Recalling of Multiple Grasping Methods From an Object Image With a Convolutional Neural Network

Makoto SANADA (gr0320ki@ed.ritsumei.ac.jp)
Ritsumeikan Daigaku Joho Rikogakubu Daigakuin Joho Rikogaku Kenkyuka https://orcid.org/0000-0001-5905-0318

Tadashi MATSUO
Ritsumeikan Daigaku

Nobutaka SHIMADA
Ritsumeikan Daigaku

Yoshiaki SHIRAI
Ritsumeikan Daigaku

Research Article

Keywords: object grasping, convolutional neural network, recalling grasping method, clustering

Posted Date: August 13th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51700/v1

Version of Record: A version of this preprint was published at ROBOMECH Journal on July 6th, 2021. See the published version at https://doi.org/10.1186/s40648-021-00206-4.

Recalling of Multiple Grasping Methods from an Object Image with a Convolutional Neural Network

Makoto SANADA (corresponding author)
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
gr0320ki@ed.ritsumei.ac.jp

Tadashi MATSUO
itsumeikan University
1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
matsuo@i.ci.ritsumei.ac.jp

Nobutaka SHIMADA
itsumeikan University
1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
shimada@i.ci.ritsumei.ac.jp

Yoshiaki SHIRAI
itsumeikan University
1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
ykshirai@gmail.com

Abstract
In this study, a method for a robot to recall multiple grasping methods for a given object is proposed. The robot learns grasping methods using a convolutional neural network to observe the grasping activities of human without special instructions. For this setting, only one grasping motion is observed for an object at a time. By acquiring multiple grasping methods, the robot learns by automatically clustering the observed grasping postures. In the proposed method, the grasping methods are clustered during the process of learning the grasping position. The method first recalls the grasping positions. The network for recalling the grasping position estimates the multi-channel heatmap that indicates one grasping position in each channel. The method then checks the graspability for each estimated position. Finally, it recalls the hand shapes that are based on the estimated grasping position and the object's shape. This study describes the results of recalling multiple grasping methods and demonstrates the effectiveness of the proposed method.

Keywords: object grasping, convolutional neural network, recalling grasping method, clustering

1. Introduction

Recently several studies have been conducted on robots grasping objects. In order for a robot to perform a grasping motion, huge amount of information is needed, which includes the object shape, grasping hand shape, and arm motion information. It is troublesome for the user to provide this information to the robot. Therefore, it is desirable for the robot to automatically generate the action of grasping an object. Considering that the robot manipulates the object after grasping, it is important to generate a grasping motion that is convenient for the following manipulation.

In several related studies, a variety of approaches have been proposed to recall the object grasping method. Ekvall et al. [1] proposed a method to select the grasping type with the highest grasping quality for the object shape, which is approximated by shape primitives among the multiple grasp types based on the prior database. Nagata et al. [2] proposed a method to find multiple grasping methods for an indicated object based on shape primitives and they were presented to the user. Huebner et al. [3] proposed a method to determine the grasping method for an object model represented by box primitives. Ekvall et al. used a prior database and reported that the robot cannot handle situations that are not defined in the database. In three different studies [1-3], the robot used object models represented by shape primitives; thus it is difficult to recall a grasping method for complex object shapes. To solve this problem, it is desirable to learn the grasping method using the

object shape without using a prior database and shape primitives. Several studies have been performed for estimating the various grasping methods and its confidence score using a neural network (NN) [4,5]. In addition, investigations have been performed to estimate the grasping method from the features of the posture and the color information of the object using the random forest classification algorithm [6,7]. These studies dealt with grasping that only carried objects. However, there are several ways to manipulate the object after grasping. For example, when we are carrying a cup, we usually grasp the upper part of the cup, and during drinking, we grasp the side of the cup. Because the grasping method is determined by the manipulation after grasping the object, it is necessary to learn to recall multiple types of grasping methods for a given object.

There is a method for learning the optimal grasping method by using reinforcement learning [8]. However, since reinforcement learning generally learns the optimal action for a single problem, it is difficult to learn multiple actions, such as multiple types of grasping methods. There is a method for learning multiple types of grasping methods by using supervised learning. When learning multiple outputs to a single input with supervised learning, it is generally necessary to provide multiple correct data to a single input [9, 10]. When learning a grasping method, the correct data is the data of the actual or simulated grasping of the object. It is difficult to acquire the correct data by using 3D models of the hand and the object. This is because a precise and realistic simulation environment is required. Therefore, it is necessary that the correct data is obtained by observing the grasping motions of the person. By using object and hand detection, visual information, such as an object shape, a grasping hand shape, and a grasping position on the object can be obtained as correct data. However, only one grasping motion can be observed in a single observation, and it is not possible to obtain the correct answers of multiple grasping methods at the same time. In order to learn multiple grasping methods for one object by observing the human grasping motion, it is necessary to cluster the grasping methods in the learning process and to recognize that there are multiple types of grasping methods.

In this study, we propose a method that uses a convolutional neural network (CNN) to learn multiple grasping methods. This is achieved by automatically clustering the grasping methods while learning human grasping motions through observation. We have divided the grasping method into the grasping position and the grasping hand shape. Therefore, learning the grasping method is divided into two steps: learning the grasping position for the object shape and learning the grasping hand shape for the object shape. These steps are performed using different networks. To cluster the grasping methods in the learning process, the network for the grasping position is designed to output multiple grasping positions for a single input. This network learns clustering by providing the correct position only to the channel output that is closest to the correct position for each learning sample.

When different objects have similar shapes such as a cup with and without a handle, the learned grasping method for one object might be recalled for a similar object that is different. Therefore, a grasping method may fail in some cases. For instance, grasping the side of a cup may fail due to the interference of the handle. When a person grasps an unknown object, the person recalls multiple grasping methods, and it simulates these grasping methods to judge the graspability. Even when the robot recalls the grasping methods, an approach is needed to determine the graspability the recalled grasping positions. Whether or not it can be physically grasped at the recalled grasping position depends on the physicality of the person or the robot. Therefore, it is desirable to identify the graspability by performing grasping or a simulation. However, it takes time to perform grasping or simulating each grasping positions. We learned that the NN can estimate the graspability multiple grasping positions with certainty. This proposed method recalls the grasping method that can grasp an object by inputting only the grasping positions with a high estimated certainty to the network, which estimates the grasping hand shape network.

