International Journal of Control, Automation and Systems 19(10) (2021) 3428-3434
http://dx.doi.org/10.1007/s12555-019-0758-1
ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

Irregular Depth Tiles: Automatically Generated Data Used for Network-based Robotic Grasping in 2D Dense Clutter

Da-Wit Kim, HyunJun Jo, and Jae-Bok Song*

Abstract: Recent advances in deep learning have enabled robots to grasp objects even in complex environments. However, a large amount of data is required to train the deep-learning network, which leads to a high cost in acquiring the learning data owing to the use of an actual robot or simulator. This paper presents a new form of grasp data that can be generated automatically to minimize the data-collection cost. The depth image is converted into simplified grasp data called an irregular depth tile that can be used to estimate the optimal grasp pose. Additionally, we propose a new grasping algorithm that employs different methods according to the amount of free space in the bounding box of the target object. This algorithm exhibited a significantly higher success rate than the existing grasping methods in grasping experiments in complex environments.

Keywords: Data generation, deep learning, grasping, manipulation.

1. INTRODUCTION

Robot arms, which were previously able to perform only certain tasks in confined spaces, can now organize items in boxes in home environments and sort objects that are irregularly placed at industrial sites. This is due to the development of deep-learning technology that enables machines to learn rules that are difficult to define by humans through data learning. In particular, deep learning-based algorithms exhibit excellent performance in various fields such as object recognition and text recognition [1]. Additionally, deep learning has been used to perform robotic grasping. Although grasping is the most basic task of manipulation, it is a difficult problem to define clearly, because contact with real objects is required. Therefore, studies are being actively conducted to enable robots to grasp objects using deep learning.

Deep learning-based robotic grasping can be categorized into two groups. The first involves grasping an object in a stable pose appropriate for the application [2, 3]. Recognizing and grasping the handle of a kettle to pour water into a cup corresponds to this group. The second involves grasping an object in a complex environment [4, 5]. Three-dimensional (3D) bin picking, which is a representative problem of robotic grasping, corresponds to this group. In a complex environment, estimating the grasp pose requires a large amount of data because several variables must be considered, such as the shape of the target object and positions and number of surrounding obstacles.

Most high-performance deep learning-based grasping algorithms use a large amount of data to train the network [6, 7]. However, because it is difficult to collect such a large amount of data, determining what data to use is critical. In one study based on supervised learning, the grasping quality was measured using a network trained with grasp data collected with a simulator [7]. This method has a high grasping success rate. However, it is difficult to collect additional data for training, because millions of data must be collected using the CAD models in a simulator. In contrast, in one study based on reinforcement learning, the network was trained by constructing a virtual environment in a simulator [8]. However, the training results for the simulator were not identical to the real-world results, owing to the differences between the two environments.

The novelty of the present work lies in the proposal of a method for transforming a depth image into the optimized data for grasping, as shown in Fig. 1(a), and a method for generating data automatically to reduce the data collection cost. Because the optimized data are used, the network structure for estimating the grasp pose is simple, and the data are clearly distinguishable according to the situation. Additionally, because the data have regularity, the network can learn the rules and determine quickly and accurately. Since this processing needs to know the area

Manuscript received September 9, 2019; revised January 5, 2021; accepted January 6, 2021. Recommended by Editor Euntai Kim. This research was supported by the MOTIE under the Industrial Foundation Technology Development Program supervised by the KEIT (No. 20008613).

Da-Wit Kim is affiliated with the Program in Mechatronics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Korea (e-mail: ekdnltrla@korea.ac.kr). HyunJun Jo and Jae-Bok Song are affiliated with the School of Mechanical Engineering, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, Korea (e-mails: {jhj0630, jbsong}@korea.ac.kr).
* Corresponding author.

