An Robot Vision Grasping Network Based on Inception-Lite

Liang Xu1, Ziwei Zhou1* and Chaoyang Wang1

1 School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan, Liaoning, 114051, China
*Corresponding author’s e-mail: zzw@ustl.edu.cn

Abstract. Using computer vision algorithms to find the grasping position of unknown objects is a key issue in the development of intelligent robot technology. In order to address the problem of robot grasping, firstly, based on the analysis of Inception-v1, a novel Inception-Lite convolution module is proposed, and then a grasping network GRASPNET-CLF is designed based on the convolution module. Finally, the grasping network GRASPNET-CLF is trained and verified by GPU on Cornell grasping dataset. The experimental results show that the GRASPNET-CLF has better performance in the same kind of network, and meets the real-time requirements.

1. Introduction

With the progress and development of science and technology, industrial robots play a very important role in the industrial production environment. The control program can be written by the programmer in order to make the robot complete the task according to the predetermined program in the designated position. However, it is a very challenging problem to use a robot to grasp the target, because it is very difficult to use sensors to understand the surrounding environment.

The early research on robot grasping can be traced back to the discussion of interaction planning between a multi fingered robot and environment by Grupen et al. [1]. Kamon et al. [2] proposed a visual inspection method, which divided the grasping problem into grasping point selection and grasping quality prediction. [3-7] all find the right grasping position by 3D simulation.

In the past two decades, the theory and technology of robot grasping have been developing continuously. One of the core problems of grasping detection is how to determine a suitable grasping position. Early detection of grasping position is to detect a grasping point. Until 2011, Jiang et al. [8] proposed that it would be more efficient to use an oriented rectangle to represent the grasping position of the robot, which has been widely used by researchers. Another core problem of robot grasping is how to extract features. Before the emergence of artificial neural networks, people needed to spend a lot of time designing features manually to solve robot tasks [9-10]. In 2014, the artificial neural network was first introduced by Lenz et al. [11] to solve the problem of robot grasping. At the same time, in the field of computer vision, convolution neural network has proved to be a powerful model for image feature extraction [12-13]. In 2015, Joseph et al. [14] first applied convolutional neural network to robot grasping task. At present, convolution neural network is widely used to solve the problem of robot grasping [15-17].

Due to the low price of the RGB-D sensor on the market, it has been widely used in robot vision tasks, and has also been used by researchers in the research of grasping tasks [18]. It has been proved that using colour data and depth data together in RGB-D sensor has better grasping effect than using colour data alone [19-20]. In addition to using depth image and colour image as 3D data, 3D point cloud
is also a kind of 3D data that can be obtained from RGB-D sensor. However, the development of deep learning based on 3D point cloud is not as fast as that based on image, the successful point cloud models at this stage are PointNet [21] and PointNet ++ [22] proposed by Qi et al. in 2016 and 2017 respectively, they can directly consume the point cloud network, and effectively perform object classification, part segmentation and semantic segmentation.

In recent years, Chu et al. [23] proposed a deep network structure to solve the multi-target grasping problem. Asif et al. [24] proposed hierarchical cascaded forests to infer categories and grasping positions from RGB-D point clouds. Guo et al. [25] proposed a hybrid depth neural network combining tactile and vision for grasping detection. In 2019, NVIDIA company Mousavian et al. [26] used 3D point clouds to extract features and calculate captured 3D positions, and solved the problem of 6-DOF grasping posture, this research is a relatively advanced research so far, but the method of this research needs huge computing power to support.

2. Inception module
Because most of the neural connections in nature are sparse, and human brain can be regarded as the repeated accumulation of sparse neurons, so the Inception series network uses sparse Inception module to solve the visual problems.

2.1. Inception-v1
The Inception network was first proposed from the design of GoogLeNet [27], it is a new deep learning architecture. Similar to VGGNet, they all construct deep neural networks with stronger expression ability by stacking convolution modules. Unlike from VGGNet, the Inception network uses a branch structure composed of multiple small channel convolution layers to replace the convolution layer used in traditional convolution neural network. The design of this structure can effectively reduce the parameters and calculations. The number of parameters of GoogLeNet using the Inception module is about 36 times less than that of VGGNet and about 12 times less than that of AlexNet. As shown in Fig. 1 and Fig. 2, the basic structure and bottleneck structure of Inception-v1 are shown respectively.

Figure.1 Basic structure of Inception-v1.

Figure.2 Bottleneck structure of Inception-v1.

