

References

[1]. A. Zeng, K. T. Yu, S. Song, et al., “Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge,” IEEE Robotics & Automation Magazine, vol. 39, no. 6, pp. 1383-1386, Jun. 2017.

[2]. J. M. Wong, V. Kee, T. Le, et al., “SegICP: Integrated Deep Semantic Segmentation and Pose Estimation,” IEEE/RSJ International Conference on Intelligent Robots & Systems, Vancouver, Canada, 2017, pp. 5784-5789.

[3]. Y. L. Cun, B. Boser, J. S. Denker et al., “Handwritten digit recognition with a back-propagation network,” Advances in Neural Information Processing Systems, vol. 2, no. 2, pp. 396-404, Feb. 1990.

[4]. A. Krizhevsky, I. Sutskever, G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA, 2012, pp. 1097-1105.

[5]. B. A. Olshausen, D. J. Field, “Sparse coding with an overcomplete basis set: a strategy employed by V1?” Vision Research, vol. 37, no. 23, pp. 3311-3325, Dec. 1997.