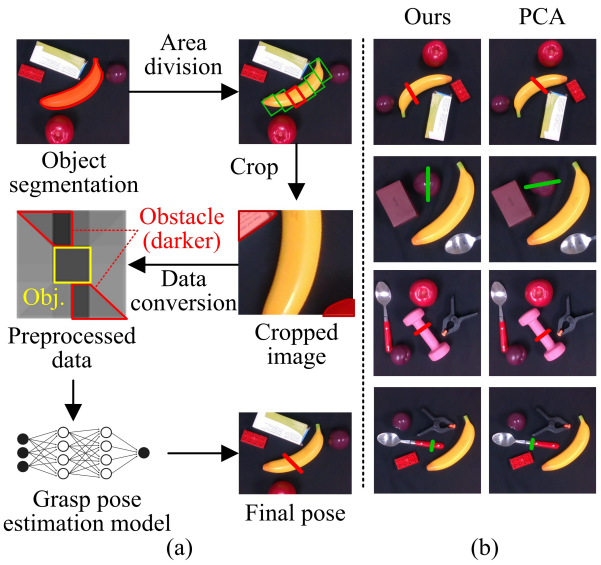

Fig. 1. (a) Process of grasping an object in dense clutter; (b) comparison of our grasping method and the principal component analysis (PCA)-based grasping method.

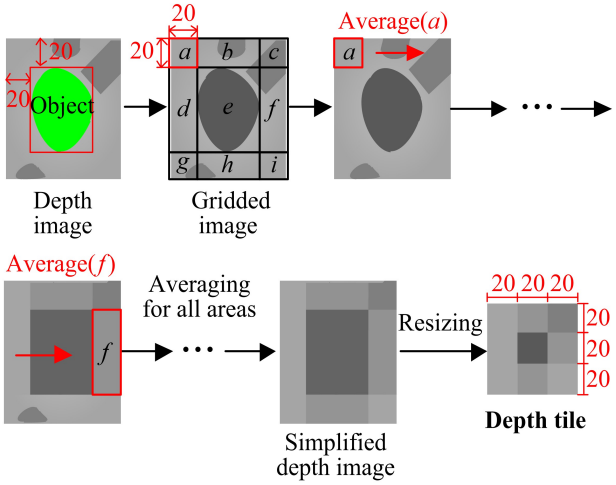

Fig. 2. Concept of modified average filtering.

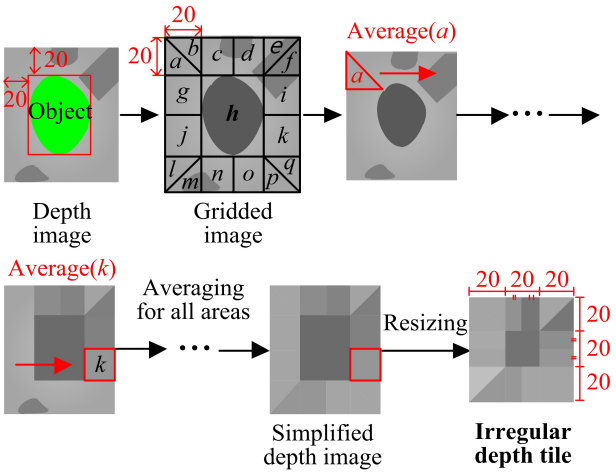

Fig. 3. Concept of irregular average filtering.

of the object on the RGB image, the pre-trained Mask R-CNN model was used [9]. In the subsequent explanations, it is assumed that an object mask is obtained using the Mask R-CNN scheme. Fig. 1(b) presents the grasp poses estimated in an actual complex environment. As shown, the grasp poses estimated using the proposed method did not cause interference with surrounding obstacles.

The remainder of this paper is organized as follows: Section 2 describes how the depth image is transformed into the optimized grasp data, and Section 3 describes how the training data are generated automatically and how the optimal grasp angle is estimated. Section 4 describes the overall grasping algorithm, and Section 5 presents the verification of the performance of the proposed grasping method in a complex environment. Finally, conclusions are drawn in Section 6.

2. IRREGULAR AVERAGE FILTERS

In the previous deep learning-based robotic grasping method developed in our laboratory, the depth image was converted into a simple form to solve the problem with a large amount of data [10]. In the corresponding work, the modified average filter (MAF) shown in Fig. 2 was proposed. The MAF divides the depth image of the object and its surroundings into several areas and replaces the pixel values of each area with the average pixel value of the area. All areas are then resized to 20×20 pixels. The depth image simplified by the MAF is called a "depth tile" because it is similar to a tile.

