# 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 Joho Rikogakubu Daigakuin Joho Rikogaku Kenkyuka

**Nobutaka SHIMADA**
Ritsumeikan Daigaku Joho Rikogakubu Daigakuin Joho Rikogaku Kenkyuka

**Yoshiaki SHIRAI**
Ritsumeikan Daigaku Joho Rikogakubu Daigakuin Joho Rikogaku Kenkyuka

**Research Article**

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

**Posted Date:** December 29th, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-132179/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
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu Shiga 525-8577 JAPAN
matsuo@i.ci.ritsumei.ac.jp

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

Yoshiaki SHIRAI
Ritsumeikan 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 automatically clustering the observed grasping postures, the robot learns multiple grasping methods. In the proposed method, the grasping methods are clustered during the process of learning of the grasping position. The method first recalls grasping positions. The network for recalling the grasping position estimates the multi-channel heatmap such that each channel heatmap indicates one grasping position. The method then checks the graspability for each estimated position. Finally, it recalls the hand shapes based on the estimated grasping position and the object's shape. This study shows 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 grasping method depends on the succeeding manipulation, it is important to generate multiple patterns of grasping methods.

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. In order to use the prior database of the grasping type, it is necessary to prepare the object shape patterns and the grasping type patterns in advance and define the relationship between them. Nagata et al. [2] proposed to approximate an object by shape primitives and find multiple grasping methods that are presented to the user. In order to use shape primitives as in [1, 2], it is necessary to prepare shape primitives for expressing various object shapes. As a method that does not require the shape primitives, there is a method of recalling the grasping method from a realistic object shape using machine learning. Several studies have been performed

for estimating the various grasping methods and its confidence score using a neural network (NN) [3,4]. 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 [5,6]. These studies recall only one type of grasping method for an object. 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. Huebner et al. [7] estimates a suitable grasping method for post-grasping manipulation from among multiple grasping methods for the object shape approximated by the box primitive. However, this study requires prior knowledge of the number and type of grasping method and the relationship between the grasping method and the manipulation. Our study aims to learn multiple grasping methods by observing the grasping motion of a person without prior information such as the number and type of grasping methods.

Korkmaz [8] proposed to learn the optimal grasping method using reinforcement learning. However, since reinforcement learning generally learns the optimal action for a single problem, it is difficult to learn multiple types of grasping methods. Mueller et al. [9] and Cao et al. [10] used supervised learning to learn multiple outputs for a single input. Generally supervised learning requires one correct data for one answer. Correct data of the grasping method can be obtained by the actual or simulated grasping of the object. For simulating the grasping, a precise and realistic simulation environment with 3D models of the hand and the object is required. It is difficult to implement complex physics enough to simulate realistic correct data. In actual grasping, correct data are obtained by observing the grasping motions of a person. Once object and hand interaction is observed in daily life, visual information (i.e. 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 other grasping methods from the observed motion. Multiple grasping methods for one object is learned by clustering the grasping methods in the learning process.

In this study, we propose to use 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. Grasping positions are clustered by giving the correct position for each learning sample only to the network channel that the position closest to the correct position.
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 with a different shape. This is because the number of grasping positions estimated by the grasping position network is set to a fixed number. 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 the NN to estimate the graspability of multiple grasping positions as a certainty. This network can estimate the certainty from the object shape and the specified grasping position, and is used to select only the grasping positions that can actually be grasped among the estimated grasping positions. In this proposed method, the grasping method that can grasp an object are recalled by inputting only the grasping positions with a high estimated certainty into the network for estimating the grasping hand shape.

The grasping hand shape network outputs a depth image of the grasping hand shape by inputting an object shape and one grasping position. This network is a model that extracts the relationship between the local object shape and the hand shape, and can recall a three-dimensional hand shape according to the local shape feature of the object at the specified grasping position.

