Robust Robot Grasp Detection in Multimodal Fusion

Qiang Zhang[1,2,3,*], Daokui Qu[1,2,3], Fang Xu[1,2,3], and Fengshan Zou[3]

[1] *State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, No.114 Nanta Street, Shenhe District, Shenyang 110016, P.R. China*
[2] *University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, P.R. China*
[3] *SIASUN Robot & Automation Co., LTD., NO.16 Jinhui Street, Hunnan District, Shenyang 110168, P.R. China*

Abstract. Accurate robot grasp detection for model free objects plays an important role in robotics. With the development of RGB-D sensors, object perception technology has made great progress. Reach feature expression by the colour and the depth data is a critical problem that needs to be addressed in order to accomplish the grasping task. To solve the problem of data fusion, this paper proposes a convolutional neural networks (CNN) based approach combined with regression and classification. In the CNN model, the colour and the depth modal data are deeply fused together to achieve accurate feature expression. Additionally, Welsch function is introduced into the approach to enhance robustness of the training process. Experiment results demonstrates the superiority of the proposed method.

1 Introduction

With the development of visual sensors, robot vision perception technology has made great progress. Especially in recent years, robots can perceive the colour and the distance properties of the environment due to the development of RGB-D sensors like Kinect and Xtion. In the field of home service robot applications, vision based robot grasp detection has been an important research direction in robot technology for the reason that it can improve the level of human-computer interaction.

For the service robot, the objects to be operated are divided into two categories, the known and the unknown objects. The known objects mean that robots have stored the model of the object. Under this circumstance, robot grasp detection procedure can be separated into two processes. The first process is to detect the specific object and the second one is to estimate the object's pose and find out the proper grasp point. However, on the other hand, the robot cannot always store models for all objects. For example, when the service robot enters a strange environment, objects in this exception would be unknown for the robot. Then, grasping the unknown object or the model free object can be a difficult task for the robot.

During last few years, several significant approaches [1-9] have been proposed to solve the model free grasp detection problem. In 2006, 2D grasping point representation method was proposed by A. Saxena [1]. In 2010, Q. V. Le [2] introduced a method by using multiple contact points to represent grasp locations. In 2011, Y. Jiang [3] presented a new representaiton mehtod which describe the robot grasp as a 5-dimensiaonal oriented rectangle. In order to learn the representation, the SVM ranking algorithm was introduced in the learning process. With the development of deep learning, convolutional neural networks has been introduced as powerfull visual models [4-5]. In 2013, Lenz [6] proposed a convolutional networks based method for robot grasp detection. The method using sliding windows to generate multiple candidate grasps firstly and true grasps could be retained by the classifier. In 2016, J. Wei [7] proposed a multimodal fusion based deep extreme learining machine for robot grasp recognition. L. Trottier presented a dictionary learning method in the same year. However, these three methods have disadvantages of high complexity. In 2015, J. Redmon [9] introduced a real-time detection method by using single-stage regression. The end-to-end solution reduced the training difficulty. However, the coarse data fusion leads to low detection accuracy. Meanwhile, the training process has slow convergent speed because of outliers.

In order to improve robot grasp detection accuracy, we adopted an improved approach. The contribusions are the following. First, we proposed a robust loss function by taking Welsch function into consideration. Second, we introduced Atrous convolution algorithm to our architecture to improve the local expression ability of features. Both these two contributions can improve robot grasp detection accuracy.

The rest of the paper is orgnised as follows. Section 2 describes the deep regression model for robot grasp detection. Details of the proposed method are discussed

* Corresponding author: zhangqiang@sia.cn

in Section 3. Section 4 evaluates performance of the proposed method. Finally, Section 5 gives the summary and conclusion.

2 Related works

Generally, object models should be constructed before robotic grasping. For instance, the specific object can be expressed by invariant local features [10], sparse 3D point cloud model [11] and dense 3D point cloud [12]. However, building object model is difficult and time consuming. This limits the ability of the robot to adapt to the environment. Recently, category based object detection [13-15] has been applied into robot applications. It is still difficult to determine object's pose.