In this study, an irregular average filter (IAF) is proposed, which is an advanced form of the MAF. It transforms a depth image into a simple but grasp-optimized form. Fig. 3 describes the operation of the IAF. The depth image is divided into 17 areas from 'a' to 'q' with an object at the center, and the pixel values of each area are replaced by the average pixel values of the area. Then, resizing is performed such that each area has a constant size. The depth image processed by the IAF is similar to depth tiles that are irregular in shape; thus, it is denoted as the "irregular depth tile" (IDT) herein. Because the IDT is divided into a larger number of sections than the depth tile, the environment around the object can be described more elaborately.

3. GRASP POSE ESTIMATION MODEL

A grasp pose estimation model (GPEM) is proposed to estimate the optimal grasp angle. The network architecture is shown in Fig. 4. Two pairs of convolutional and pooling

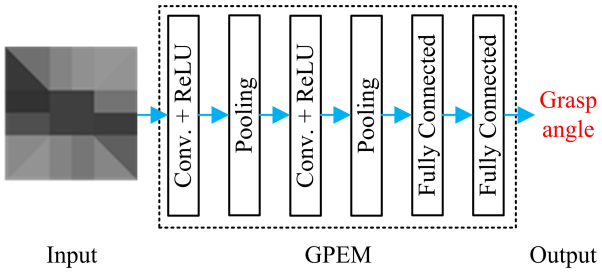

Fig. 4. GPEM (grasp pose estimation model).

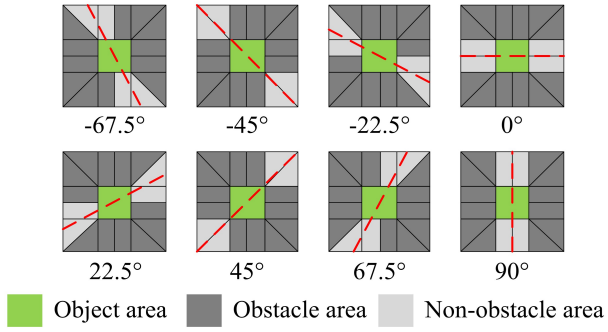

Fig. 5. Typical IDTs according to the grasp angle.

Table 1. List of angles.

Index	Angle (°)	Index	Angle (°)
1	−67.5	5	22.5
2	−45	6	45
3	−22.5	7	67.5
4	0	8	90

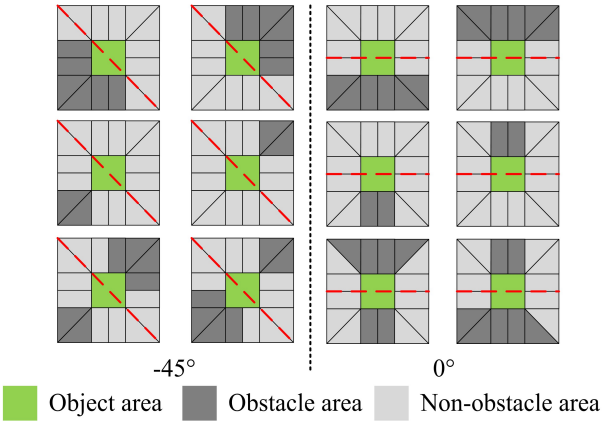

Fig. 6. Examples of possible IDTs.

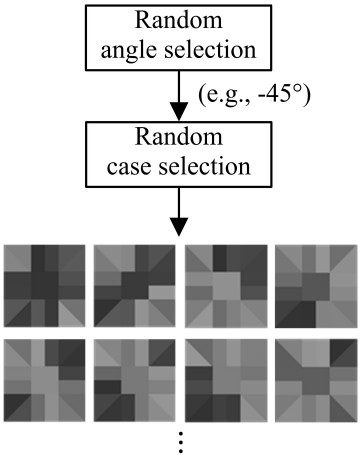

Fig. 7. Procedure of automated data generation.

layers extract features from the depth image and estimate the optimal grasp angle through the fully connected layers. The grasp angle is between −90° and 90° because the gripper is symmetrical.

The GPEM requires an IDT as an input and the grasp angle as an output for training. Although it is possible to randomly generate IDTs and label the grasp angles in an obstacle-free direction one-by-one, this method takes considerable time and effort. Therefore, in the present study, the training data were automatically generated by setting the grasp angle first and creating an IDT having no obstacles in that direction. Fig. 5 shows the typical IDTs according to the eight different grasp angles defined in Table 1. The angle was set at 22.5° intervals because this interval was suitable for clearly distinguishing the area of the IDT. As shown in Fig. 5, an IDT can be classified into an obstacle area, non-obstacle area, and object area. Numerous combinations of obstacle and non-obstacle areas are possible for each grasp angle, and all the combinations are automatically generated for training. Fig. 6 shows examples of some cases.