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 a threshold, it is determined that the grasping is possible at that grasping position.
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 the dataset, the network needs to automatically learn clustering by similar grasping types and object shapes. A cluster of similar grasping types is created by giving the ground truth of the grasping position only to the channel closest to the ground truth of the grasping position during learning. 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

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

estimated grasping positions must be as different as possible. This is because 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 heatmap; $w$ is a weight parameter; $y_{pos}$ is the ground truth heatmap of the grasping position; $P$ is a set of all the pixel coordinates for the one-channel heatmap; and $p$ and $q$ are coordinate indices;

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

$Binary(\cdot)$ is the function that binarizes the pixel values larger than a threshold to one and it or less to zero. 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 overlap between the estimated heatmap and the ground truth of grasping position. Once a channel is trained by feeding a ground truth, the channel becomes to generate the heatmap with a single high peak and a small variance for the other similar input object image, while the other channels still generate a heatmap with low peaks and a large variance. Therefore, when a different grasping position as the ground truth, the other channel tends to have higher responses around the ground truth and then that channel rather than pretrained channel is easy to be selected. In other words, every channel greedily learns similar grasping positions and the same typical grasping pattern is aggregated for each channel.

If the grasping position network has just one output channel (i.e., the grasping position is not clustered), the trained network outputs a two-peak heatmap for the grasping position as shown in Fig. 3 (b). When recalling the grasping hand shape from such a heatmap indicating two grasping positions, the hand shape overlaid with the two types of the grasping hand shape is estimated as shown in Fig. 3 (c). With such an overlaid hand shape image, it is difficult to determine which pixel represents which type of hand shape. Therefore, our method prepares a sufficient number of output channels for the number of correct grasping position and clusters the grasping positions so that individual grasping methods can be recalled.

Fig.3 Effect of the number of output channels of the grasping position network

2.2 Grasping Hand Shape Network

This network takes an object depth image and a one-channel heatmap, which indicates one grasping position, 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, only the depth values in the correct hand region are given to the corresponding pixels in this channel. In predicting, the hand shape image is recalled by masking the hand depth image with the binarized hand region image.

If the hand shape is generated by only one channel depth image, the depth value for the background region is learned in addition to that for the hand region. Since the background region is much larger than the hand region, loss value is mainly determined by the background depth than the hand and the details inside the hand tends to be neglected as shown in Fig. 4 (c). The two-channel representation (depth and mask) can recall the three-dimensional hand shape more accurately as shown in Fig. 4 (b).

Fig.4 Effect of the number of output channels of the grasping hand shape network. (a) the correct hand shape, (b) the estimation result by the two-channel output network, (c) the estimation result by the one channel output network.

The loss function of this network is presented in Eq. (3).

$$Loss_{\text{region}} = \frac{1}{n(P)} \sum_{p \in P} \left\{ -y_{handR_p} \log \psi(x, \varphi(x)^k)_{handR_p} - (1 - y_{handR_p}) \log(1 - \psi(x, \varphi(x)^k)_{handR_p}) \right\}$$

$$Loss_{depth} = \frac{1}{n(Q_{hand})} \sum_{q \in Q_{hand}} \frac{\left( \psi(x, \varphi(x)^k)_{handD_q} - y_{handD_q} \right)^2}{\left( y_{handD_q} \right)^2}$$

$$Loss_{hand} = Loss_{\text{region}} + Loss_{depth} \quad (3)$$

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. $Loss_{\text{region}}$ 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. $Loss_{depth}$ 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. 5 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. 5. 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. However, it is difficult to prepare the data of the graspability for all grasping positions. 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 use the learned grasping position network to learn this network. We classified the training data into a few object types, such as cups with and without a handle and 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 and the weight parameter $w$ in Eq. (1) to one. 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 Structure of Networks

Table 1 Details of the Structure of Each Network

This method uses three networks, and each network was designed with referent to the lightweight model of Resnet in [9]. Table 1 shows the details of the structure of each network. The grasping position network estimates the three-channel heatmap from the object depth image. The grasping hand shape network and the grasping position certainty network have two input data, an object depth image and a grasping position heatmap, and estimate a hand shape image or a certainty.

### 3.2 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 simultaneously. If an object has multiple grasping methods, another type of grasping method may be observed at the next opportunity for the same object. However, a new object image should be obtained in every observation. To simulate this scene observation, we defined each training sample as a triplet that consists of an input object image, 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 19 types of objects: cups with and without a handle, watering cans, teapots, and containers. These objects have different bottom depths, different sizes, and different shapes such as cylinders or inverted truncated cones, and are captured from eight different viewing angles. To increase the variety of object images, we randomly translated the objects by shifting them in 25 patterns, rotating them in the range of -20 to 20 degrees, and scaling 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 15 types of objects, and a validation set of the four object types. We prepared 640,000 training data and 1,600 validation data in total.