The grasping hand shape network outputs a depth image of the grasping hand shape by inputting an object shape and one grasping position. Subsequently, it estimates the three-dimensional hand shape. However, the pixel of the depth image indicates only the depth distance, but it is not known which pixel in the image represents the hand. Therefore, the grasping hand shape network is designed to output two answers. This network learns the bitmap that represents the likelihood of the hand region in the image and the depth value in the hand region, respectively. By combining the

estimated hand region image and the depth value image, this network presents a three-dimensional grasping hand shape.

Section 2 describes the simultaneous recall method for multiple grasping methods, the network's learning method that learns the relationship between the object shape and the grasping method, and the learning method of the network that determines the graspability. In Section 3, we present the results of recalling multiple grasping methods while proving the usefulness of this study.

2. Methods

Fig. 1 The recalling flow of the multiple grasping methods

This study describes an approach that simultaneously recalls multiple grasping methods from one object image using the CNN. As illustrated in Fig.1, this approach consists of three networks: the grasping position network, the grasping hand shape network, and the grasping position certainty network. The grasping method is recalled using the grasping position network and the grasping hand shape network. The grasping position network estimates multiple grasping positions for an object. The grasping hand shape network estimates the hand shape for each estimated grasping position. The grasping position certainty network estimates the certainty, which represents the possibility of being able to grasp an object based on its position. The estimated certainty is used to determine whether the object can be grasped at the estimated position.

The process of recalling multiple grasping methods from an object image is as follows.

1. The multi-channel heatmap indicates one grasping position for each channel, which is generated by inputting the object depth image into the grasping position network.
2. The certainty for each grasping position candidate is estimated by inputting the combination of the object image and one channel of the multi-channel heatmap into the grasping position certainty network.
3. If the estimated certainty is greater than the threshold, it is determined that the grasping position is possible to achieve grasping.
4. The grasping hand shape image is generated by feeding the object image and the one-channel graspable position heatmap with high certainty into the grasping hand shape network.

The grasping position heatmap is an image that represents the likelihood of the grasping position for each pixel. An object image, a grasping position heatmap, and a grasping hand shape image are represented in the same image coordinate.

The learning method for each network is described in the following section.

Fig. 2 The learning flow of the grasping position network and the grasping hand shape network

2.1 Grasping Position Network

This network takes an object depth image and outputs the multi-channel heatmap that indicates one grasping position in each channel. Each channel represents a typical grasping position cluster for the objects. In our learning setting, the training dataset provides one correct answer for one input. This is because all the training data are assumed to be acquired in daily life scenarios where humans grasp objects. To recognize the different types of grasping methods by learning from a dataset, the network needs to automatically learn clustering by similar grasping types and object shapes. To create a cluster of similar grasping types during learning, the ground truth of the grasping position is given to the channel closest to the ground truth of the grasping position. This is among the multi-channel heatmaps that are tentatively estimated for the training input image in each network update iteration. In addition, a constraint is introduced where each channel image for the estimated grasping positions must be as different as possible. The reason for this is that the different types of

grasping positions are clustered for each of the channels. The loss function of this network is presented in Eq. (1), and the channel selection method is described in Eq. (2).

where: x is an input object image; $\varphi(x)$is the multi-channel heatmap that is estimated for x by the grasping position network $\varphi(\cdot)$; i, j, and k are the channel indices of the multi-channel

$$Loss_{position} = \frac{\left\|\varphi(x)^k - y_{pos}\right\|^2}{\left\|\varphi(x)^k + y_{pos}\right\|^2} + \left\{\sum_{i \neq j} \frac{\left\|\varphi(x)^i + \varphi(x)^j\right\|^2}{\|\varphi(x)^i - \varphi(x)^j\|^2 + \varepsilon}\right\} \quad (1)$$

$$k = \underset{i}{\operatorname{argmax}} \frac{\sum_{q \in Q_{pos}} \varphi(x)_q^i}{\sum_{p \in P} \varphi(x)_p^i} \quad (2)$$

heatmap; y_{pos} is the ground truth heatmap of the grasping position; Q_{pos} is the set of coordinates with pixel values greater than 0 in y_{pos}; P is a set of all the pixel coordinates for the one-channel heatmap; and p and q are coordinate indices. The first term in Eq. (1) is the expression that normalizes the squared error between the ground truth of the grasping position and the one-channel heatmap that is selected in Eq. (2). By minimizing this expression, this network is trained to estimate the one-channel heatmap that is closer to the ground truth. The second term is the expression that normalizes the inverse of the squared error between each channel of the output multi-channel heatmap. By minimizing this expression, this network learns to output different estimations (i.e. different grasping position) for each channel. Eq. (2) selects the channel that has the largest sum of the pixel values at coordinates that are included in Q_{pos} among the channels of the multi-channel heatmap that are normalized so that the sum of the pixel values is one for each channel.

2.2 Grasping Hand Shape Network

This network takes an object depth image and a one-channel heatmap, which indicates one grasping position candidate, and outputs a two-channel image, as shown in Fig.2. The first channel of the output image estimates the likelihood of the hand region for each pixel. In training, the binary image representing the hand region is given as the ground truth. The second channel estimates the depth value in the hand region. In training, the depth value is given to each pixel in the correct hand region. During the estimation, the hand shape image is recalled by multiplying the binarized hand region image with the hand depth image. The loss function of this network is presented in Eq. (3).

$$Loss_{hand} = \frac{1}{n(P)} \sum_{p \in P} \left\{-y_{handR_p} \log \psi(x, \varphi(x)^k)_{handR_p} - (1 - y_{handR_p}) \log(1 - \psi(x, \varphi(x)^k)_{handR_p})\right\} + \frac{1}{n(Q_{hand})} \sum_{q \in Q_{hand}} \frac{\left(\psi(x, \varphi(x)^k)_{handD_q} - y_{handD_q}\right)^2}{\left(y_{handD_q}\right)^2} \quad (3)$$

where: x is an input object image; $\varphi(x)^k$ is the k th-channel heatmap that is selected by Eq. (2); $\psi(x, \varphi(x)^k)_{handR}$ and $\psi(x, \varphi(x)^k)_{handD}$ are the hand region likelihood image and the hand region depth image, respectively, which are estimated by the grasping position network $\psi(\cdot)$;y_{handR} is the ground truth of the hand region likelihood; y_{handD} is the ground truth of the hand region depth; P is a set of all the pixel coordinates in the hand shape image; and Q_{hand} is the set of pixel coordinates in the hand region for the correct hand shape image. The first term is the expression of the cross entropy loss for the first channel. By minimizing this expression, this network learns the likelihood of the hand region for each pixel. The second term is the expression that normalizes the mean squared error of the depth value at the pixel coordinates that are included in Q_{hand} for the second channel. By minimizing this expression, this network learns to estimate the depth values that are closer to the ground truth in the hand region.

