Proc. of the 9th Intl. Conf. of Asian Society for Precision Engg. and Nanotechnology (ASPEN 2022)
15–18 November 2022, Singapore. Edited by Nai Mui Ling Sharon and A. Senthil Kumar
doi://10.3850/978-981-18-6021-8_OR-02-0217

Intelligent 6-DoF Robotic Grasping and Manipulation System Using Deep Learning

You-Rui Chu[1], Haiyue Zhu[2#] and Zhiping Lin[1]

1 School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Ave., Singapore 639798
2 Adaptive Robotics and Mechatronics Group, Singapore Institute of Manufacturing Technology (SIMTech), 2 Fusionopolis Way, Singapore 138634
Corresponding Author / Email: zhu_haiyue@simtech.a-star.edu.sg

KEYWORDS: Deep Learning, Robotics, Grasping

Random object grasping in unstructured environment is a crucial problem in robotics which is yet to be solved but highly demanded. In this paper, we focus on the prediction of 6-DoF grasp poses using end-to-end deep learning approach based on RGB-D images. Most of the current approaches for 6-DoF grasp are generated from point clouds or unstable depth images, which may lead to undesirable results in some cases. The proposed method divides the 6-DoF grasp detection into three sub-stages. The first stage is the LocNet, a convolutional-based encoder-decoder neural network to predict the location of the objects in the image. Besides, ViewAngleNet is also a convolutional-based encoder-decoder neural network that predicts the 3D rotation groups of the gripper at the image location of the objects, similar to LocNet but with a different output head. Afterwards, a feasible grasp search algorithm will determine the gripper's opening width and the gripper's distance from the grasp point. Real-world experiments are conducted with a UR10 robot arm, an Intel Realsense camera and a Robotiq two-finger gripper on single-object scenes and cluttered scenes, which show satisfactory success rates.

NOMENCLATURE

g = 6-DoF grasp pose with 6 parameters, (x, y, z, r_x, r_y, r_z)
x, y, z = 3D position of grasp pose
r_x, r_y, r_z = 3D rotation of grasp pose
w = opening width of the two-finger gripper
V = number of approaching vectors (views)
A = number of in-plane rotations (angles)

1. Introduction

Grasping objects has always been an instinctive behaviour for humans but is a relatively difficult task for robots. The robot's grasping ability is not as good as that of a human, especially when it comes to novel objects. Hence, object grasping is a challenge in the field of robotics. With more solutions for this, applications of robotic grasping can be implemented in various areas such as agriculture, manufacturing and assembly.

Robotic grasping requires visual perception, motion planning and control [1-3]. The first step would be for a robot to recognise and locate the object. Furthermore, pose estimation is also an important step in robotic grasping. Subsequently, the detection of all grasp candidates facilitates the planning of robot movement paths and orientation. However, the reliability of object grasping is still a concern for objects in unstructured environments and unseen objects [4].

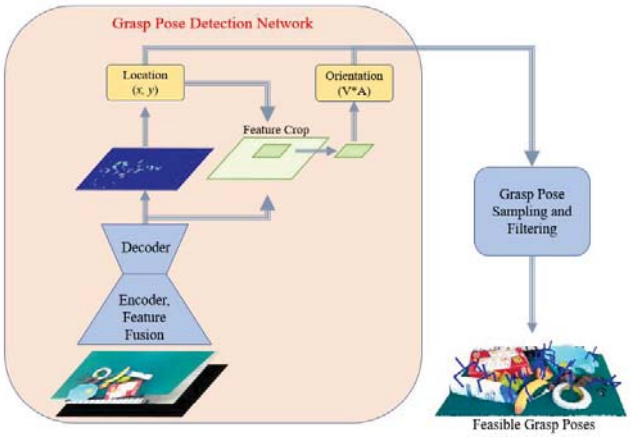

Fig. 1 Overview of Grasp Pose Detection

Traditionally, physics analysis methods [5, 6] are used to determine appropriate grasp poses. However, the object models required may not always be available, and applying them to unseen objects can be difficult. Furthermore, these procedures are typically time-consuming and computationally costly.

In order to achieve autonomous robotic manipulation, the trend

has shifted towards the use of deep learning [7, 8]. The prediction of grasp detection with deep learning is based on computer vision and artificial intelligence. The detection of 2D rectangular grasps for objects with RGB-D input data is the initial approach investigated, and it is also the most common. Some datasets [9, 10] have been published, and numerous techniques [9, 11, 12] have been developed to produce 2D planar grasps on those dataset. However, the 2D rectangular-based grasp has some limitations in terms of the grasp pose. The gripper can only approach the object from the vertical direction and it would not be optimal for objects that are lying horizontally on the surface.

