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-02-0186

Segmentation for Grasping: An Approach toward Autonomous Table Clearing

Ka-Shing Chung[1,#], Marcelo H Ang Jr[1], Wei Lin[2] , Haiyue Zhu[2] , Joel Short[2] and Pey Yuen Tao[2]

1 Department of Mechanical Engineering, National University of Singapore, 9 Engineering Drive 1, Singapore 117575, Singapore
2 SIMTech, A*STAR, 2 Fusionopolis Way, Singapore 138634, Singapore
Corresponding Author / Email: chung@u.nus.edu

KEYWORDS: Robotic grasping, Instance Segmentation

Abstract -- Autonomous clearing of food trays and crockery at hawker centres involve robotic tasks such as item recognition, grasp point estimation and grasp execution. Assuming there are no dynamic obstacles in the manipulator workspace, we focus on the machine intelligence required for the first two tasks above. Our problem statement is as follows: Given a RGBD view of a cluttered scene consisting of one or more known objects at rest on a flat surface, we seek a way to determine a feasible grasp pose for each of these objects.

Recent approaches treat the whole pipeline as one black box and try to predict grasp poses directly from the RBGD input. However, the pipeline is intricate and contain many sub-tasks that could be explored with greater complexity. The related work that are successful in the end-to-end training, on the other hand, will merely output highest ranked grasp poses of any reachable object, without any semantic concept of the item being picked up. This lack of scene understanding would eventually inhibit the optimization of the grasping algorithms as there is no way to willfully select a particular object for manipulation.

In our work, we break down the grasp pose determination into several components and focus on solving them individually. First, we parse the input scene by passing it through a convolutional neural network trained for instance segmentation. The network outputs an image and depth mask for each object that has been detected in the scene, as well as the object class. We assume that we have a database of known objects. Next, we use the object masks to project a partial point cloud, which is registered to a complete point cloud of the corresponding object in our library. A transformation is needed to align the two point clouds. Finally, we apply the same transformation to the grasp pose from the library to produce a grasp pose for the object as seen in the initial scene. Our approach allows the user to select the object to be grasped, and also lays the groundwork for an automated object selection strategy in the future.

NOMENCLATURE

P_{obs} = homogenous coordinates of a point in the object point cloud as observed by the robot camera
P_{CAD} = homogenous coordinates of a point in the complete object point cloud as projected from its CAD model
T = 4x4 transformation matrix from library to scene

1. Introduction

Table clearing at food and beverage establishments is a research problem that has gathered some attention in recent years. It is a natural extension of smart delivery robots, which navigate their way through a crowded restaurant to deliver food to diners [10], but rely on diners to manually place the empty trays and used dishes back onto the robot for the returning journey. Adding table clearing capabilities will allow the robot to collect these dishes from the table autonomously, with the use of one or more manipulator arms mounted on the mobile platform.

The robotic capabilities can be broken down into several main tasks [12, 18]. In order to collect the empty dishes and cutlery, the robot will first need to identify and localize them within the table. Next, one of the identified objects will be selected for grasp planning, which includes determining suitable grasping points on the object

surface and planning a trajectory for the arm to move into position. Finally, the fingers of the gripper will be closed and the object should be secured with no relative motion to the fingers when the manipulator arm is raised from the table. We assume the robot has a head mounted RGBD sensor and executes this grasp without moving its base.

2. Our method

The overview of our approach is as follows: given the aligned RGBD images of the scene, we perform image segmentation to crop out an object that we are interested in grasping, then register its pose in the scene to a stored CAD model of that object, and finally apply the computed 6DOF transformation to a known grasp pose to obtain a feasible grasp pose in the scene. This pipeline will be repeated for subsequent objects if the scene is cluttered.

2.1 Segmentation

Related work on 6DOF grasping [1-4, 13-16] directly generated grasp proposals with a deep neural network from a raw depth image or point cloud input. Most of the recent work filter feasible contact points from the RGBD image or point cloud and infer the most confident grasp pose from the filtered set. [3] separates the modalities by using RGB to find approach angles and using depth to obtain the gripper width. In these approaches, the cluttered scene is not interpreted semantically and grasp points are proposed whereever visible to the sensor and feasible for approach. We aim to segment the image before determining the grasp pose, to aid in future pipelines that may employ object selection strategies to optimize the table clearing. Secondly, knowing the object that we are targeting also gives us better information on the object geometry and hence propose stable grasp points.