2.3 Grasping Position Certainty Network

Fig. 3 Learning flow of the grasping position certainty network

This network takes an object depth image and a one channel heatmap that indicates one grasping position candidate, and outputs a certainty that represents the graspability at that grasping position, as displayed in Fig. 3. As shown in the grasping possibility gate block of Fig. 1, the network estimates the certainty at each grasping position that is proposed by the grasping position network. To learn this network, it is necessary to prepare a sufficient number of training data that includes the graspable and the ungraspable positions for the objects to recall the grasping type. However, it is difficult to prepare the data for the graspability for each set of training data. Since the grasping position network clusters similar grasping positions during learning, it is expected that the grasping positions corresponding to typical grasping patterns will be output for each channel of the multi-channel heatmap. Therefore, we classified the training data into a few object types, such as cups with and without a handle. In addition, we selected in advance the channel that outputs the grasping position that can be grasped for each object type. When training this network, a probability of 1 is assigned as the ground truth if the input heatmap is the heatmap of the selected channel; otherwise, 0 is given. The loss function of this network is described in Eq. (4).

$$Loss_{certainty} = \frac{1}{c}\sum_{i=1}^{c}\left\{-{y_{cert}}^{i}\log\Phi(x,\varphi(x)^{i}) - (1-{y_{cert}}^{i})\log(1-\Phi(x,\varphi(x)^{i}))\right\} \quad (4)$$

where: x is an input object image, $\Phi(x,\varphi(x)^{i})$ is the estimated certainty for the i-th channel heatmap that is estimated by the grasping position certainty network$\Phi(\cdot)$, y_{cert} is the ground truth of the certainty, i is the channel index, and c is the number of channels of the multi-channel heatmap. This equation represents the average of the cross entropy loss for the estimated certainty. By minimizing this equation, this network learns the graspability at the input grasping position.

3. Experiment

To prove the usefulness of the proposed method, multiple grasping methods are recalled by using networks that learned the grasping methods as described in Section 2. In this experiment, we set the number of output channels of the grasping position network to three. Since it takes times to observe the human's motion and to collect the data of the grasping method, the artificial data of the various grasping methods are used for learning.

3.1 Dataset

The dataset consists of the object depth images as the input, and the grasping position heatmaps and the depth images of the grasping hand shape as the ground truth. When the training images are obtained by observing daily life scenes, only one grasping motion can be observed. If an object has multiple grasping methods, another type of grasping may be observed at the next opportunity for the same object; however, a new object image should be acquired in every observation. To simulate this scene observation, we defined each training sample as a triplet that consists of an input object image with only one grasping position heatmap and one hand shape image that is obtained by one observation. Since collecting many samples requires time-consuming efforts, we employed artificial training samples in which different grasping methods are associated with the same synthesized object images.

In this study, we prepared objects with two grasping types: grasping from above and from the side. The object and hand shape regions are extracted from a 16-bit depth image that are taken by Kinect for Windows with a depth sensor. Then, the object images are augmented by overlaying them on random background images that are taken by Kinect for Windows. The background of the hand shape images are set so that all their pixel values are 5,000. The grasping position heatmaps have an 8-bit pixel depth which have a peak value at the pixel specified as the grasping position and profiles like the Gaussian function.

The dataset is prepared by capturing 16 types of objects: eight types of cups with a handle and eight types of cups without a handle. These objects have different bottom depths, different sizes, and different shapes such as cylinders or inverted truncated cones, which were captured from eight different viewing angles. To increase the variety of object images, we randomly translated the objects by using 25 patterns. We rotated these objects in the range of -20 to 20 degrees and scaled them in

the range of 0.94 to 1.06. A grasping position heatmap and a hand shape image are also processed by the same transformation as an object image for consistency.

The dataset is divided into a training set, which includes 13 types of objects, and a test set of the three object types. We prepared 1.04 million training data and 1,200 test data in total.

The examples of the dataset are illustrated in Fig. 4.

Fig. 4 Examples of the dataset. Each row shows examples of the different grasping methods for the same object image. The object and hand shape image are displayed by a colormap for the visibility of the depth.

3.2 Results and discussion

Fig. 5 Results of recalling the grasping method with certainty for the training data. Each row in the second to rightmost columns corresponds to each typical grasping type that is clustered in the grasping position network. The third column displays the estimated grasping position and the ground truth are red and green, and the intersection pixels are yellow. The ground truth is displayed only on the image of the channel that is selected in Eq. (2). The pixel color of the object image and the hand shape image means that the depth values are converted by the colormap.

Fig. 6 Point cloud of the object and recalled grasping hand for the second object in Fig. 5. The object and hand points are blue and yellow, respectively.

Fig. 5 shows the results of recalling the grasping method and estimating its certainty for the training data, which includes five different object shapes. Although the hand shape image should be estimated only for the grasping position channel that is determined to be "graspable", namely with high certainty, the recalled hand shapes for any other grasping position candidates are presented here for analyzing the estimation process. Fig.6 shows the 3-D point cloud representation of the object and the recalled grasping hand shapes for the second object of Fig.5.

As depicted in Fig.5, multiple different grasping methods are estimated for a variety of object shapes. It can be observed that the grasping positions were automatically clustered into each channel of the multi-channel heatmap through the training process.

The second and third columns show that each estimated grasping position indicates each of the typical grasping positions that are observed in the training dataset. This includes the upper body, the handle, or the body's side from the input object. The grasping positions that are not seen in the dataset, such as an imaginary handle position of an object without a handle and the position of the handle's inside, are also recalled by the incorrect channels because their partial shape is similar to each other. It is determined that these impossible grasping positions have low grasping certainties and they can be rejected.

The mean error and standard deviation of the estimated position for all the training data are 0.75 [pixel] and 0.76 [pixel]. The average object height for the training data is approximately 8 cm and the pixel size projected on the image is around 20 pixels. Since this error value corresponds to 0.3 cm in real space, it can be observed that the estimated positions are near the ground truth. The largest error in the training data is about 3.6 [pixel], which is shown as a result for the fifth object in Fig. 5. This error value corresponds to 1.4 cm in real space, which is comparable to the size of a fingertip.