Recently, more studies have been published on the detection of 6-DoF grasp poses. The estimation of 6D poses [13, 14] allows the prediction of 6-DoF poses for seen objects but this cannot be generalised to unseen objects. Another sampling and evaluation approach is used by GPD and PointNetGPD in a two-step process. However, the unsatisfied sampling findings result in the examination of a large number of samples to find the accurate grasp poses. This procedure is not time-efficient. Besides, grasp poses can also be transferred [15, 16] from existing objects to others. However, if the objects are unseen and the geometries are different from the existing ones, this approach will fail. Furthermore, 6-DoF grasp poses can be generated by passing the partial view point cloud, obtained from RGB-D images, through the networks [17, 18]. But, due to the possibility of sensor failure, depth image data is less reliable than RGB image data.

Therefore, this paper aims to explore the inference of the 6-DoF grasping pose for random objects using RGB-D image data. A novel deep network framework is proposed for grasp pose detection as shown in Figure 1. The input for the proposed framework is a RGB-D image and 6-DoF grasp poses are produced from the framework as the output. The problem is divided into three subtasks. Both the RGB image and the depth image are used to generate a heatmap indicating feasible object grasp locations. This is achieved by passing the RGB and depth images through an encoder-decoder like network, LocNet. Similarly, the rotation matrix of the gripper for each grasp location is produced by the same RGB-D input and an encoder-decoder like network, ViewAngleNet, together with the feasible grasp locations. Afterwards, a feasible grasp search is carried out to obtain the opening width of the gripper and the distance to the grasp position. The search is made up of sampling and filtering. The grasp pose samples are calculated from the grasp locations and rotation matrices. The filtering consists of collision detection and empty-grasp detection between the gripper and the scene. Given the deep neural network and the big data from the GraspNet-1Billion dataset [19], the approach may not just apply to seen objects but also generalise to unseen objects. The performance of the developed framework will be simulated on a UR10 collaborative robot arm.

2. Method

The proposed grasp pose detection comprises three parts which work together in a pipeline as illustrated earlier. The first two parts of the pipeline, LocNet and ViewAngleNet, form the grasp pose detection network.

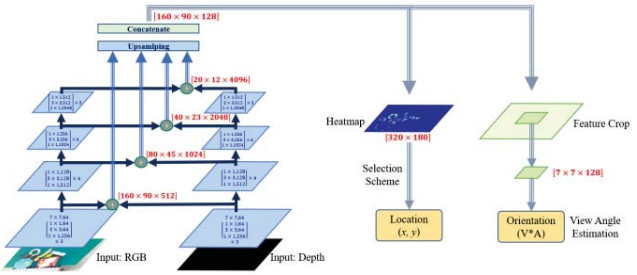

Fig. 2 Architecture of Grasp Pose Detection Network

Both LocNet and ViewAngleNet share the same backbone network which includes the encoder-decoder based convolutional neural network with Feature Pyramid Network (FPN) structure as seen in Figure 2. For the encoder-decoder structure, an encoder extracts features from input images through a series of convolutions, while the decoder semantically transfers the discriminative features onto the pixel space. The baseline encoder employed for image processing is ResNet50 due to its deep architecture and the ability to solve complex problems. FPN is also applied to better use the RGB-D image information for feature extraction through the fusion approach.

2.1 LocNet: Grasp Location Heatmap

LocNet is the first component of the grasp pose detection network which generates the feasible grasp location heatmap. The heatmap is then used to determine the optimal grasp location (x, y) at pixel level, where higher intensity represents a higher probability of grasping. LocNet consists of the backbone network with a classification output head as it is treated as a binary classification problem for each pixel (x, y) in the image. In practice, every annotated grasp location is considered as a small target circular area, because it is impossible to acquire the exact ground truth for all the possible grasp locations. The advantage of employing this encoder-decoder structure in the backbone network is that it is a generative detection method, even though the target map is discrete in each annotation. It can extrapolate from the sparse grasping location annotations and attempt to map out all conceivable grasp locations in a dense and continuous way. The grasp location samples are selected from the predicted heatmap by setting a threshold. After filtering the heatmap, the location samples that are above the threshold will be regarded as the feasible grasp locations. The samples can be further selected through the Non-maximum suppression (NMS) process to finalise the grasp location.

