Human-like Object Manipulation Based on Object Affordance Detection and 3D Shape Analysis for Social Robot

Jeongho Lee, Myoungha Song, SeungHyun Kang,
Changjoo Nam, Changhwan Kim and Dong Hwan Kim

Abstract— **With the development of modern vision technology using deep learning, robots can detect objects in various ways such as classification, detection, segmentation and so on. In addition, the methods in which the robot grasps objects are becoming more detailed and accurate. In this paper, we propose a human-like object manipulation based on object affordance detection and 3D shape analysis on the detected grasp affordance area. The grasp synthesis provides several grasp information on an object. The grasp information involves how to use the object. This information, which properly identifies the characteristics of the object, is suitable for social robots. In experiments, affordance detection module achieves about 88% accuracy on custom dataset. We verified that our grasp synthesis using affordance is useful and Jaco robot arm grasped an object using generated grasp information.**

I. Introduction

Modern vision technology has been making tremendous progress through deep learning. Classification, detection, and segmentation have been greatly performed, furthermore networks have been developed such as instance segmentation, panoptic segmentation, and scene understanding.

Due to the development of this vision, many changes have been taking place in robotics. In particular, in the field of grasp detection, there were studies in deep learning [2, 3, 4, 5] helped to generate grasp information in real-time.

However, the above approach mainly provides only information for grasping the object. For example, the mug generates grasp information only on the handle, and the bowl generates grasp information for grasping the inner and outer parts. When holding the mug, the person holds the handle, but sometimes hold the body part only, and the bowl rarely is used the inner and outer parts. In addition, because those studies did not consider the characteristics of objects by focusing on grasping objects, they are insufficient on additional motions after grasping the object. Thus they are not suitable for social robots.

To solve these problems, we applied the concept of affordance to the method of grasping objects. The affordance of an object is aligned for each part of the object and is defined as general actions by which people can act on the part of the object. For example, if you aligned a knife affordance, you can define the handle part as grasp and the blade part as cut. The affordance that is applied to network [6] detects the object as a bounding box and the affordance for each part of the object in the form of a pixel-wise mask.

To grasp objects, we utilize the grasp affordance. Since AffordanceNet [6] output is 2D masks, 3d information corresponding to the mask of the grasp affordance is extracted. At this time, background or noise may be mixed, so remove them by performing clustering and over a certain distance. The refined 3D information of the grasp affordance is fitted to several primitive shapes to select the shape that best fits. From the selected shape, grasp information is generated. This information is necessary for the robot to grasp the object.

The grasp information generated by applying grasp affordance is different from other grasp information generation methods. First, this paper gives several candidate information for grasping objects. In addition, this paper does not end with grasping an object, but also provides information on how to use the object. This information is provided through other affordances than the grasp affordance. This information, which properly identifies the characteristics of the object, is suitable for social robots.

II. Related works

A. Grasp detection

There has been a lot of research in robotics to grasp objects. In the era before deep learning, [8, 9] researched to grasp novel objects. In [2], there was a detection research using deep learning. Redmon et al. [3] detected grasp information in real-time, and Chu et al. [4] detected multiple grasp information on many objects. Mousavian et al. [5] formulated grasp information generation using VAE for high grasp success rate. These studies focused on grabbing objects and lacked research on the use of objects.

B. Affordance detection

The affordance detection has been studied in computer vision and robotics in recent years. There was an affordance detection study on indoor scenes [10], but the most active study was on object affordance detection. Myers et al. [11] learned affordance from local shape and geometry primitives. Nguyen et al. [12] proposed an end-to-end network using deep learning with RGB-D as input.

In [13], training was performed with only affordance annotation of key points. In [14], affordance was detected using few annotated data.

Nguyen et al. [7] have provided IIT-AFF dataset to increase detection rate and add CRF to improve segmentation accuracy. Do et al. [6] proposed affordance detection network

The authors are with the Center for Robotics Research, Korea Institute of Science and Technology, Seoul, Korea (e-mail: kape67@kist.re.kr, choo2969@kist.re.kr, 091892@kist.re.kr, cjnam@kist.re.kr, ckim@kist.re.kr, gregorykim@kist.re.kr, respectively)

close to real-time by transforming Mask R-CNN [15]. We generate grasp information using AffordanceNet [6].

