Grasping Prediction Algorithm Based on Full Convolutional Neural Network

Kaiyue Feng¹, Wu Wei¹, Qiuda Yu¹*, Qinglun Liu²*
¹ South China University of Technology, Guangzhou, Guangdong, 510000, China
¹*Corresponding author’s e-mail: 201910102738@mail.scut.edu.cn
²*Corresponding author’s e-mail: liuqinglunzd@zstp.cn

Abstract. Robot grasping is a very frontier and important research direction in the field of robotics. In order to solve the problem of the robot's real-time capture, reduce the time of the visual processing, we proposed a two-stage Convolutional Neural Network structure whose design is simple, with less training parameters, improving the efficiency of the visual system. Using rotation and translation to expand the Cornell fetching dataset. The best model at Cornell grasp test set has achieved 88% of forecast accuracy compared with 73% accuracy rate on one stage network. Moreover, our model size is 0.51MB, speed at 30 FPS on GPU inferencing.

1. Introduction
Among all kinds of robots with various functions, a mobile robot that can grasp is undoubtedly the most needed one in the current industrial market. Mobility gives the robot a larger movement space and a wider working space, while grasping makes the robot have more flexible operation ability, which is a concentrated reflection of technology transforming into productivity.

Robot grasping involves the design of robot body structure, motion control, motion planning, visual perception and intelligent prediction. It has been investigated for decades, yielding a multitude of different techniques [1, 2, 3, 4]. It is a composite research direction that integrates multiple disciplines and fields. Recently, deep learning has made amazing achievements in many ways including robotic grasping. These methods have advantages in improving the precision of grasping [5, 6, 7, 11, 12, 13].

However, most of the methods use relatively large network structures, resulting a long running time of the order one second to ten second [8, 9], thus making it hard to do real-time processing. The University of Berkeley laboratory has done a series of capture studies based on Dex-Net. Their methods are based on point cloud images, and the data set is huge [9, 10]. As a result, these methods are based on specific environmental conditions and require cameras with high accuracy and low real-time performance.

We proposed an end-to-end solution to generate the grasp quality of each pixel, meanwhile, find the best gripping point, the angle of the gripper direction and the information of the gripper opening width. We generated a lightweight fully Convolutional Neural Network, borrowing the idea of GGCNN [17, 18], inputting a depth image, outputting the quality, angle, and width information of each pixel, and the information is the same as the input size. With this method, we can accurately obtain the probability of successful pixel-level capture in real time.
2. Related works

Robot grasping means that given any one or more objects, the robot can obtain the object information (color image, depth information, geometric shape, surface texture, etc.) through itself or sensors and other equipment, and get suitable for grasping. The grasping pose includes the three-dimensional position that the fixture or manipulator needs to reach and the width that the fixture or manipulator needs to open when grasping the object after reaching the specified three-dimensional position. In order to grasp more confidently, the fixture or manipulator needs to be based on the vertical axis.

Lee [12] considered the use of deep learning methods to solve the problem of detecting robot grasping in the RGB-D view of the scene containing the object, thereby avoiding the time-consuming manual design function. Pinto [13] spent 700 hours on robotic grasping experiments, and finally collected a data set of 50,000 data points, which increased the available training data to 40 times the previous work. Their work makes it possible to train a Convolutional Neural Network (CNN) to predict the location of grasping points. Redmon [14] proposed a method based on convolutional neural network to accurately obtain real-time robot grasping pose detection. Varley [15] generated a deep learning architecture that can directly detect the position of a stable grasp from a partial view of the grasped object. These technologies all have a common process: grasp candidate objects sampled from images or point clouds are classified, and then they are sorted individually using CNN.

In order to solve the problem of 6-DOF object posture display, Tremblay [16] from the research team of NVIDIA proposed that the reality gap can be successfully eliminated through a simple combination of region randomization and realism data. Using synthetic data generated in this way, a one-off deep neural network is introduced that can compete with the latest networks that combine well-trained real data and synthetic data.

3. Grasping description

In the field of computer vision, semantic segmentation integrates target classification, image detection, and target segmentation. By extracting the characteristics of different objects in the image, various objects are distinguished, and they are labelled by different color regions. Based on the above considerations, we can reduce training parameters, improve computing efficiency, and achieve pixel-
level grasping pose prediction, using a CNN structure with fewer layers, so as to achieve real-time rapid generation of the best grasping pose effect.

3.1. The prediction algorithm structure of grasp

The network structure of the grasp prediction algorithm in this article adopts the "encoder-decoder" structure, which mainly includes a top-down path as an encoder and a bottom-up decoder. It adopted the classic full convolutional neural network idea, using a deconvolution layer instead of a fully connected layer to achieve higher computational efficiency. However, if an unequal architecture of the convolutional layer and the deconvolutional layer is adopted, although the resolution of the picture can be restored to a certain extent, it cannot be guaranteed that the position information of the features in the original picture will not change. Therefore, in order to ensure that the image resolution is restored while still retaining the position information of the original image features, a symmetric encoding-decoding network structure based on a full convolutional neural network is adopted. The number and parameters of the convolutional layer are consistent with those of the deconvolutional layer. In addition, such a structure is not sufficient to predict optimal fetching, so we decided to use two similar structures.