2.2 ViewAngleNet: Gripper Orientation Estimation

ViewAngleNet is the other component of the grasp pose detection network that predicts the pixel level-based gripper orientation. Directly regressing the rotation matrix is one simple approach to tackling this problem. However, there are multiple possible rotations that can achieve successful grasping at the same location. Thus, it would not be suitable to use regression. As a result, with reference to GraspNet-1Billion [19], the gripper rotation matrix is split into two parts: the approaching vector (view) and the in-plane rotation (angle).

Since multiple rotations can be feasible in the same location, this

problem is treated as a multi-label classification problem. There are a total of V A classes. For every possible grasp location, ViewAnlgeNet predicts the confidence scores for each of the V A classes independently and outputs them in the 1-dimensional vector format with the length of V A. Similar to LocNet, ViewAngleNet also contains a backbone network with a classification output head but the classification head in this case is for multi-label classification.

2.3 Feasible Grasp Search

After obtaining the grasp location and the gripper orientation from LocNet and ViewAngleNet respectively, the feasible grasp search is applied to determine the gripper opening width and the distance from the gripper to the image plane.

The main idea is to sample the grasp candidates first, followed by filtering out the non-viable grasp poses. This is under the assumption that the gripper point should be as near to the partial-view point cloud rebuilt from the RGB-D image as possible. In order to estimate the width, different widths are sampled given the grasp location and gripper orientation. Similarly, with the grasp location and gripper orientation obtained previously, distances are sampled in a uniform manner from the position above the gripper point to the position below the gripper point. The gripper point can be computed from the grasp location and depth in the depth image. A group of grasp pose samples can be generated with the gripper point, grasp rotation matrix, and various widths and distances. Subsequently, the filtering involves collision and empty-grasp detections. For collision detection, the sample grasp poses will be filtered out if there are points contained in the gripper space. For empty-grasp detection, the sample grasp poses will also be filtered out if there are no points in between the grasping space of the parallel jaw gripper. Afterwards, grasp pose non-maximum suppression (GPNMS) [19] is conducted on the remaining grasps to eliminate the duplicates and find the optimal grasp poses.

4. Implementation Details

4.1 Dataset and Preprocessing

The data required for training involves the RGB-D image with the corresponding ground truth for the heatmap, as well as view vector and angle. The generated Cornell dataset [9] and the Jacquard dataset [10] are first looked into. They provide the RGB-D image with ground truth for the 2D planar grasp. However, the grasp poses in these datasets are all approaching the objects in the vertical direction only and are not defined in various directions.

Hence, the dataset is obtained from GraspNet-1Billion [19], which is the largest known 6-DoF grasping dataset available to the public. It includes richly annotated grasp poses in complex scenes captured by two commercial RGB-D cameras (Kinect Azure and RealSense D435), which helps to generate the ground truth.

Due to the enormous amount of data available, Farthest Point Sampling (FPS) [20] method is adopted to reduce the training data sample. FPS is carried out to obtain evenly distributed data points from the dense data sample. 5% of the total grasp locations are sampled which results in a total of 10 million grasp locations and 191 million grasp pose labels. Therefore, each image contains roughly 195 grasp locations and each location has around 19 grasp poses.

As mentioned previously, the gripper orientation is decoupled into view and angle. Furthermore, from the gripper rotation matrix in the dataset, the nearest view and angle are computed and combined as the ground truth label for the gripper orientation. Only the grasp orientations for the corresponding grasp locations are considered. Besides, the radius for circular area centering at the grasp location (x, y) is set to 4.5 pixels.

During feasible grasp search, the gripper opening widths are sampled from the range of 0.01m to 0.1m with a step size of 0.01m The distance to the image plane from the gripper point is selected from 0.02m above to 0.02m below with a step size of 0.01m. For each predicted grasp location, the closest neighbouring grasp poses are produced from the various widths and distances to form the sample group for filtering.

4.2 Network and Training

During training, V=100 views and A=4 angles are sampled evenly, which generates a product of V A=400 classes of combined view and angle. The network is trained on four Nvidia GeForce RTX 2080 Ti GPUs with a batch size of 16. With the ADAM optimiser, the learning rate is initialised to 0.001. The learning rate applied is the step learning rate which will decrease the learning rate by a factor of 10 for every 20 epochs. ResNet pretrained weights are used to start the encoder section of the network. Data augmentation is employed during training by performing color jittering to prevent overfitting.

5. Experiments

The experiment setup includes an UR10 robot arm, an Intel Realsense L515 camera and a Robotiq 2F-140 gripper.