Fig. 7 shows the procedure of automated data generation. First, one angle is selected randomly from the eight angles in Table 1. Then, one case is selected by considering various combinations. After the angle and case are selected, the pixel values are specified according to the obstacle and non-obstacle areas. The pixel values of the obstacle area are 50–100 smaller than those of the non-obstacle area because the obstacle has smaller depth values than the non-obstacle area. Additionally, the depth values of the object area are set randomly to generate the data for objects of various heights.

The automatically generated data having pairs of IDTs and optimal grasp angles can be used for training. In this study, 90000 IDTs were generated, and the hyperparameters for training were set as shown in Table 2.

The following mean squared error (MSE) was used as the loss function for training, and the RMSprop algorithm was used as the optimizer [11].

$$L = \frac{1}{n}\sum(\hat{y} - y)^2, \tag{1}$$

where n is the number of data, $\hat{y}$ is the predicted output of the neural network, and y is the label representing the cor-

Table 2. Hyper-parameters.

Parameter	Value
Learning rate	0.01
Epoch	30
Steps per epoch	2000
Batch size	1000

rect answer. The reason for using the MSE as a loss function is that the network estimates a single value that can be compared with the label directly. It took approximately 5 min to complete the training with the aforementioned settings (using GTX 1080 Ti).

4. GRASPING ALGORITHM

The bounding box, which is obtained as a result of object segmentation, is composed of a target object area and free space, as shown in Fig. 8. If the amount of free space is large, there may be obstacles as well as an object inside the bounding box. Therefore, in this study, the free-space ratio of the object was derived through the following process. First, the major axis of the object mask was found using principal component analysis (PCA) [12]. In PCA, the major axis corresponds to the longest axis of the object mask. To obtain a bounding box that fits the object, the object mask was rotated so that the major axis was aligned with the vertical axis. Then, the free-space ratio r_f was defined as

$$r_f = \frac{A_{box} - A_{obj}}{A_{box}} \times 100\ (\%), \tag{2}$$

where A_{box} and A_{obj} represent the areas of the bounding box and the object, respectively. Note that the area is computed using the number of pixels of the mask.

In this study, the situations are classified into two cases based on $r_f = 30\%$. The first case ($r_f < 30\%$) is shown in Fig. 9. In this case, because there is insufficient free space in the bounding box, the bounding box itself is considered as the object area corresponding to h in Fig. 3. Then,

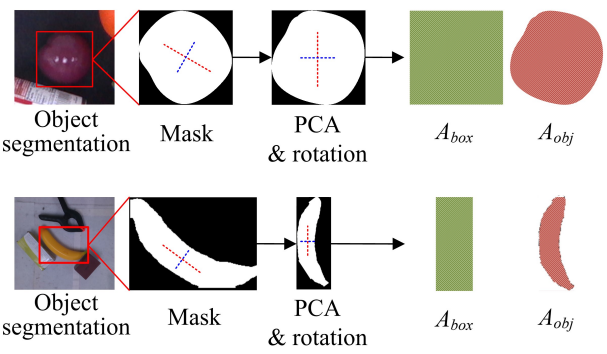

Fig. 8. Bounding box and object.

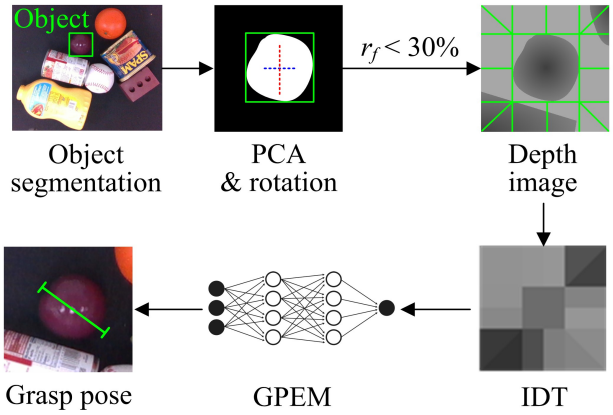

Fig. 9. Grasp pose estimation (free-space ratio < 30%).