The examples of the object used in this experiment is shown in Fig. 6, and the examples of the dataset are show in Fig. 7.

Fig. 6 Example of object used in this experiment

Fig. 7 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.

### 3.3 Results and discussion

Fig.8 Loss of the grasping position, the hand shape, and the certainty in each epoch

Fig.9 Mean error and standard deviation per a sample of the grasping position, the hand depth, and the certainty

Fig.8 shows the change of the loss during the learning process of each network and Fig.9 shows the mean error and standard deviation of the grasping position, the hand shape, and the certainty, in the final epoch. In Fig.8, all the losses of them converge to a constant value as the learning progresses. In Fig.9, there are no large difference between the mean error of the training set and the validation set.

Fig. 10 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 in red and green, and the intersection pixels in 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 and the hand shape images encodes the depth values by the "red(near)-blur(far)" colormap.

Fig. 11 Point cloud of the object and recalled grasping hand for the second object in Fig. 10. The object and hand points are blue and yellow, respectively.

Fig. 10 shows the estimated results of the grasping method and its certainty for the training data, which includes five different object shapes. While it may be enough that the hand shape only for the grasping position channel judged as "graspable" (i.e., with high certainty) is obtained, for analyzing the estimation process, the recalled hand shapes for any other grasping position candidates are presented here. Fig.11 shows the 3-D point cloud representation of the object and the recalled grasping hand shapes for the second object of Fig.10.

As depicted in Fig.10, 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, such as the upper body, and the handle or the body's side of the input object, that are observed in the training dataset. 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 one of the channels because their partial shape is similar to each other. These grasping positions are judged as having low grasping certainty and then rejected.

As shown in Fig. 9, the mean error and standard deviation of the estimated grasping position for all the training data are 0.53 [pixel] and 0.55 [pixel]. The average object height for the training data is approximately 8 cm and the size projected on the image is around 20 pixels (corresponding to 0.2 cm in 3-D space scale) and the estimated positions may be regarded as near the ground truth enough. The largest error in the training data is about 2.2 [pixel], which is shown as a result for the fifth object in Fig. 10. This error value corresponds to 0.9 cm in 3-D space scale, which is comparable to the size of a fingertip.

As shown in the fourth column in Fig. 10, the estimated hand shapes for each grasping position that is included in the dataset, such as the second and third rows of the first object, are close to the ground truth. For the grasping positions that are not seen in the dataset, such as the first row of the first object and the second row of the fifth object, the recalled hand shapes are plausible for grasping. However, that grasping is actually impossible because the hand is apart from the object or it interferes with the object. These ungraspable methods can be appropriately rejected by evaluating the grasping position certainty (=0.011) shown in the right-most column of Fig. 10.

As shown in Fig. 9, the mean error and standard deviation of the recalled hand "depth" for all the training data are 21.7 [mm], which is about the width of a finger, and 10.4 [mm]. An example of a poor result for the hand depth recall is shown in the first row of the second object in Fig. 10. This result has a 30 [mm] error on average for all the pixels in the hand region, 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 first row in Fig. 11, 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. 10, the certainty value is close to one when grasping is possible at the input position; otherwise, it is close to zero. In this experiment, when the grasping position is recalled inside of the handle or in the air instead of an object, it is learned that it cannot be grasped at those positions because the handle interferes with the grasping hand or there is no object part to grasp. In Fig. 10, it can be confirmed those grasping positions have a certainty close to zero. From the above results, it can be seen that the trained model is clustered as different grasping patterns when there is a handle and not a handle, and the certainty can be estimated appropriately for the object shape and grasping position. As shown in Fig. 9, the grasping position certainty for all the training data is accurately predicted with a mean estimation error of 0.03 and its standard deviation of 0.13. In this study, we determined that the grasping is possible for the input object if the certainty is over the threshold. We set the threshold to 0.5 for the highest F-value, which gives the high accuracy of 97.8% and the precision of 99.1%.

Fig.12 displays the results of recalling the grasping method and estimating its certainty for the validation data in the same way as Fig. 10 for the training data. Fig. 13 shows the 3-D point clouds of the object and the recalled grasping hand shapes for the fourth object in Fig. 12.