Y. Jiang [3] proposed a rectangle representation method for robot grasp, which skips object detection and pose estimation process. Each grasp is described by a rectangle with its central coordinates, size, and orientation. This expression method greatly simplifies the complexity of the model. With the development of deep learning, convolutional neural networks based grasp detection method was proposed by I. Lenz [6]. However, the sliding window approach decreases the detection efficiency. L. Trottier [8] proposed a dictionary learning method for robot grasp detection. Although it takes the advantage in high accuracy, proposal procedure makes it inefficient. J. Redmon [9] presented an end-to-end solution for real time grasp detection. Nevertheless, the shallow data fusion architecture can affect the accuracy of the detection. At the same time, full-connected layers reduced the local feature representation ability. D. Guo [16] introduced a pipeline using reference rectangles on the feature map. For that only colour information was employed, the detection accuracy is still not high enough.

3 Robust robot grasp detection

3.1. Architecture

In our work, the robot grasp is represented by a grasping rectangle which is a 5-dimontional vector $\{x, y, w, h, \theta\}$ the same as in [5]. In this vector, x and y denotes the central coordinates, w and h denotes the rectangle size, θ represents the angle between the rectangle and the horizontal direction. Inspired by [9], robot grasp detected in the regression way. The proposed architecture is shown in Figure 1.

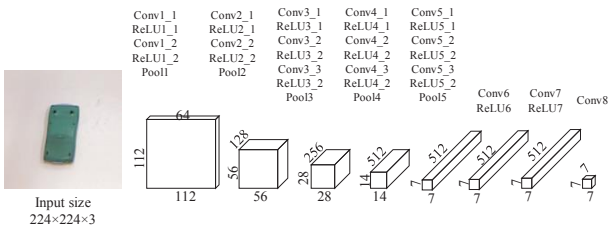

Fig. 1. The robot grasp detection architecture

The proposed architecture is a modification of the VGG-16 networks [17]. The full connection layers of VGG-16 were replaced by three convolutional networks aiming at improving the expression ability of local information according to hole filling algorithm [18]. In the architecture, all convolutional layers have the same kernel size 3×3 and the same stride 1. After each convolutional layer, batch normalization was employed to improve detection accuracy. All pooling layers are maxpooling layers and the same stride is 2. In the architecture, the input colour image size is 224×224×3 and the output matrix size is 7×7×7.

3.2. Detection methodology

The detection architecture is designed aiming at combining classification with regression. For each input image, we detect 49 (7×7) results, and each result is a 7 dimensional vector. As can be seen in Figure 2, the vector indicates the graspable probability of the detection result. Meanwhile, it represents the location of the graspable rectangle. The detection result is illustrated in Figure 2.

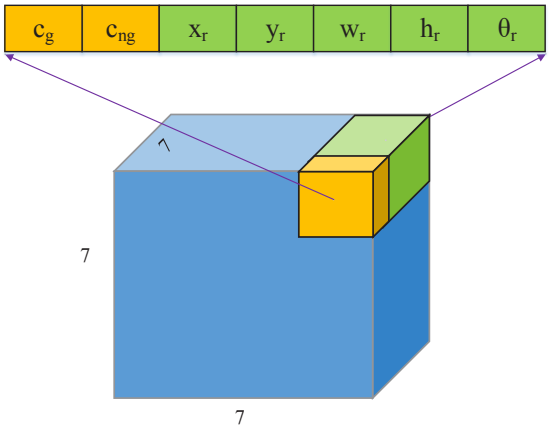

Fig. 2. Detection result in each vector

The loss function of the proposed architecture is as follows.

$$L(c,r)=\sum_{i=1}^{N}\left[L_c(c_i)+\lambda L_r^g\left(x_i,y_i,w_i,h_i,\theta_i\right)\right] \quad (1)$$