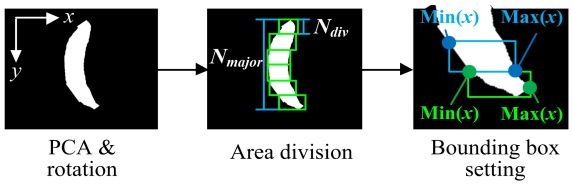

Fig. 10. Procedure of area division.

applying the IAF to the bounding box creates the corresponding IDT, which is given as an input to the GPEM.

In the second case ($r_f > 30\%$), there may be obstacles other than the object inside the bounding box because of the sufficient free space. Therefore, it is necessary to estimate the grasp pose more elaborately by dividing the object. In this case, the object area is divided into several subareas along the major axis as shown in Fig. 10. The subareas that fit the object area can be obtained by cutting the object mask along the major axis. Note that the grasp success rate increases when the grasp pose lies along the minor axis of the object. Therefore, after the angle of the major axis is determined via PCA, the image is rotated such that the major axis is vertical. The number of subareas in the rotated mask is obtained by

$$s = \mathrm{round}\,(N_{major}/N_{div}), \tag{3}$$

where s represents the number of subareas ($s = 6$ in Fig. 10), N_{major} is the maximum number of pixels of the major axis, and N_{div} is the number of vertical pixels of the subarea ($N_{div} = 30$ in this study), which happens to be the height of the subarea. The width of the subarea is the difference between the maximum and minimum positions in the x-axis for preventing the object from coming out of the bounding box and being recognized as an obstacle. The IAF is then applied to each of the subareas, as described in Section 2. Note that each subarea corresponds to the area h in Fig. 3.

Because the object is divided into several subareas, a criterion is needed to determine which of the estimated grasp poses will be the most stable. To this end,

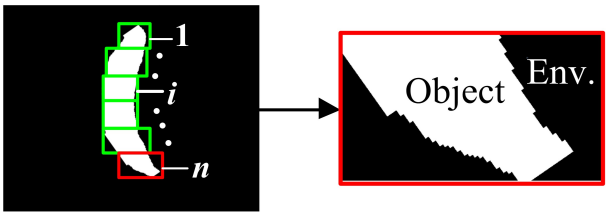

Fig. 11. Object area and environment area.

a grasping-quality (GQ) function is defined such that a larger GQ value leads to a higher grasping success rate. That is, the GQ function increases as the depth difference between the object and the environment increases, and the grasp position is located near the center of the object. Therefore, the GQ function well reflects the characteristics of humans grasping objects stably.

Suppose that the object area is divided into n subareas, as shown in Fig. 11. The depth difference $d(i)$ for subarea i is defined as

$$\Delta d(i) = d_{env}(i) - d_{obj}(i), \tag{4}$$

where $d_{env}(i)$ represents the average of the depth of the area outside subarea i, and $d_{obj}(i)$ represents the average of the depth of subarea i. The depth difference of (4) is usually positive because the distance is measured from the camera. Then, the GQ function of subarea i is defined as

$$GQ(i) = \begin{cases} \Delta d(i) & \text{if } 1 < i < n, \\ r \times \Delta d(i) & \text{if } i = 1 \text{ or } n, \end{cases} \tag{5}$$

where r is the discount factor for the edge areas which ranges between 0 and 1. If the robot grasps the edge of the object, it may fail to grasp the object firmly, owing to the weight of the object. Thus, the discount factor is used to account for the disadvantage in such a case. In this study, by setting r to 0.9, a certain subarea can be selected if it has a larger depth difference than other subareas, even if it is an edge area. Through this process, the robot performs grasping in the subarea having the maximum GQ. Fig. 12 shows the procedure of dividing the object area and performing grasping. The application of the IAF to the divided subareas creates n IDTs. The GPEM estimates the optimal grasp angle by taking each IDT as an input, and the best grasp angle with the highest GQ value is selected through the GQ function evaluation.