As shown in the fourth column in Fig. 5, the estimated hand shapes for each grasping position that is seen in the dataset are close to the ground truth, such as the first and second rows of the first object. For the grasping positions that are not seen in the dataset, such as the third row of the first object and the first row of the third object, the recalled hand shapes are plausible for grasping. However, grasping is impossible because the hand is apart from the object or it interferes with the object. These impossible solutions can be appropriately rejected by evaluating the grasping position certainty as shown in the right-most column of Fig. 5.

The mean error and standard deviation of the recalled hand "depth" for all the training data are 13.2 [mm], which is about the width of a finger, and 6.4 [mm]. An example of a poor result for the hand depth recall is shown in the second row of the second object in Fig. 5. The recalled grasping hand depth for that object has a 19 [mm] error on average for all the pixels in the image, but the pixels around the fingertip have a 10 [mm] error on average, which is more precise than the mean error for all the training samples. As shown in the point cloud in the second row of Fig. 6, the fingertips are precisely in contact with the object part to the handle. These results explain that the grasping hand shape network successfully learned the graspable hand shape at the grasping position for the various objects.

As shown at the right-most column in Fig. 5 which displays the grasping position certainty, the certainty value is close to 1 when grasping is possible at the input position; otherwise, it is close to 0. The grasping position certainty for all the training data is accurately predicted with a mean estimation error of 0.02 and its standard deviation of 0.13. In this study, we determined that the grasping type is possible for the input object if the certainty is over the threshold. We set the threshold as 0.8, which gives the highest accuracy of 98.3%, and a precision of 98.0% was achieved under that threshold setting.

Fig.7 displays the results of recalling the grasping method and estimating its certainty for the test data in the same way as Fig. 5 for the training data. Fig. 8 shows the 3-D point clouds of the object and the recalled grasping hand shapes for the second object in Fig. 7.

Fig. 7 Results of recalling the grasping methods with certainty for the test data

Fig. 8 Point cloud of the object and recalled grasping hand for the second object in Fig. 7

As shown in the second column in Fig. 7, multiple grasping positions are estimated for the unknown object images. The mean error and its standard deviation of the estimated grasping position for all of the test data are 1.16 [pixel] and 0.79 [pixel]. The average object height for the test data is approximately 7.5 cm and the pixel size that is projected on the image is around 20 pixels. Since this error value corresponds to 0.44 cm in real space, which is smaller than the width of a finger, the estimated positions are considered to be near the ground truth.

The fourth column in Fig. 7 shows the grasping hand shapes that are estimated for each estimated grasping position. The mean error and standard deviation of the recalled hand depth for all the test data are 19.5 [mm], which is larger than the width of a finger, and 7.3 [mm]. As shown in Fig.8, since the fingertips are in contact with the object and it is precise with the training data, it can be observed that the grasping hand shape network successfully estimated the grasping hand shape for the unknown object images.

As shown in the right-most column in Fig. 7, which displays the grasping position certainty, the estimated certainty value is close to 1 when grasping is possible at the input position; otherwise, it is close to 0. The grasping position certainty for all the test data is accurately predicted with a mean estimation error of 0.03 and its standard deviation of 0.15. The grasping position certainty network estimates the grasping position certainty for the unknown object images with a high accuracy of 97.3% and a high precision of 97.1%.

Fig. 9 The recall results of the multiple grasping methods for a real image. The region cropped by the white frame in the left image is the input object image. Each row of the right image corresponds to the recalled grasping method.

Fig. 9 shows the recall results of the multiple grasping methods for a real image. The left image is captured by a depth camera. Each row of the right image is the input object image that is overlaid with the recalled grasping method. This includes a grasping position, a grasping hand shape, and a grasping position certainty. In the lowest row, since the estimated certainty value is exceedingly

lower than the threshold of 0.8, grasping was determined to be impossible at that position and the grasping hand shape did not receive more estimations.

4. Conclusions

This study proposes a method to recall grasping methods for objects having multiple graspable positions and grasping hand shapes. This technique trains the CNN to recall multiple grasping methods by automatically clustering the object shapes and grasping types in the learning process without prior knowledge of the type and number of grasping methods for each object. The grasping positions common to each of the typical grasping methods are automatically clustered into one of the multi-channel heatmaps during learning. In addition, the CNN generates the grasping positions corresponding to the learned typical grasping methods. The plausible grasping methods for the input object are chosen by evaluating the estimated grasping position certainty as the graspability. The proposed method was applied to cups with and without a handle and the suitable grasping methods were successfully recalled with their certainties.

Future issues:

1. Extend the proposed method to objects that have grasping types not distinguished by the grasping position, such as holding a pen when writing and pinching it when the pen is being carried.
2. Develop a method to generate a motor command for grasping by the robot hand based on the recalled hand shape image.

Fig. 10 The recalling results for a variety of real images. The results that are determined to be ungraspable are based on the estimated certainty, which are displayed by the dark images.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The dataset used used during the current study is available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Ritsumeikan Global Innovation Research Organization (R-GIRO) and JSPS KAKENHI Grant Number 18H03313.

Authors' contributions

MS devised the concepts and design of the study, collected and analyzed data, and drafted the manuscript. TM contributed to analyze the estimated results of networks. NS contributed concepts and ideas, analyzed and interpreted the estimated results, and revised the manuscript. YS contributed concepts and ideas, interpreted the estimated results, and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Editage (www.editage.com) for English language editing.

Authors' information

Makoto SANADA (corresponding author)

Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu Shiga, JAPAN
gr0320ki@ed.ritsumei.ac.jp

Tadashi MATSUO, Nobutaka SHIMADA, Yoshiaki SHIRAI
College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, JAPAN

References

[1] S. Ekvall and D. Kragic (2007) Learning and Evaluation of the Approach Vector for Automatic Grasp Generation and Planning. IEEE International Conference on Robotics and Automation. doi: 10.1109/ROBOT.2007.364205.
[2] K. Nagata et al. (2010) Picking up an indicated object in a complex environment. IEEE/RSJ International Conference on Intelligent Robots and Systems. doi: 10.1109/IROS.2010.5651257.
[3] K. Huebner and D. Kragic (2008) Selection of robot pre-grasps using box-based shape approximation. IEEE/RSJ International Conference on Intelligent Robots and Systems. doi: 10.1109/IROS.2008.4650722
[4] F. Chu, R. Xu and P. A. Vela (2018) Real-World Multiobject, Multigrasp Detection. IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 3355-3362. doi: 10.1109/LRA.2018.2852777.
[5] H. Zhang et al. (2019) ROI-based Robotic Grasp Detection for Object Overlapping Scene. IEEE/RSJ International Conference on Intelligent Robots and Systems. doi: 10.1109/IROS40897.2019.8967869..
[6] U. Asif, M. Bennamoun and F. A. Sohel (2017) RGB-D Object Recognition and Grasp Detection Using Hierarchical Cascaded Forests. IEEE Transactions on Robotics, vol. 33, no. 3, pp. 547-564. . doi: 10.1109/TRO.2016.2638453
[7] J. Zhang et al. (2020) Robotic grasp detection based on image processing and random forest. Multimedia Tools and Applications 79:2427–2446. doi: 10.1007/s11042-019-08302-9
[8] S. Korkmaz (2018) Training a Robotic Hand to Grasp Using Reinforcement Learning. ReseachGate.
[9] F. Mueller et al. (2017) Real-Time Hand Tracking Under Occlusion from an Egocentric RGB-D Sensor. IEEE International Conference on Computer Vision Workshops (ICCVW). doi: 10.1109/ICCVW.2017.82.
[10] Z. Cao et al. (2017) Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR.2017.143.