2.2. Inception-Lite
In the calculation process of grasping detection, in order to add and modify the model conveniently, based on the basic design criteria of convolution neural network structure, this paper proposes Inception-Lite, which is a lightweight convolution network module based on the Inception structure. This structure transforms convolution from dense connection structure to sparse connection structure, which reduces the network computation without weakening the network expression ability. In order to improve the overall performance of the network, the convolution module can be stacked to form a larger scale convolution neural network. Similar to other design ideas, the basic version and reduction version module of Inception-Lite are proposed, as shown in Fig. 3 and Fig. 4 respectively.
3. Grasping Network Structure

3.1. Grasping position
If the depth stereo camera is used on the robot, the pixel coordinates can be converted to the robot coordinates through the transformation relationship, and then the manipulator can be guided to move. According to this premise, the problem of grasping the visual object can be transformed into the prediction of the grasping position in the image. Specifically, as shown in Fig. 5, a rectangle can be used in the image to represent the grasping position of the robot gripper, so that the rectangle contains the information of position, direction and opening distance of the gripper device.

![Grasping position](image)

Where, $\theta$ represents the direction of the rectangle relative to the horizontal axis, $(x, y)$ represents the center of the rectangle, $H$ represents the height of the rectangle, $W$ represents the width of the rectangle.

3.2. Grasping Network Structure
With the purpose of completing the grasping detection task, the GRASPNET-CLF network is designed by combining with the Inception-Lite convolution module. The network model takes an image containing the target as the input, and outputs a grasping position for the target after the network calculation. In the process of practical application, the model needs to assume that each image has a target to grasp, and only needs to predict one of the possible positions of the target.
Generally, there are two stages in common grasping network. The first stage needs to judge all possible grasping positions, and the second stage selects the highest score as the final result. The two-level structure needs to design two kinds of neural networks, the first stage is a small-scale neural network, the second stage is a relatively large-scale neural network. Different from the general grasping network, GRASPNET-CLF is an end-to-end grasping network, and the results can be obtained after the network is executed once. The network structure of the GRASPNET-CLF is shown in Fig. 6.

The input of the GRASPNET-CLF network can be any image, such as RGB colour image or three channel depth image. The network regards the grasping detection problem as a regression problem, it needs to take the five parameters of the grasping rectangle as the final prediction result. On the whole, the network is composed of ordinary convolution layer, Lite convolution layer and full connection layer. There are two convolution layers in the ordinary convolution layer, and the kernel size of each convolution layer is 3×3; The convolution module and the reduction module in Lite convolution layer are proposed in Section 2.2. Firstly, it passes through three Lite convolution modules, then through a down sampling module, then through two Lite convolution modules and a down sampling module; Finally, the network uses three full connection layers, and the number of hidden layer nodes is 1024, 256 and 5 respectively. After the whole network calculation is finished, the final prediction of each parameter output is obtained.

4. Experiment

4.1. Dataset
In order to compare the experimental results, Cornell grasping dataset [8] is selected to train and test the network model. The dataset is specially designed for parallel grippers, it provides image data and point cloud data of 240 different grasping targets. There are 885 images in total, and each image is marked with grab rectangle.

4.2. Training
Before the network training, it is necessary to enhance the data processing. The enhancement methods include Gaussian noise, contrast adjustment, random translation and rotation. After the sample enhancement processing, each original image will generate 2500 training samples. The model of
graphics processing unit used in the experiment is NVIDIA GTX1650, the operating system is Ubuntu 16.04, and the deep learning framework is MXNET. The parameters set in training process of grasping network are shown in Table 1.

| Learn rate | Dropout rate | Regularization | Batch size |
|------------|--------------|----------------|------------|
| 5e-4       | 0.6          | 1e-3           | 16         |

Similar to other work, the data is divided into two parts, one is random segmentation based on image split, the other is random segmentation based on object split, and the segmented images are sent to the same batch for training. After 500 iterations, the training stopped. Table 2 shows that the GRASPNET-CLF network performs better in grasping detection.

| Image-wise split detection accuracy | Object-wise split detection accuracy | Detection speed |
|------------------------------------|-------------------------------------|-----------------|
| Fast Search [8] 6.7%               | 6.7%                                | -               |
| SAE [11] 73.9%                    | 75.6%                               | 1350ms          |
| AlexNet MultiGrasp [14] 88.0%     | 87.1%                               | 76ms            |
| GRASPNET-CLF 89.7%                | 88.1%                               | 123ms           |

At the end of the experiment, six result images are selected to display, as shown in Fig. 7 to Fig. 12.