In equation(1), the loss function is composed by two parts: the classification part and the regression part. L_c is the Softmax Loss between the classification result and the truth label. The truth label is 1.0 when the graspable rectangle centre is located in the block, otherwise the value is 0. L_r^g indicates differences between the regression result and the true rectangle locations. The weight term λ is set to 1.0 according to validation result. In the training process, truth values of $\{x, y, w, h\}$ are all normalized as:

$$\begin{cases} x_n = \mathrm{mod}\left(x,32\right)/32 \\ y_n = \mathrm{mod}\left(y,32\right)/32 \end{cases} \quad (2)$$

$$\begin{cases} w_n = w / w_{max} \\ h_n = h / h_{max} \end{cases} \tag{3}$$

In equation (2), the value 32 indicates the sizes of each separate block. In equation (3), w_{max} and h_{max} are the largest width and height of the graspable rectangles.

During training process, back propagation is to adjust parameters to meet detection requirements. Outliers can lead to slow convergence rates. Meanwhile, small residuals contribute less to the back propagation process. Inspired by [19], we introduced a more robust loss function named Welsch to retain relative large residuals. The function is designed as follows:

$$f(x) = \frac{\alpha^2}{2}\left[1 - e^{-(x/\alpha)^2}\right] \tag{4}$$

In equation (4), α=2.9846. The function and its derivatives are shown in Figure 3.

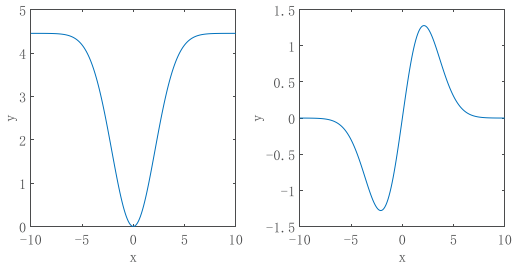

Fig. 3. Welsch function and its derivative

Then, the full loss function L_r^g is designed as follows.

$$\begin{aligned} L_r^g = \sum_{reg \in pos} \{ & f(x_n - x_r) + f(y_n - y_r) \\ & + f[\log(w_n) - \log(w_r)] + f[\log(h_n) - \log(h_r)] \\ & + f[\sin(\theta_n - \theta_r)]\} \end{aligned} \tag{5}$$

3.3. Multimodality fusion

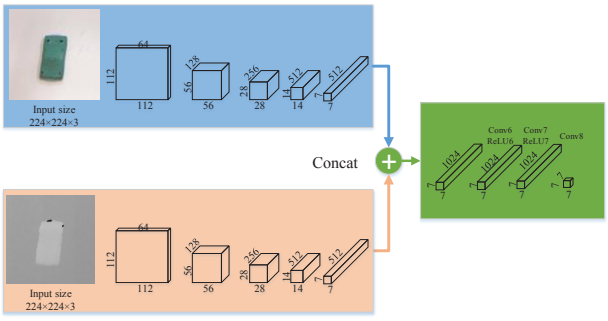

Fig. 4. Feature fusion principle in the regression architecture

Nowadays, RGB-D sensors make it possible for robots to get distance information of the scene. For robot grasp, colour and depth data fusion has been a research hotspot. In our work, we proposed a data fusion method based on feature fusion. We first trained the RGB based neural networks. The same as RGB image based networks training, we feed depth image to the similar networks. In order to simplify the training complexity, each depth image was expanded to 3D. After all, we concatenate the features from the RGB image and the depth image, and retrain layers Conv6, Conv7 and Conv8. Feature fusion methodology is illustrated in Figure 4.

4 Experiments

4.1. Experimental methodology

We adopted the Cornell Grasping Dataset to evaluate the proposed method. The dataset contains 240 different objects. The total number of the samples is 885. For each sample, RGB image and 3D dense cloud image is provided. The 3D dense cloud images were first transformed to depth images. For the reason that each image contains multiple graspable and non-graspable locations, the proposed method can be applied to these multiple label detection task. In order to improve the robustness of the networks, we performed an extension for the original dataset in image cropping, rotation. The number of the training examples was extended to 3540 and the number of the graspable rectangles was over 20,440. The dataset was separated into two parts, 80% was the training set and the remaining 20% was test set. Moreover, all images were cropped into the size of 224×224.