Fig. 12 Results of recalling the grasping methods with certainty for the validation data

Fig. 13 Point cloud of the object and recalled grasping hand for the fourth object in Fig. 12

As shown in the second column in Fig. 12, multiple grasping positions are estimated for the unknown object images. As shown in Fig. 9, the mean error and its standard deviation of the estimated grasping position for all of the validation data are 2.04 [pixel] and 1.80 [pixel]. The average object height for the validation data is approximately 8 cm and the pixel size that is projected on the image is around 20 pixels (corresponding to 0.9 cm in 3-D space scale, which is smaller than the finger width), the estimated positions are considered to be near the ground truth enough. Although there is another estimated grasping with a large position error, such as shown in the result of the second row of the fifth object, even such an example successfully recalls the graspable handle-shape part.

The fourth column in Fig. 12 shows the grasping hand shapes recalled for each estimated grasping position. As shown in Fig. 9, the mean error and standard deviation of the recalled hand depth for all the validation data are 33.3 and 21.1 [mm]. Hand shapes with a very large depth error, such as the result of the third row of the fourth object, were rare case. The mean error and standard deviation of the recalled hand depth except those bad cases are 24.9 [mm] and 10.5 [mm], which are close to the values for the training data. As shown in Fig.13, since the fingertips are in contact with the object, it can be observed that the grasping hand shape network successfully estimated the grasping hand shape for the unknown object images.

The right-most column in Fig. 12 shows the estimated certainty. As shown in Fig. 9, the grasping position certainty for all the validation data is accurately predicted with a mean estimation error of 0.07 and its standard deviation of 0.20. The grasping position certainty network estimated the certainty for the unknown object images with an accuracy of 93.8% and a precision of 96.5% under the condition of threshold of 0.5 at which the F-value is maximized.

Fig. 14 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. 14 and 15 show examples of recall results of the multiple grasping methods for real images. The recalling procedure is performed in real time by employing CUDA-driven GPU board (Geforce GTX1080). The depth images for a total of 20 kinds of unknown objects were taken for 5 seconds at 15 fps. The left side of Fig. 14 is a captured depth image and each row of the right side is the recalled grasping method overlaid on the input image which 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 0.5, it was judged as impossible to grasp at that position and no grasping hand is overlaid. Fig. 15 shows recall results for the other objects in the same manner as Fig. 14.

The second column in the first row of Fig. 15 shows the recall result for the elephant-shaped watering can with a handle similar to a cup, and grasping methods with high certainty are recalled for the handle and the upper part, and a grasping method with low certainty is recalled for the inside of the handle. It can be seen that the grasping methods can be recalled in response to similar local shape features even for an object having large different shape from the learned object.

## 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 heatmap 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 different objects with similar shape features, such as cups with and without a handle, watering cans, teapots, and containers, 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. 15 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] 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.
[4] 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..
[5] 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
[6] 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
[7] 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
[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