C. Grasp detection using affordance

In [16], grasp affordance is used for grasp detection. However, since they are only looking for grasp affordance, they do not understand the use of the object. In [1], the grasp affordance of the object is changed according to the task. For example, if the recognized object is a spoon and the task is a poke, the grasp affordance becomes the flat part of the spoon not general handle.

III. METHODOLOGY

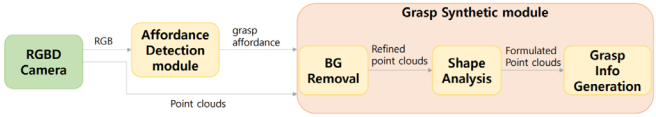

Figure 1. The proposed architecture of the grasp synthesis.

The proposed grasp synthesis algorithm is shown in the Fig. 1. The affordances of each part of the object (such as cut, contain, grasp and hit) are detected using the RGB. Among them, the grasp affordance is necessary for robot to grasp objects. The output of grasp affordance is a 2D mask. It is combined with its corresponding 3D point cloud data so that the 3D grasp information can be generated for robot to grasp the object. In the proposed grasp synthesis module, a proper set of grasp information can be provided to support various object manipulation according to the environment where the robots are located.

A. Affordance Detection

The affordance of an object is defined as the general actions people can take on the object. We pre-define affordance for each part of some objects and train the model to detect the affordance. We employ AffordanceNet [6] to detect affordance.

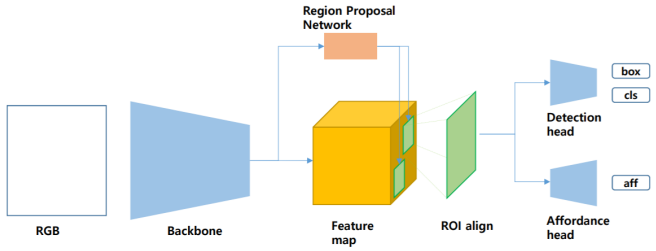

Figure 2. The Affordance Detection module.

AffordanceNet [6] is a network based on instance segmentation [15] (Fig. 2). In the AffordanceNet [6], a feature map is created through the backbone when the RGB image comes in. From this feature map, the region where the object is likely to be is extracted through the Region Proposal Network [18]. The feature map corresponding to this area is scaled to a fixed size through ROI align [15]. After that, the detection head finds the class and box for this area. If an object is found in the area, the affordance head finds the affordance mask in the area.

We used IIT-AFF dataset [7] with 10 object classes and 9 affordances (Table 1) for training. Affordance defined parts of an object as things that can be expressed in use. For example, a cup consists of contain and side-grasp affordances, and a hammer consists of pound and top-grasp affordances.

TABLE I. THE TYPES OF OBJECTS AND AFFORDANCES

Objects	Bowl	Monitor	Hammer	Pan	Knife
	Cup	Drill	Racket	Spatula	Bottle
Affordances	Contain	Hit	Cut	Display	Engine
	Support	Pound	w-Grasp	Grasp	

B. Grasp Synthesis

We referred Nam et al. [17] to generate grasp information. The grasp synthesis consists of 3 steps. First, refined 3D information of grasp affordance is obtained. After that, it fits into the most suitable pre-defined primitive shape. Finally, by generating grasp information using primitive shape, it provides the information necessary for robot to grab the grasp affordance of the object.

B.1. Background Removal

If affordances are detected through the affordance detection, check is needed whether the object that the robot is trying to grasp has grasp affordance. If the grasp affordance is detected, point clouds corresponding to the affordance are extracted. In this process, the point clouds may contain background due to the error between the 2D and 3D sensors. To eliminate the point clouds of the background, hierarchical clustering with single-linkage is used to remain only the largest cluster and remove the others. In addition, point clouds away from the camera along the camera z-axis are considered noise and removed. The refined point clouds are used to shape analysis

B.2. Shape Analysis

The point clouds are refined through above the process, but not organized. Therefore, we adopted the notion of the primitive shape. The primitive shape is the simplest geometric objects such as sphere, cube and cylinder. To organize the refined point clouds, the primitive shape that best matches the refined point clouds are found. We use 3 primitive shapes: sphere, cube and cylinder. When matching is failed, PCA algorithm is used to replace it with cubic. The result of shape analysis is in Fig. 3.