In order to evaluate the performance of the proposed method and the conventional methods, the grasp intersection over union (IoU) metric was employed in our work. The IoU is defined as:

$$IoU = \frac{G_{detection} \bigcap G_{truth}}{G_{detection} \bigcup G_{truth}} \tag{6}$$

In the upper equation, the numerator and the denominator of the right part denotes the overlap area and the union area respectively. The grasp was seen as a right one only when the IoU was above 0.25 and the angle difference between the detected grasp and the truth was less than 30°. Two experiments were conducted in order to evaluate the image-wise and object-wise detection accuracy.

4.2. Training details

As mentioned previously, our proposed architecture is a modification of VGG-16 networks. Parameters of the last three layers were generated randomly, other parameters were initialized by the pre-trained VGG-16. Then we converted the model to perform robot grasp detection on the RGB and depth image dataset respectively. Features corresponding to the RGB and depth images were extracted and concatenated to perform multiple modalities fusion. The last three convolutional neural network layers were re-trained accordingly.

During training and validation process, Caffe deep learning library was employed. The experiment was performed based on two NVIDIA Tesla K80 GPUs with 24GB memory. In order to speed up the training process, the NVIDIA CuDNN library was taken into our work. In

the training procedure, we took stochastic gradient descent method to optimize the model.

4.3. Results and analysis

We compared our proposed method with algorithms proposed by Y. Jiang [3], I. Lenz [6] and J. Redmon [9]. Experimental results can be seen in Table 1. In Table 1, the proposed method has the better detection accuracy of 88.90% in image-wise split and 88.20% in object-wise split. The detection speed of the proposed method is 117ms per image. Some detection results can be seen in Figure 5.

Table 1. Detection results of different algorithms.

Algorithm	Detection accuracy		speed
	Image wise split	Object wise split	
Jiang	60.50%	58.30%	-
Lenz	73.90%	75.60%	13,500ms
Redmon	88.00%	87.10%	76ms
Ours	88.90%	88.20%	117ms

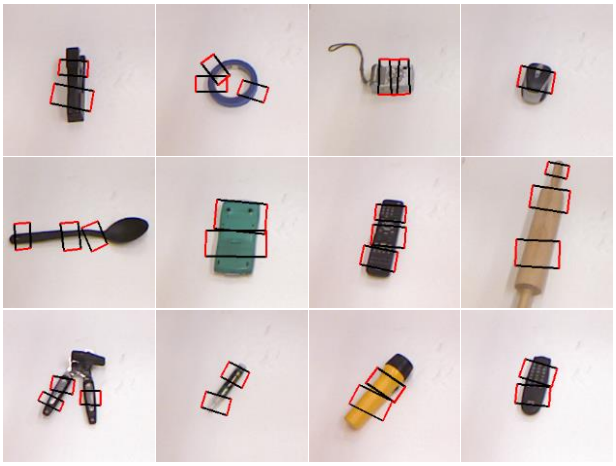

Fig.5 Detection results of the proposed method

According to the experimental results, the detection accuracy is higher than that of the traditional convolutional neural networks. The accuracy of the proposed method is increased for the reason of the reasonable loss function and the modified neural networks. The speed of our proposed method is slightly slower than Redmon's because that the proposed architecture is much more complex. However, the speed is still acceptable. Moreover, the end-to-end convolutional neural network structure takes the advantage of high speed compared with auto-encoder based learning method.

We also applied the proposed method to our application. The detection accuracy reached to 82%. Some detection examples are shown in Figure 6.

5 Conclusions

Robot grasp improves the ability of human robot communion. This paper proposes a robust method to detect robot grasps. The main contributions of the paper are: (1)We propose a new robust loss function for robot grasp detection using deep neural networks. By taking Welsch function into consideration, back propagation by interior points with small contributions and outliers can be reduced. This can lead to robust training progress. (2)Hole filling algorithm is introduced to improve the expression ability of local features. Experimental results show that the proposed method achieves the superior performance in robot grasp detection.