5. EXPERIMENTS

To verify the performance of the IDT-based grasping algorithm, experiments were performed in which a robot emptied a workbench where objects were placed close together. In the first experiment, only objects with a small free-space ratio were used, as shown in Fig. 13(b). In the

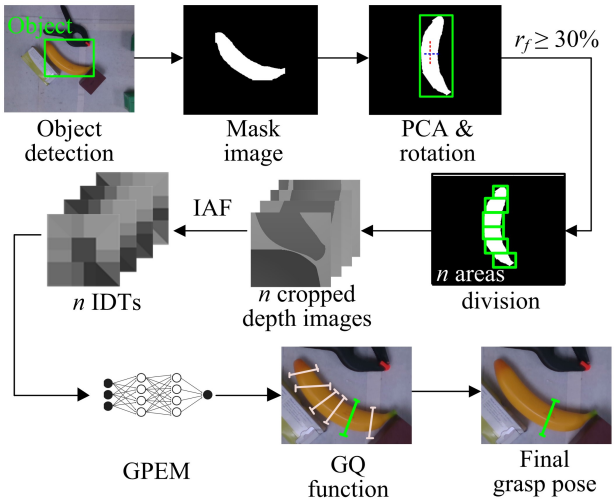

Fig. 12. Grasp pose estimation (free-space ratio $r_f \geq 30\%$).

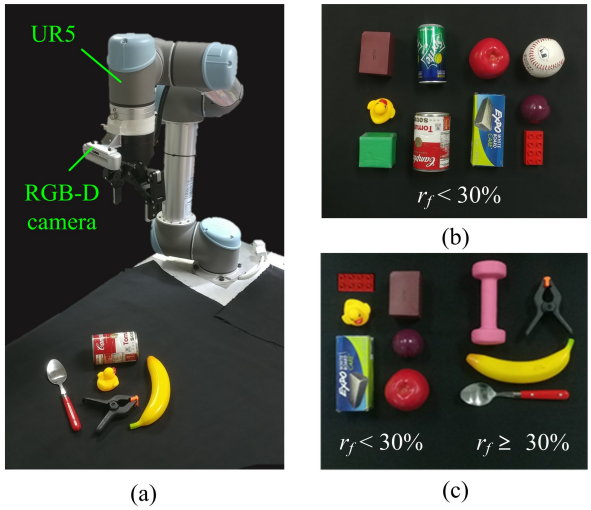

Fig. 13. (a) Experimental set-up, (b) test objects with free-space ratios (r_f) less than 30%, and (c) test objects with various free-space ratios.

second experiment, objects with both small and large free-space ratios were mixed together, as shown in Fig. 13(c). RealSense D435, an RGB-D camera manufactured by Intel, was used to acquire the RGB images of the objects, and UR5 was used as a robot arm to perform grasping, as shown in Fig. 13(a). In each experiment, the robot attempted to grasp five randomly chosen objects, and a total of 20 experiments were conducted. The five objects were placed close to each other in a ransom arrangement, and the objects that were grasped successfully were removed from the workbench.

In both experiments, three grasp pose estimation methods were compared; the PCA-based method, the MAF-based method, and the proposed IDT-based method. In the PCA-based method, PCA was applied to the object mask

Table 3. Experimental results.

Object name	Success rate		
	IDT-based	MAF-based	PCA-based
Brown brick	100	90	70
Sprite	90	80	60
Apple	100	100	60
Baseball	100	100	70
Plum	100	100	80
Red block	90	70	70
Board eraser	100	90	70
Tomato soup	90	70	70
Green box	90	90	80
Yellow duckie	90	80	60
Spoon	**90**	**60**	**30**
Banana	**90**	**40**	**10**
Dumbbell	**100**	**40**	**30**
Clamp	**100**	**50**	**30**

obtained through Mask R-CNN and the robot grasped the center of the object in the direction of its minor axis. In the MAF-based method, modified average filtering was applied to the depth image and the grasp pose was estimated from the processed image [10]. In the first experiment consisting of 100 trials involving objects with a small free-space ratio, the IDT-based method succeeded 95 times, while the PCA-based and MAF-based methods succeeded 69 times and 87 times, respectively. In the second experiment consisting of 100 trials involving objects with small and large free-space ratios, the IDT-based method succeeded 96 times, while the PCA-based and MAF-based methods succeeded 51 times and 72 times, respectively. Table 3 lists the grasping success rate for each object. As indicated by the results, the IDT-based grasping was far more successful than the PCA-based grasping and MAF-based grasping for all the objects. For the objects with a large free-space ratio (i.e., the spoon, banana, dumbbell, and clamp), the difference in performance was particularly large.