Figures

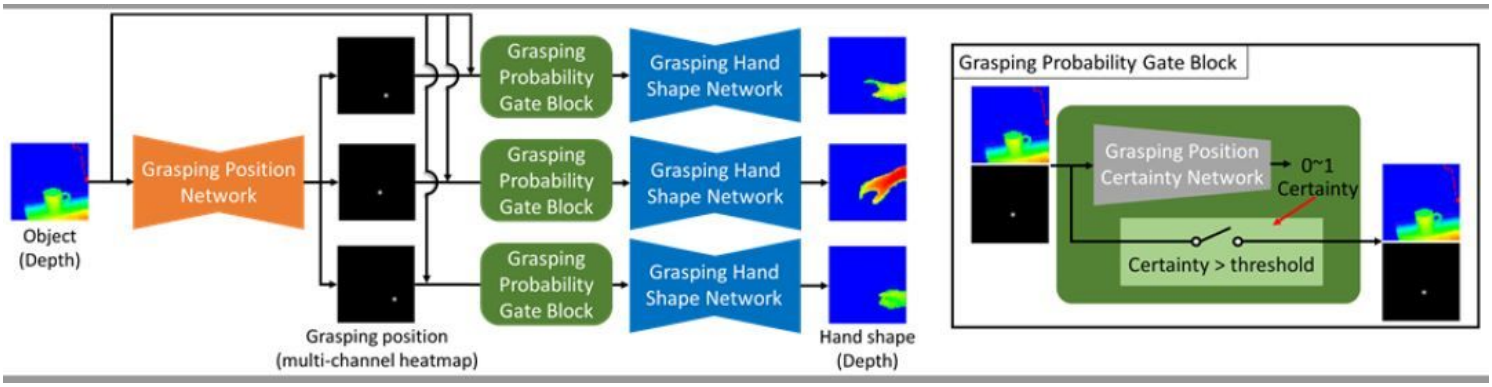

Figure 1

The recalling flow of the multiple grasping methods

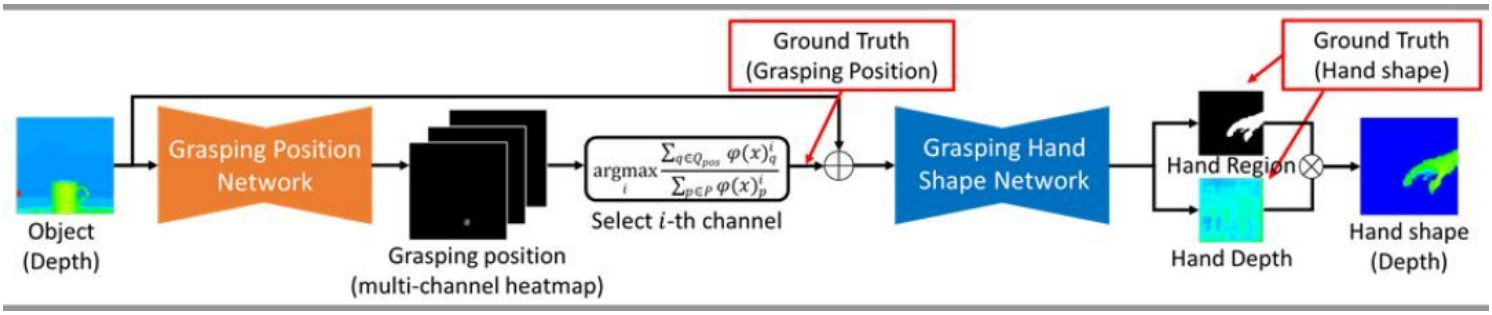

Figure 2

The learning flow of the grasping position network and the grasping handshape network