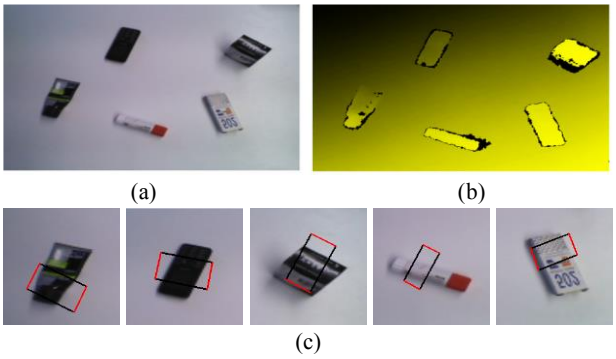

Fig.6 Detection results in our application: (a) The RGB image; (b)The depth image; (c)Detection results

Acknowledgements

We give our thanks to Shenyang SIASUN Robot Automation Co., LTD for funding the research program. This work was supported by the National Key Research and Development Program of China under Grant No. 2016YFF0202701.

References

[1]Saxena A, Driemeyer J, Kearns J, et al. Robotic grasping of novel objects[C]//Advances in neural information processing systems. 2007: 1209-1216.

[2]Le Q V, Kamm D, Kara A F, et al. Learning to grasp objects with multiple contact points[C]//Robotics and Automation (ICRA), 2010 IEEE International Conference on. IEEE, 2010: 5062-5069.

[3]Jiang Y, Moseson S, Saxena A. Efficient grasping from rgbd images: Learning using a new rectangle representation[C]//Robotics and Automation (ICRA), 2011 IEEE International Conference on. IEEE, 2011: 3304-3311.

[4]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[5]Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[6]Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps[J]. The International Journal of Robotics

Research, 2015, 34(4-5): 705-724.

[7]Wei J, Liu H, Yan G, et al. Robotic grasping recognition using multi-modal deep extreme learning machine[J]. Multidimensional Systems and Signal Processing, 2017, 28(3): 817-833.

[8]Trottier L, Giguère P, Chaib-draa B. Dictionary Learning for Robotic Grasp Recognition and Detection[J]. arXiv preprint arXiv:1606.00538, 2016.

[9]Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks[C]//Robotics and Automation (ICRA), 2015 IEEE International Conference on. IEEE, 2015: 1316-1322.

[10]Kurzejamski G, Zawistowski J, Sarwas G. Robust method of vote aggregation and proposition verification for invariant local features[J]. arXiv preprint arXiv:1601.00781, 2016.

[11]Hsiao E, Collet A, Hebert M. Making specific features less discriminative to improve point-based 3D object recognition[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 2653-2660.

[12]Yi Z, Li Y, Gong M. An Efficient Algorithm for Feature-Based 3D Point Cloud Correspondence Search[C]//International Symposium on Visual Computing. Springer International Publishing, 2016: 485-496.

[13]Tychsen-Smith L, Petersson L. DeNet: Scalable Real-time Object Detection with Directed Sparse Sampling[J]. arXiv preprint arXiv:1703.10295, 2017.

[14]Song S, Xiao J. Deep sliding shapes for amodal 3D object detection in RGB-D images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 808-816.

[15]Kehl W, Tombari F, Navab N, et al. Hashmod: a hashing method for scalable 3D object detection[J]. arXiv preprint arXiv:1607.06062, 2016.

[16]Guo D, Sun F, Kong T, et al. Deep vision networks for real-time robotic grasp detection[J]. International Journal of Advanced Robotic Systems, 2016, 14(1): 1729881416682706.

[17]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[18]Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. arXiv preprint arXiv:1606.00915, 2016.

[19]Lee C C, Chiang Y C, Shih C Y, et al. Noisy time series prediction using M-estimator based robust radial basis function neural networks with growing and pruning techniques[J]. Expert Systems with Applications, 2009, 36(3): 4717-4724.