3.2. Datasets

Cornell grasping data set [19] is a grasping data set with artificially labelled grasping frames made by Cornell University for research on visual grasping. Its greatest contribution is to provide each image with a feasible grasp frame annotation. The grasp frame is parallel to the plane where the object is located, and the images in the data set are taken by real cameras and contain noise, which can be trained without adding environmental noise. The data set contains 280 common objects in daily life, each of which contains 4 to 7 pictures, for a total of 1035 pictures. Meanwhile, since the number of images in the data set is a little bit small, we use the method of data set enhancement to obtain a higher amount of data, which includes cropping, flipping, zooming in and zooming out.
3.3. Training

The training in this paper is carried out on the NVIDIA server. The server graphics card group consists of 4 pieces of 1080ti, with a total of 11G of video memory. The server system is Ubuntu16.04, and the programming language version is python3.6. We use 80% of the Cornell graspbed data set as the training set, and use the remaining 20% as the test training set. In this paper, the rectangle measurement method in [19] is used to evaluate the accuracy of prediction, and the IOU criterion is used. IOU translates as intersection and ration, describing the effect of overlapping between the predicted rectangle and the real rectangle. Assuming that the predicted rectangle is A and the real rectangle is B, the IOU between the two is:

\[ IOU = \frac{A \cap B}{A \cup B} \]

The actual evaluation index is that when the angle between the predicted rectangle and the real rectangle is less than 30°, and the IOU between them is greater than 25%, then the predicted grasping rectangle is judged to be an accurate grasp, otherwise it is an incorrect grasp. In addition, this paper chooses the mean square error loss function as another indicator of training performance.

The figure shows that the best accuracy value was 88% at 32th epoch, while the original GGCNN got 73% accuracy. It is obvious that adding a stage can achieve better results. After training, a lightweight model is obtained, which is just 0.5 MB and can reach 30 FPS when running on TX2, thus the real-time performance is guaranteed.
4. Conclusion
In this paper, we proposed a two-stage convolutional neural network to predict the best grasp through depth image. The network gets a high accuracy rate 88% at a speed of 30 FPS. Furthermore, our model size is 0.51MB, which can be easily ported to embedded devices. Besides, our future plan is to transfer the results to a complex environment for real grasping. Through real grasping, we can adjust our algorithm to get better performance.

Acknowledgments
This work is supported by International Industrial Technology Innovation Alliance of Artificial Intelligence and Digital Manufacturing, Guangdong Vocational Colleges Production-education integration innovation Platform Project: Zhongshan Torch Artificial intelligence and robot application production-education integration innovation platform and Typical Manufacturing Line Fault Prediction and Health Service System Project based on Information Physical System.

Reference
[1] Jae-Han Park, Ji-Hun Bae, Yong-Deuk Shin, Sung-Woo Park and Moon-Hong Baeg, "Framework of grasping planning for multi-fingered robot hands," 2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Incheon, 2011, pp. 455-458, doi: 10.1109/URAI.2011.6145863.
[2] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-Driven Grasp Synthesis – A Survey. IEEE Transactions on Robotics, 30(2):289–309, 2014.
[3] Anis Sahbani, Sahar El-Khoury, and Philippe Bidaud. An overview of 3D object grasp synthesis algorithms. Robotics and Autonomous Systems, 60(3):326–336, 2012.
[4] Karun B Shimoga. Robot Grasp Synthesis Algorithms: A Survey. The International Journal of Robotics Research (IJRR), 15(3):230–266, 1996.
[5] Redmon J, Angelova A. Real-Time Grasp Detection Using Convolutional Neural Networks[J]. Proceedings - IEEE International Conference on Robotics and Automation, 2014, 2015.
[6] Varley J, Weisz J, Weiss J, et al. Generating multi-fingered robotic grasps via deep learning[C]// 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015
[7] Mahler, Jeffrey, Matl, Matthew, Liu, Xinyu. Dex-Net 3.0: Computing Robust Robot Vacuum Suction Grasp Targets in Point Clouds using a New Analytic Model and Deep Learning[J].
[8] Lenz I, Lee H and Saxena A (2015) Deep learning for detecting robotic grasps. The International Journal of Robotics Research 34(4–5): 705–724.
[9] Mahler J, Liang J, Niyaz S, et al. (2017) Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In: Proceedings of Robotics: Science and Systems (RSS).
[10] Mahler J, Pokorny FT, Hou B, et al. (2016) Dex-Net 1.0: A cloud based network of 3D objects for robust grasp planning using a Multi-Armed Bandit model with correlated rewards. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), pp. 1957–1964.
[11] He Z, Wu G, Zhang S, et al. Moment-based 2 1/2D visual servoing for texture-less planar part grasping[J]. IEEE Transactions on Industrial Electronics, 2018:1-1.
[12] Lee, Honglak, Saxena, et al. Deep learning for detecting robotic grasps[J]. International Journal of Robotics Research, 2015
[13] Pinto L, Gupta A. Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours[J]. 2015.
[14] Redmon J, Angelova A. Real-Time Grasp Detection Using Convolutional Neural Networks[J]. Proceedings - IEEE International Conference on Robotics and Automation, 2014, 2015.
[15] Varley J, Weisz J, Weiss J, et al. Generating multi-fingered robotic grasps via deep learning[C]// 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2015.

[16] Tremblay J, To T, Sundaralingam B, et al. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects[J]. 2018.

[17] Morrison D, Corke P, Leitner J. Learning robust, real-time, reactive robotic grasping[J]. The International journal of robotics research, 2020, 39(2/3):183-201.

[18] Morrison D, Corke P, Leitner, Jürgen. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach[J]. 2018.

[19] Z. Cao, G. Hidalgo Martinez, T. Simon, S.-E. Wei and Y. A. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, doi: 10.1109/TPAMI.2019.2929257.