In our work we assume a known CAD model of each object and only consider objects that are in our library. As there is no specialized dining dataset, we use the Graspnet-1Billion dataset [5] as it contains daily use objects in a cluttered scene as well as the 6DOF pose required to grasp these objects.

For segmentation, we apply a Mask R-CNN [5] network with a ResNet-50 backbone to classify and segment the objects in the given scene. A model pre-trained on the MS COCO dataset [17] is used for the segmentation without further training on our dataset. We filter the object classes to only keep those classes in our interest, such as bowls, cups, plates, etc, and project a point cloud from the masked depth and RGB images. The result is seen in Figure 1.

2.2 Pose Estimation

The projected point cloud of the object is incomplete and suffers from occlusion due to itself as well as any surrounding object which may be overlapping or blocking the camera view. Hence, the object class predicted by Mask R-CNN is vital for accurate identification of the object, so that the complete point cloud can be retrieved from an offline database. With a partial point cloud (scene) and the complete point cloud (target), we then apply pose registration to align them.

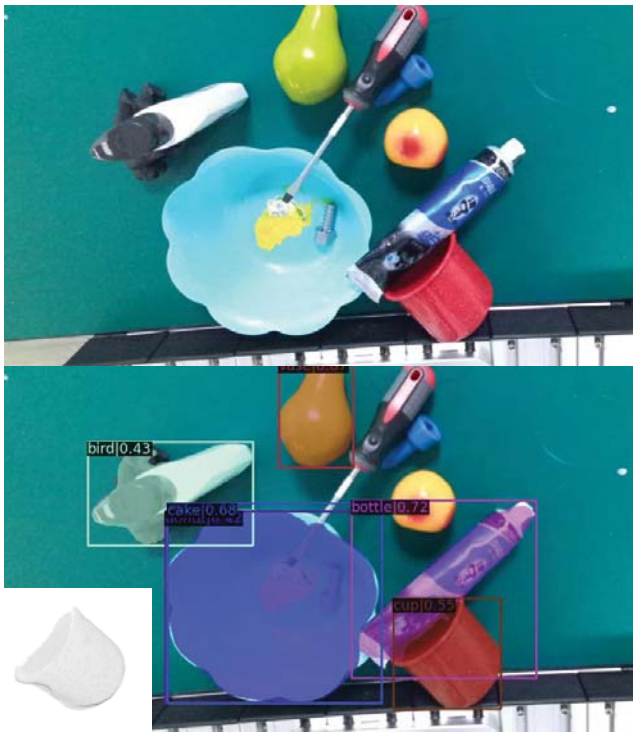

Fig. 1 The original RGB image (above) and the result of the Mask R-CNN segmentation (below) with bounding boxes, object masks, and object class. We are only interested in cups, bottles, fork, spoon, and trays, so other object classes such as birds are filtered afterwards. A partial point cloud (inset, bottom left) is then re-projected from the image and depth mask.

Fig. 2 One grasp pose of the cup as predicted by our pipeline (above) and some ground truth grasp poses visualized from our grasping library (below).

We apply RANSAC for global registration [6, 8], using fast point feature histograms for correspondence matching [9], followed by ICP for fine registration [7] to obtain a 4x4 transformation matrix T, such that $P_{obs} = TP_{CAD}$. Thus, $\hat{g} = Tg$, where $\hat{g}$ and g represent the

coordinate frame attached to the robot gripper in the camera frame and library frame respectively. In our experiments, we used the Open3D implementation [19] of RANSAC and ICP and a setting of 100000 iterations. A sample of the results is seen in Figure 2, where the blue gripper is visualized after applying transformation results.