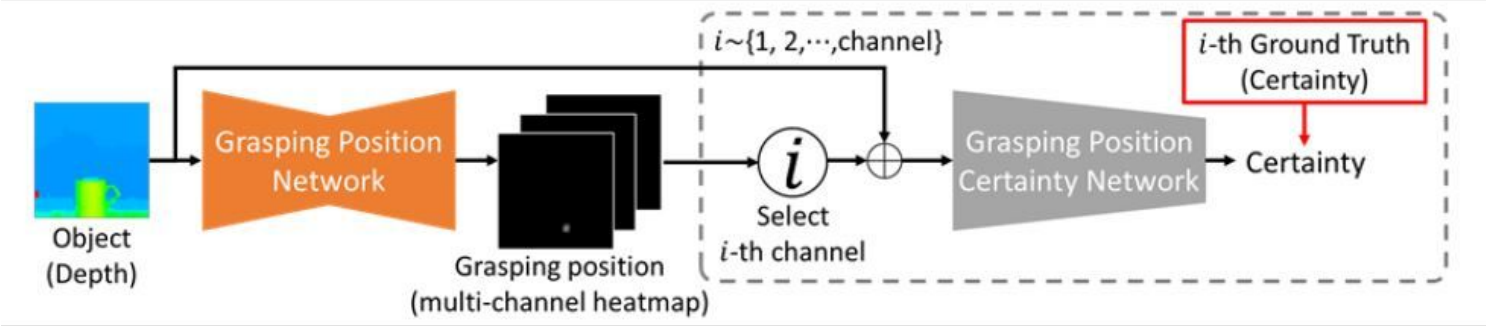

Figure 3

Learning flow of the grasping position certainty network

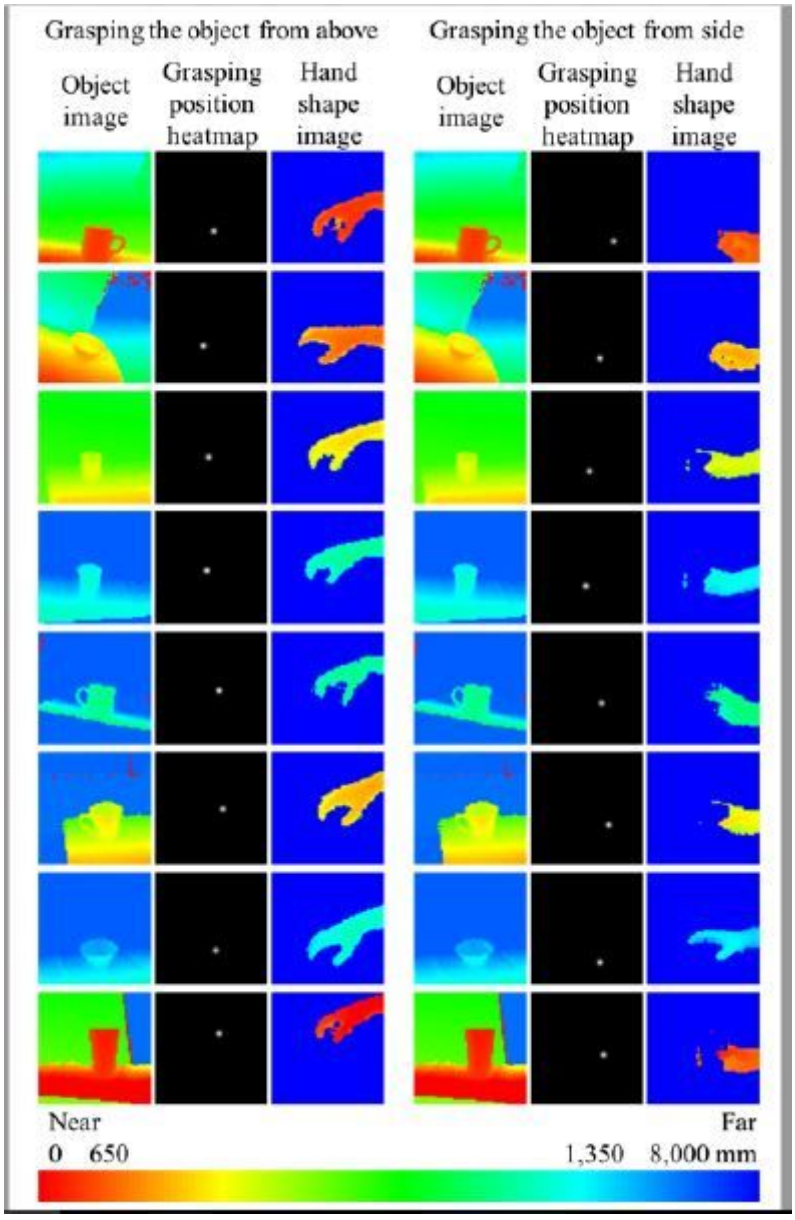

Figure 4

Examples of the dataset. Each row shows examples of the different grasping methods for the same object image. The object and hand shape image are displayed by a colormap for the visibility of the depth.

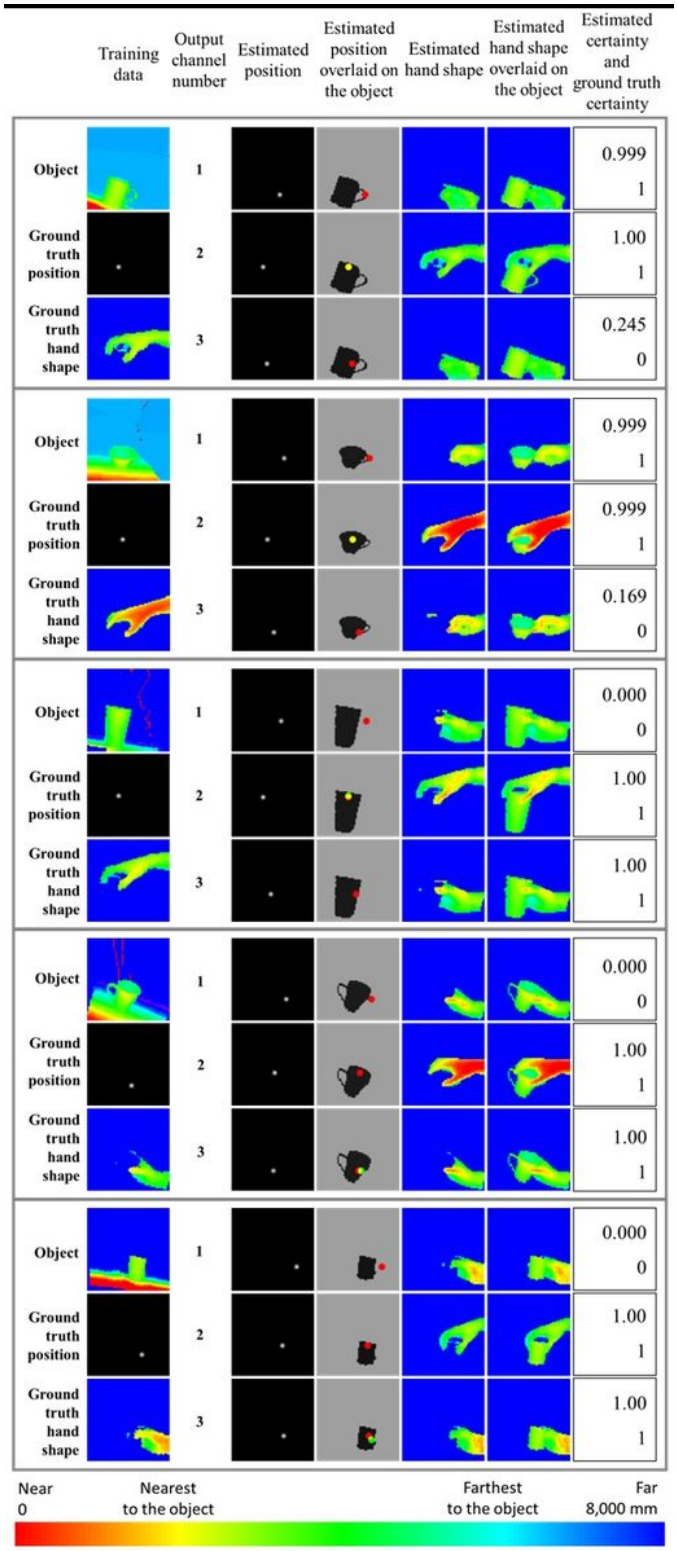

Figure 5

Results of recalling the grasping method with certainty for the training data. Each row in the second to rightmost columns corresponds to each typical grasping type that is clustered in the grasping position network. The third column displays the estimated grasping position and the ground truth are red and green, and the intersection pixels are yellow. The ground truth is displayed only on the image of the

channel that is selected in Eq. (2). The pixel color of the object image and the hand shape image means that the depth values are converted by the colormap.