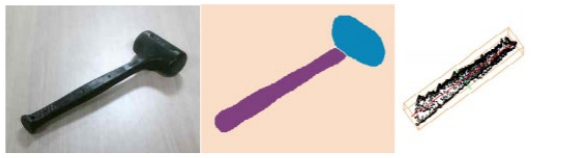

Figure 3. The result by fitting point clouds of grasp affordance to primitive shape.

B.3. Grasp Information Generation

Because the information of 3D grasp affordance can be organized, grasp information can be generated on the fitting shape. Grasp information is information necessary when a robot grasp an object.

We configure grasp information as follows.

a. grasp center point (GCP)

b. gripper approach direction (GAD)

c. gripper close direction (GCD)

To explain the composition of grasp information, since the 3D bounding volume is a cuboid, the inner center point of the cuboid becomes the GCP, and the vector from the center point of each face of the cuboid to the GCP becomes the GAD. Each GAD has one vertical GCD. This GCD is perpendicular to the long side and parallel to the short side on one square face of the associated cube when the GAD is determined. That is, we can calculate one 3D bounding volume for each object, and calculate one GCP and six GADs and GCDs for each 3D bounding volume. The result of one set of grasp information is in Fig. 4.

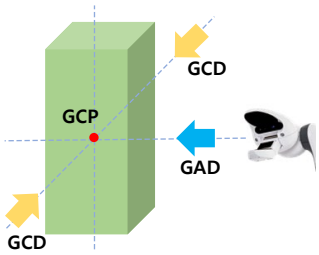

Figure 4. One set of grasp information

IV. EXPERIMENTS

In all experiments, the RGB-D Kinect2 Sensor that has 960x540 resolution was used. We experimented to evaluate affordance detection. IIT-AFF dataset [7] was used for training, and custom dataset was used for test. There is a total of 333 images in the test dataset and a total of 3,330 affordances. Metric is determined by the accuracy of the affordance corresponding to each pixel in the entire image. The accuracy is calculated as (TP + TN) / (TP + TN + FP + FN) when TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively [19]. Table 2 shows that affordance detection module achieves 88% accuracy on custom dataset. Fig. 5 shows the results of affordance detection experiment that detected 7 objects and 9 affordances.

TABLE II. EXPERIMENTS ON CUSTOM DATASET

Affordances	contain	cut	display	engine	hit
accuracy(%)	96.1	77.2	92.2	97.3	99.4
Affordances	pound	support	w-grasp	grasp	average
accuracy(%)	92.8	82.9	94.3	55.9	87.6

Figure 5. The result of affordance detection

We experimented whether a useful grasp information is generated. First of all, a tennis racket that has grasp affordance and has been easy to grasp by a Jaco robot arm was selected as the test object. In Fig. 6, grasp information was generated through grasp synthesis module. In Fig. 7, Jaco robot arm grasped the object using the generated grasp information.

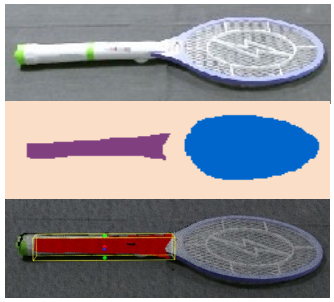

Figure 6. The result of grasp information generation. Top is RGB, middle is affordance map and bottom is the result of grasp synthesis represented by point clouds with RGB. Red point clouds are the 3D grasp affordance. Yellow box is a primitive shape that best matches the 3D grasp affordance. Dots of red, blue and green are GCP, GAD and GCD, respectively. The dots is the one set of grasp information.