Fig. 14 shows the results of the IDT-based and MAF-based grasp pose estimation. In the IDT-based method, more accurate grasping was possible because the scheme considered the free-space ratio of the object. On the contrary, in the MAF-based method, inappropriate grasp poses were estimated for objects such as banana and clamp since the space within the object was not considered.

6. CONCLUSION

An IDT-based grasping method for complex environments was proposed. The depth image is converted into simplified grasp data using the proposed IAF, and the data for various situations can be automatically generated to reduce the time and cost for data collection. Additionally, elaborate grasping is possible owing to area division and the GQ function. For objects having small free space, the 95% success rate of the IDT-based method was far better than the 69% success rate of the PCA-based method. Additionally, when objects having small free space were mixed with objects having large free space, the 96% success rate of the IDT-based method was significantly better than the 51% success rate of the PCA-based method. The results indicate that the IDT-based grasping method achieves robust performance by considering the free space of the target object and the depth difference of the surrounding environment. In the future, the proposed approach will be extended to 3D cluttered environments to solve the 3D bin picking problem.

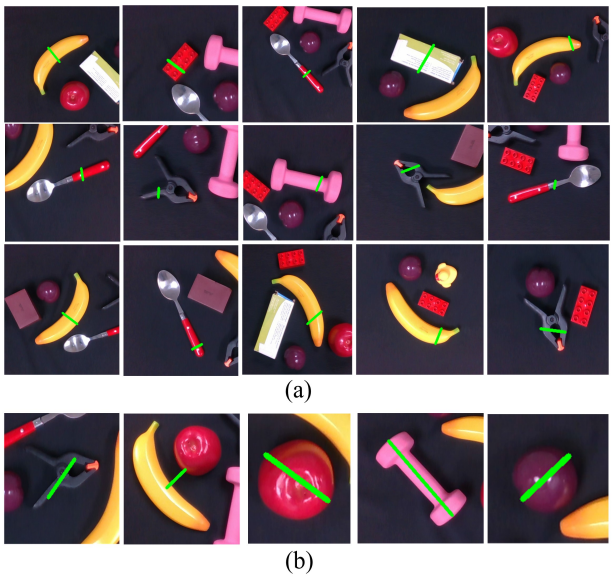

Fig. 14. Results of (a) IDT-based method and (b) MAF-based method.

REFERENCES

[1] H. Kim, J. Jo, Z. Teng and D. Kang, "Text detection with deep neural network system based on overlapped labels and a hierarchical segmentation of feature maps," *International Journal of Control, Automation and Systems*, vol. 17, no. 6, pp. 1599-1610, 2019.

[2] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 1316-1322, 2015.

[3] M. Li, K. Hang, D. Kragic, and A. Billard, "Dexterous grasping under shape uncertainty," *Robotics and Autonomous Systems*, vol. 75, pp. 352-364, 2016.

[4] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," *Proc. of*

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 598-605, 2016.

[5] A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazali, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Greem, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 1-8, 2018.

[6] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3406-3413, 2016.

[7] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *Robotics: Science and Systems (RSS)*, 2017.

[8] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2018.

[9] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961-2969, 2017.

[10] D. Kim, H. Jo, and J. Song, "Grasping method in a complex environment using convolutional neural network based on modified average filter," *Proc. of IEEE International Conference on Ubiquitous Robots (UR)*, IEEE, 2019.

[11] A. Graves, "Generating sequences with recurrent neural networks," arXiv preprint arXiv:1308.0850, 2013.

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37-52, 1987.

Da-Wit Kim received his B.S. degree in aerospace & mechanical engineering from Korea Aerospace University in 2018 and his M.S. degree in mechatronics from Korea University in 2020.

HyunJun Jo received his B.S. degree from the School of Mechanical Engineering of Korea University in 2016. He is currently pursuing a Ph.D. degree at the School of Mechanical Engineering of Korea University. His research interests include deep learning, robot manipulation, and computer vision.

Jae-Bok Song received his B.S. and M.S. degrees in mechanical engineering from Seoul National University in Seoul, Korea, in 1983 and 1985, respectively. He received a Ph.D. in mechanical engineering from M.I.T. in 1992. He has been a professor in the Department of Mechanical Engineering at Korea University since 1993. His research interests include robot safety, manipulator design and control, and AI-based robot applications. He is a senior member of IEEE.