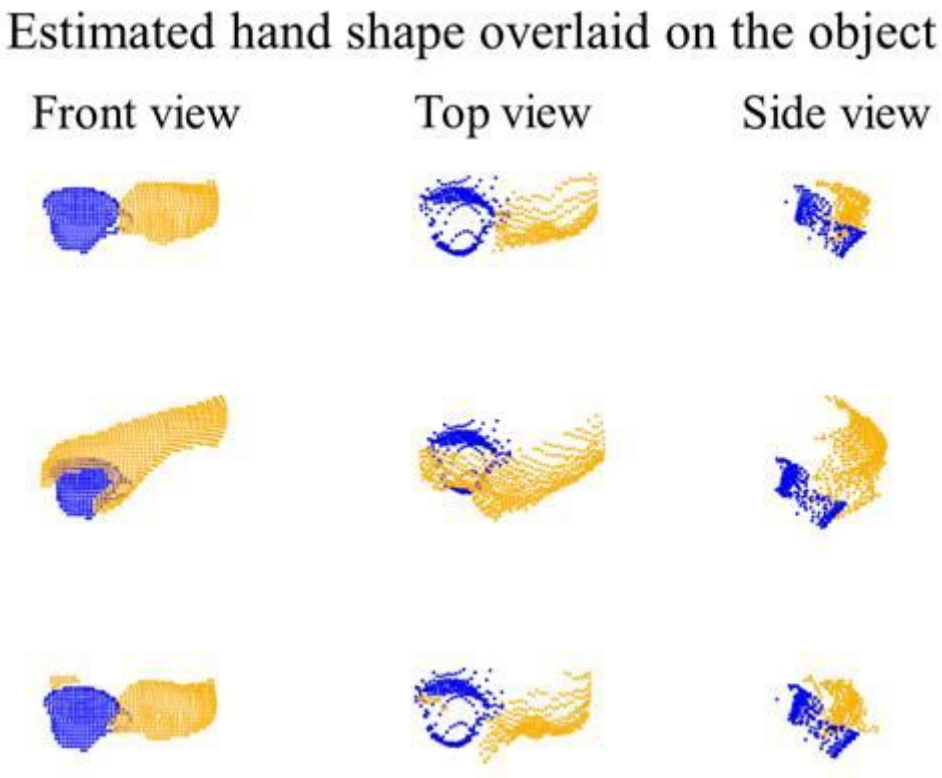

Figure 6

Point cloud of the object and recalled grasping hand for the second object in Fig.5. The object and hand points are blue and yellow, respectively.