3. Conclusion

In our work, we demonstrate a new pipeline for computing grasps by piecing together image segmentation algorithms with pose registration algorithms. We test the method on selected scenes in the Graspnet-1Billion dataset without further modification of the network in [5] or ICP registration in [7], although we note that classification accuracy can be further improved with extended training on the dataset. Our approach differs from the literature in that objects in a cluttered scene are first isolated before the grasp pose is determined. Although less direct than training an end-to-end neural network, our method generates semantic information about the objects present in the scene. We plan to exploit this data in future works by incorporating an object selection strategy in the pipeline, so that we may carry out table clearing or bin picking in a more optimal manner rather than simply picking objects from the "top of the pile".

REFERENCES

1. Sundermeyer, M., Mousavian, A., Triebel, R. and Fox, D., "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes," IEEE International Conference on Robotics and Automation (ICRA)., 2021.
2. ten Pas, A., Gualtieri, M., Saenko, K. and Platt, R., "Grasp Pose Detection in Point Clouds," International Journal of Robotics Research, 2017.
3. Minghao G., Hao-Shu F., Zhanda Z., Sheng X., Chenxi W. and Cewu L., "RGB Matters: Learning 7-DoF Grasp Poses on Monocular RGBD Images," IEEE International Conference on Robotics and Automation (ICRA)., 2021.
4. Hao-Shu F., Chenxi W., Minghao G. and Cewu L., "GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
5. He, K., Gkioxari, G., Dollár, P., Girshick, R., "Mask R-CNN," IEEE International Conference on Computer Vision (ICCV), 2017.
6. Li, J., Hu, Q. and Ai, M., "Point Cloud Registration Based on 1-point RANSAC and Scale-annealing Biweight Estimation," IEEE Trans. on Geos. and Rem. Sens., 2021.
7. Besl, P., McKay, N., "A Method for Registration of 3D Shapes," IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)., 1992.
8. Choi, S., Zhou, Q. Y. and Koltun, V., "Robust Reconstruction of Indoor Scenes," Computer Vision and Pattern Recognition Conference (CVPR), 2015.
9. Rusu, R. B., Blodow, N. and Beetz, M., "Fast Point Feature Histograms (FPFH) for 3D registration," IEEE International Conference on Robotics and Automation (ICRA)., 2009.
10. Proven Robotics, "How robotics is changing the dynamics of the food industry?", Available online: https://provenrobotics.ai/tpost/0xahc8avn1-how-robotics-is-changing-the-dynamics-of (accessed on 11 August 2022).
11. Klank, U., Pangercic, D., Rusu, R.B. and Beetz, M., "Real-time CAD Model Matching for Mobile Manipulation and Grasping", 9th IEEE-RAS International Conference on Humanoid Robots, Paris, France, pp. 290-296, 2009.
12. Zhu, H., Li, Y., Bai, F., Chen, W., Li, X., Ma, J., Teo, C. S., Tao, P. Y. and Lin, W., "Grasping Detection Network with Uncertainty Estimation for Confidence-Driven Semi-Supervised Domain Adaptation," International Conference on Intelligent Robots and Systems., 2020.
13. Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J. A. and Goldberg, K., "Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics," Robotics: Science and Systems (RSS)., 2017.
14. Yu, S., Zhai, D., Xia, Y., Wu, H. and Liao, J., "SE-ResUNet: A Novel Robotic Grasp Detection Method," IEEE Robotics and Automation Letters., 2022.
15. Paul, S. K., Chowdhury, M. T., Nicolescu, M., Nicolescu, M. and Feil-Seifer, D., "Object Detection and Pose Estimation from RGB and Depth Data for Real-time, Adaptive Robotic Grasping," Computing Research Repository (CoRR)., 2021.
16. Liang, H., Ma, X., Li, S., Görner, M., Tang, S., Fang, B., Sun, F. and Zhang, J., "PointNetGPD: Detecting Grasp Configurations from Point Sets," IEEE International Conference on Robotics and Automation (ICRA)., 2019.
17. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L., "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision (ECCV)., 2020.
18. Eppner, C., Hofer, S., Jonschkowski, R., Martin-Martin, R., Sieverling, A., Wall, V. and Brock, O., "Lessons from the Amazon Picking Challenge: Four Aspects of Building Robotic Systems," Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI), 2017.
19. Zhou, Q.-Y., Park, J. and Koltun, V., "Open3D: A Modern Library for 3D Data Processing", Computing Research Repository (CoRR)., 2018.