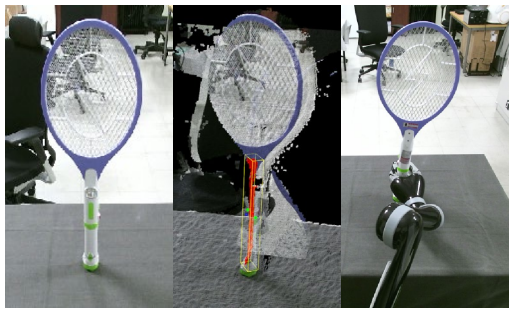

Figure 7. Left image is RGB, middle is the result of grasp synthesis and right the result of the robot grasping the object using the generated grasp information.

V. CONCLUSION

We have proposed a method of grasp information generation using object affordances. Unlike other methods, it provides various grasp information without focusing only on the grasp. It also provides the affordance of the object, helping the robot to do other tasks after grasping the object. In experiments, affordance detection module achieves 88% accuracy on custom dataset. We verified that our grasp synthesis using affordance is useful and Jaco robot arm grasped an object using generated grasp information.

ACKNOWLEDGMENT

This work was supported by the Industrial Strategic Technology Development Program (10077538, Development

of manipulation technologies in social contexts for human-care service robots) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

REFERENCES

[1] M. Kokic, J. A. Stork, J. A. Haustein and D. Kragic, "Affordance detection for task-specific grasping using deep learning." *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 91-98, 2017.

[2] Lenz, I., Lee, H., & Saxena, A, "Deep learning for detecting robotic grasps." *The International Journal of Robotics Research*, 34(4–5), pp. 705–724, 2015.

[3] Redmon, Joseph and Anelia Angelova. "Real-time grasp detection using convolutional neural networks." *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1316-1322, 2015.

[4] Chu, Fu-Jen & Xu, Ruinian & Vela, Patricio. "Real-World Multiobject, Multigrasp Detection." *IEEE Robotics and Automation Letters*. PP. 1-1, 2018.

[5] A. Mousavian, C. Eppner and D. Fox, "6-DOF GraspNet: Variational Grasp Generation for Object Manipulation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2901-2910, 2019.

[6] T. Do, A. Nguyen and I. Reid, "AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5882-5889, 2018.

[7] A. Nguyen, D. Kanoulas, D. G. Caldwell and N. G. Tsagarakis, "Object-based affordances detection with Convolutional Neural Networks and dense Conditional Random Fields," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5908-5915, 2017.

[8] Saxena, Ashutosh, et al. "Robotic Grasping of Novel Objects Using Vision." *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[9] Q. V. Le, D. Kamm, A. F. Kara and A. Y. Ng, "Learning to grasp objects with multiple contact points," *2010 IEEE International Conference on Robotics and Automation*, pp. 5062-5069, 2010.

[10] Roy, Anirban & Todorovic, Sinisa. "A Multi-scale CNN for Affordance Segmentation in RGB Images." *European Conference on Computer Vision (ECCV)*, pp. 186-201, 2016.

[11] A. Myers, C. L. Teo, C. Fermüller and Y. Aloimonos, "Affordance detection of tool parts from geometric features," *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1374-1381, 2015.

[12] A. Nguyen, D. Kanoulas, D. G. Caldwell and N. G. Tsagarakis, "Detecting object affordances with Convolutional Neural Networks," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2765-2770, 2016.

[13] J. Sawatzky, A. Srikantha and J. Gall, "Weakly Supervised Affordance Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197-5206, 2017.

[14] Sawatzky J., Garbade M., Gall J., "Ex Paucis Plura: Learning Affordance Segmentation from Very Few Examples." *German Conference on Pattern Recognition*, pp. 169-184, 2018.

[15] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980-2988, 2017.

[16] H. O. Song, M. Fritz, D. Goehring and T. Darrell, "Learning to Detect Visual Grasp Affordance," in *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 798-809, 2016.

[17] C. Nam et al., "A Software Architecture for Service Robots Manipulating Objects in Human Environments," in *IEEE Access*, vol. 8, pp. 117900-117920, 2020.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks." *28th International Conference on Neural Information Processing Systems* - Volume 1, pp. 91–99, 2015

[19] Wikipedia, Confusion matrix, https://en.wikipedia.org/w/index.php?title=Confusion_matrix&oldid=967505249 (last visited Aug. 1, 2020).