5. Conclusion
We presented a lightweight Inception-Lite module, compared with the traditional convolution module, it can effectively expand the width and depth of the network, and compared with the similar Inception module, it has less computation. We also presented a grasping network GRASPNET-CLF based on the Inception-Lite module, and experiment on the public grasping dataset. In terms of accuracy, the model achieves 89.7% in image-wise split data and 88.1% in object-wise split data; in terms of running speed, the inference speed of the model is more than 8 FPS. The experimental results show that the network has good accuracy and real-time performance in the grasping network. In the future, the network can be combined with multi-modal data and industrial robots to grasp the target in real time.
Acknowledgments

Thanks to Dr. Zhou participated in the concept, design, interpretation and commented on the manuscript and the critical revision of this paper. Thanks to Miss Wang participated in the data collection and analysis of the paper.

References

[1] Grupen R. A. (1991). Planning grasp strategies for multifingered robot hands. In: 1991 IEEE International Conference on Robotics and Automation. pp. 646-647.
[2] Kamon, I., Flash, T., & Edelman, S. (1996). Learning to grasp using visual information. In: Proceedings of IEEE International Conference on Robotics and Automation. Vol. 3, pp. 2470-2476.
[3] Bicchi, A., & Kumar, V. (2000). Robotic grasping and contact: A review. In: Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). Vol. 1, pp. 348-353.
[4] Miller, A. T., Knoop, S., Christensen, H. I., & Allen, P. K. (2003). Automatic grasp planning using shape primitives. In: 2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422). Vol. 2, pp. 1824-1829
[5] Miller, A. T., & Allen, P. K. (2004). Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine, 11(4), 110-122.
[6] Pelossof, R., Miller, A., Allen, P., & Jebara, T. (2004). An SVM learning approach to robotic grasping. In: IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004. Vol. 4, pp. 3512-3518
[7] León, B., Ulbrich, S., Diankov, R., Puche, G., Przybylski, M., Morales, A., ... & Dillmann, R. (2010). OpenGrasp: a toolkit for robot grasping simulation. In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots. Berlin. pp. 109-120.
[8] Jiang, Y., Moseson, S., & Saxena, A. (2011). Efficient grasping from rgbd images: Learning using a new rectangle representation. In: 2011 IEEE International conference on robotics and automation. pp. 3304-3311.
[9] Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., & Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In: 2010 IEEE International Conference on Robotics and Automation. pp. 2308-2315.
[10] Kragic, D., & Christensen, H. I. (2003). Robust visual servoing. The international journal of robotics research, 22(10-11): 923-939.
[11] Lenz, I., Lee, H., & Saxena, A. (2015). Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5): 705-724.
[12] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90.
[13] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580-587.
[14] Redmon, J., & Angelova, A. (2015). Real-time grasp detection using convolutional neural networks. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 1316-1322.
[15] Schmidt, P., Vahrenkamp, N., Wächtmer, M., & Asfour, T. (2018). Grasping of unknown objects using deep convolutional neural networks based on depth images. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 6831-6838.
[16] Zeng, A., Song, S., Yu, K. T., Donlon, E., Hogan, F. R., Bauza, M., ... & Fazeli, N. (2018). Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 1-8.
[17] Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., ... & Goldberg, K. (2017).
Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312.

[18] Saxena, A., Driemeyer, J., & Ng, A. Y. (2008). Robotic grasping of novel objects using vision. The International Journal of Robotics Research, 27(2): 157-173.

[19] Koppula, H. S., Gupta, R., & Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research, 32(8): 951-970.

[20] Saxena, A., Wong, L. L., & Ng, A. Y. (2008). Learning grasp strategies with partial shape information. In: AAAI. Vol. 3, No. 2, pp. 1491-1494.

[21] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652-660.

[22] Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in neural information processing systems. pp. 5099-5108.

[23] Chu, F. J., Xu, R., & Vela, P. A. (2018). Real-world multiobject, multigrasp detection. IEEE Robotics and Automation Letters, 3(4): 3355-3362.

[24] Asif, U., Bennamoun, M., & Sohel, F. A. (2017). RGB-D object recognition and grasp detection using hierarchical cascaded forests. IEEE Transactions on Robotics, 33(3): 547-564.

[25] Guo, D., Sun, F., Liu, H., Kong, T., Fang, B., & Xi, N. (2017, May). A hybrid deep architecture for robotic grasp detection. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 1609-1614.

[26] Mousavian, A., Eppner, C., & Fox, D. (2019). 6-dof graspnet: Variational grasp generation for object manipulation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2901-2910.

[27] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1-9.