5.1 Real Robot Experiment on Single Object Scenes

Random objects are chosen and placed on the table in a random order. The robustness performance of the model is evaluated in this experiment. There are 10 attempts carried out on each individual object and the success rate is computed.

Table 1 Performance Of Real Robot Experiment On Single Object Scenes

Object ID	Object Name	Type	Attempt	Success	Success Rate
1	Cardboard Box	Novel	10	7	70%
2	Tripod	Novel	10	6	60%
3	Duct Tape	Novel	10	8	80%
4	Plastic Vegetable 1	Novel	10	7	70%
5	Screwdriver	Novel	10	7	70%
6	Plastic Vegetable 2	Novel	10	7	70%
	Average	-	60	42	70%

5.2 Real Robot Experiment on Cluttered Scenes

Similarly, the real robot experiment is conducted in the cluttered scenes. Multiple unseen objects from the previous experiment are laid

out on the table randomly to form the cluttered scene. The test continues until all the objects in the scene are picked and placed in a designated position.

Table 2 Performance Of Real Robot Experiment On Cluttered Scenes

Objects ID	Attempt	Success	Success Rate
1, 2, 4	3	4	75%
2, 3, 5	3	4	75%
1, 2, 3, 4, 5	5	7	71.43%
1, 2, 3, 4, 5, 6	6	8	75%
Average	17	23	73.91%

3. Conclusions

A 6-DoF grasp pose detection pipeline is proposed in this paper. It divides the grasp pose detection problem into three sub-problems: grasp location detection; gripper orientation detection; gripper opening width and optimal gripper point detection. To address the associated subproblems, the pipeline proposes using LocNet, ViewAngleNet, and feasible grasp search. The presented procedure has introduced some novelty to the methodology. This offers a fresh perspective on the object grasping problem, which might be investigated further in the future. Although the findings do not indicate a high level of accuracy, they do show that the proposed network is capable of predicting viable grasp location heatmaps and, to a certain extent, gripper rotation matrices. Both experiments result in desirable success rates, proving that the proposed grasp pose detection is still fairly successful.

REFERENCES

[1] H. Karaoguz and P. Jensfelt, "Object detection approach for robot grasp detection," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019: IEEE, pp. 4953-4959.

[2] Y. Li *et al.*, "Few-shot object detection via classification refinement and distractor retreatment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15395-15403.

[3] H. Zhu *et al.*, "Weight imprinting classification-based force grasping with a variable-stiffness robotic gripper," *IEEE Transactions on Automation Science and Engineering,* vol. 19, no. 2, pp. 969-981, 2021.

[4] A. Sintov and A. Shapiro, "Dynamic regrasping by in-hand orienting of grasped objects using non-dexterous robotic grippers," *Robotics and computer-integrated manufacturing,* vol. 50, pp. 114-131, 2018.

[5] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, 2000, vol. 1: IEEE, pp. 348-353.

[6] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on robotics,* vol. 30, no. 2, pp. 289-309, 2013.

[7] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A Survey on Learning-Based Robotic Grasping," *Current Robotics Reports,* vol. 1, no. 4, pp. 239-249, 2020/12/01 2020, doi: 10.1007/s43154-020-00021-6.

[8] Y. Li *et al.*, "Incremental Few-Shot Object Detection for Robotics," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022: IEEE, pp. 8447-8453.

[9] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International conference on robotics and automation*, 2011: IEEE, pp. 3304-3311.

[10] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018: IEEE, pp. 3511-3516.

[11] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research,* vol. 34, no. 4-5, pp. 705-724, 2015.

[12] H. Zhu *et al.*, "Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020: IEEE, pp. 9608-9613.

[13] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review,* vol. 54, no. 3, pp. 1677-1734, 2021.

[14] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11632-11641.

[15] T. Patten, K. Park, and M. Vincze, "Dgcm-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping," *Frontiers in Robotics and AI,* p. 120, 2020.

[16] H. Tian, C. Wang, D. Manocha, and X. Zhang, "Transferring grasp configurations using active learning and local replanning, 2018," *arXiv preprint arXiv:1807.08341,* 1807.

[17] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901-2910.

[18] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: learning an end-to-end spatial grasp generation algorithm from sparse point clouds," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020: IEEE, pp. 3619-3625.

[19] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11444-11453.

[20] B. Sébastien, P. Gabriel, and D. C. Laurent, *Image Compression with Anisotropic Geodesic Triangulations*: SIAM Imaging Science 2010 2010, p. 29. [Online]. Available: https://bougleux.users.greyc.fr/articles/bougleux10siam.pdf .