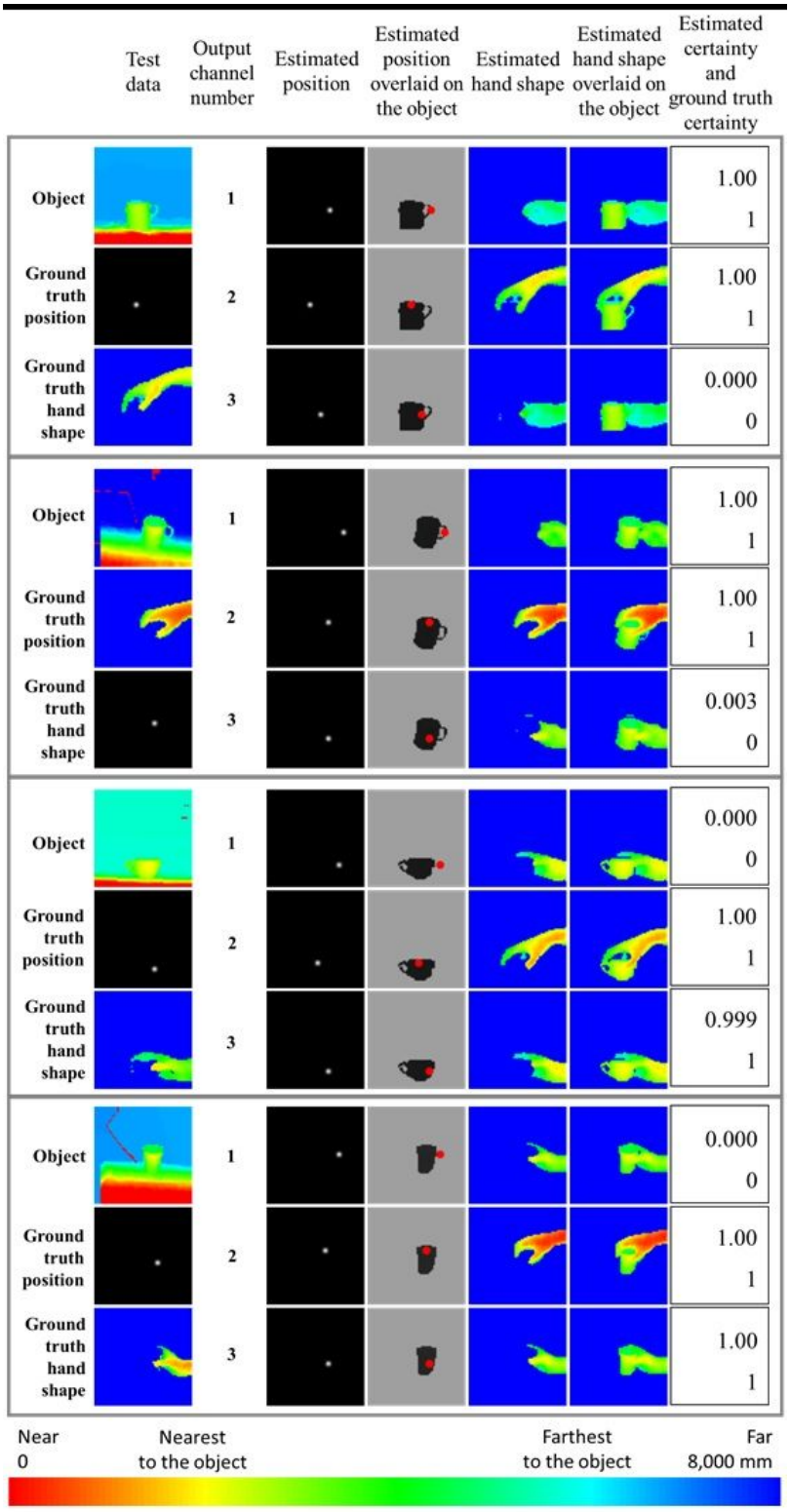

Figure 7

Results of recalling the grasping methods with certainty for the test data

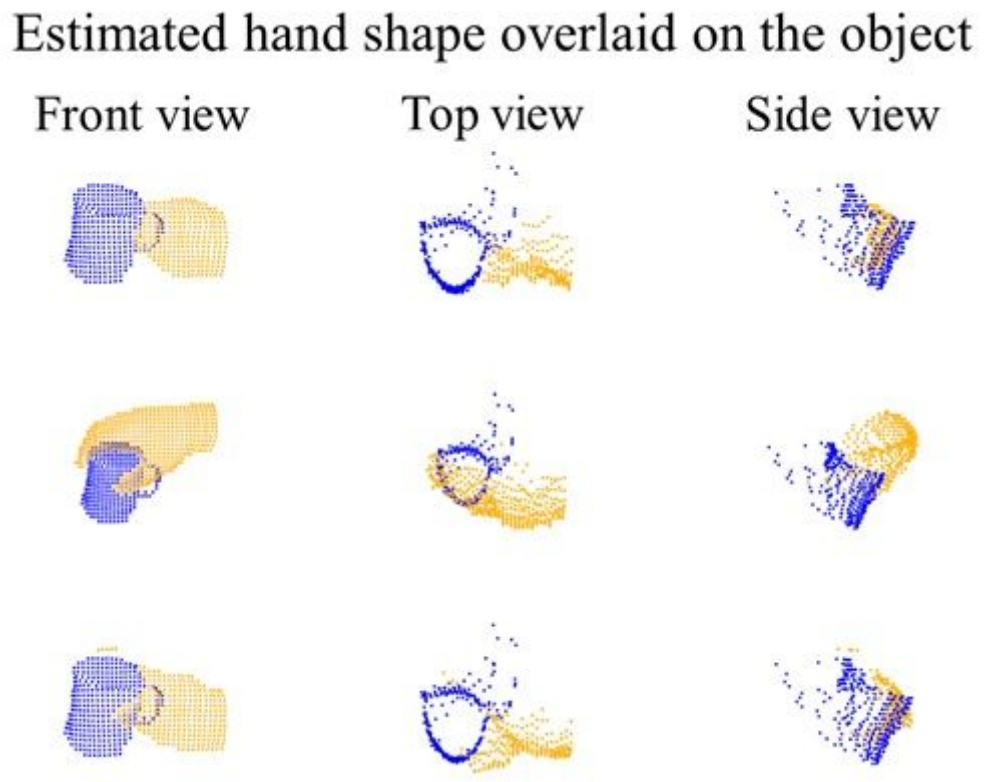

Figure 8

Point cloud of the object and recalled grasping hand for the second object in Fig.7

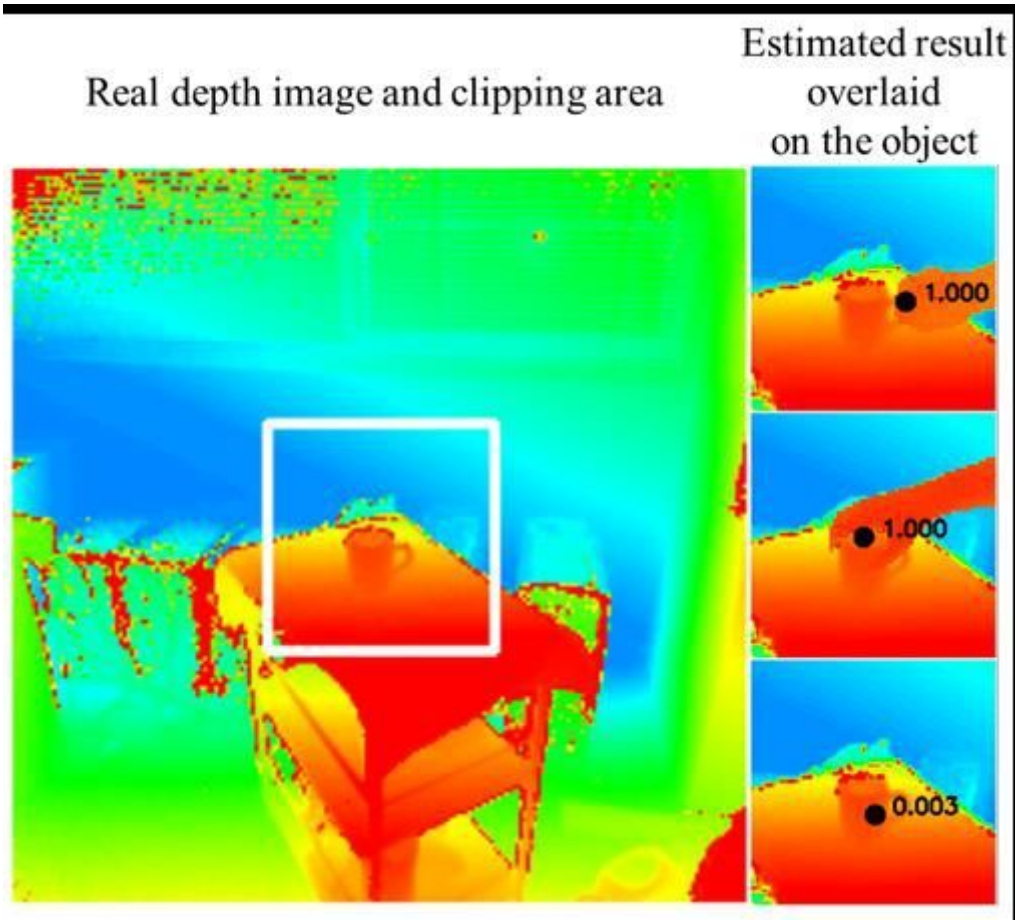

Figure 9

The recall results of the multiple grasping methods for a real image. The region cropped by the white frame in the left image is the input object image. Each row of the right image corresponds to the recalled grasping method.

Figure 10

The recalling results for a variety ofreal images. The results that aredetermined to be ungraspable are based on the estimated certainty, which are displayed by the dark images.