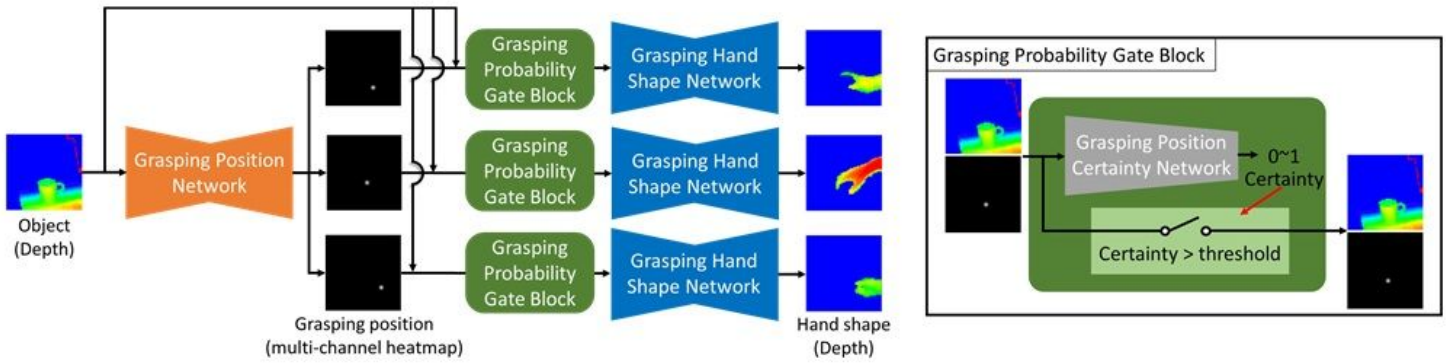



**Figure 1**

The recalling flow of the multiple grasping methods

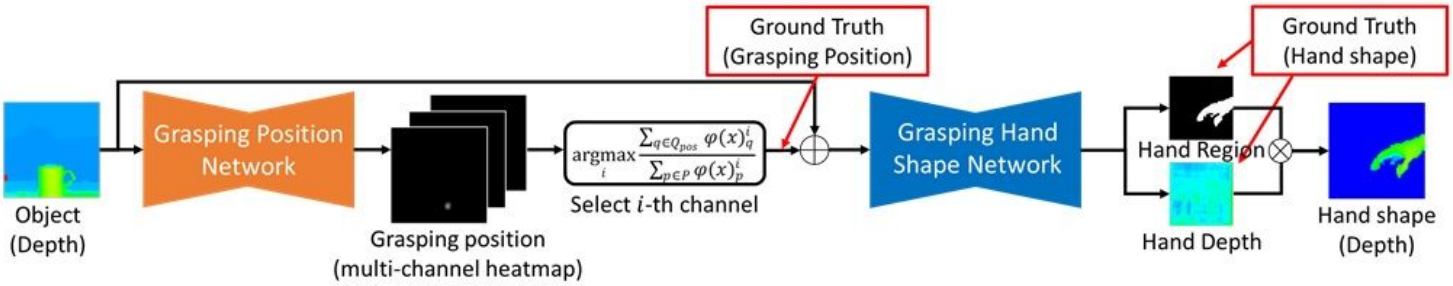



**Figure 2**

The learning flow of the grasping position network and the grasping hand shape network

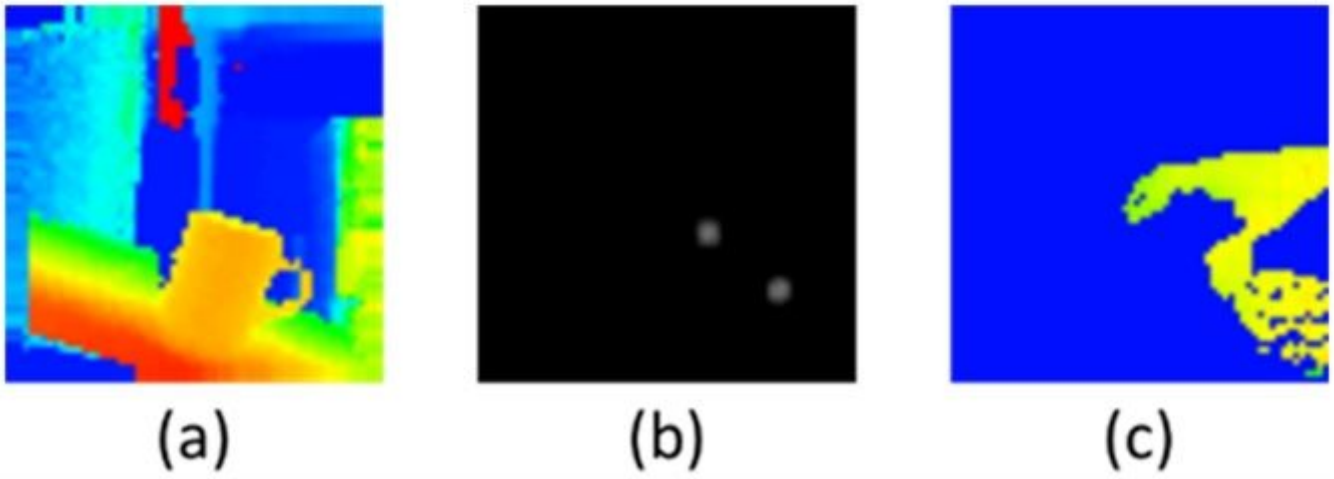



**Figure 3**

Effect of the number of output channels of the grasping position network

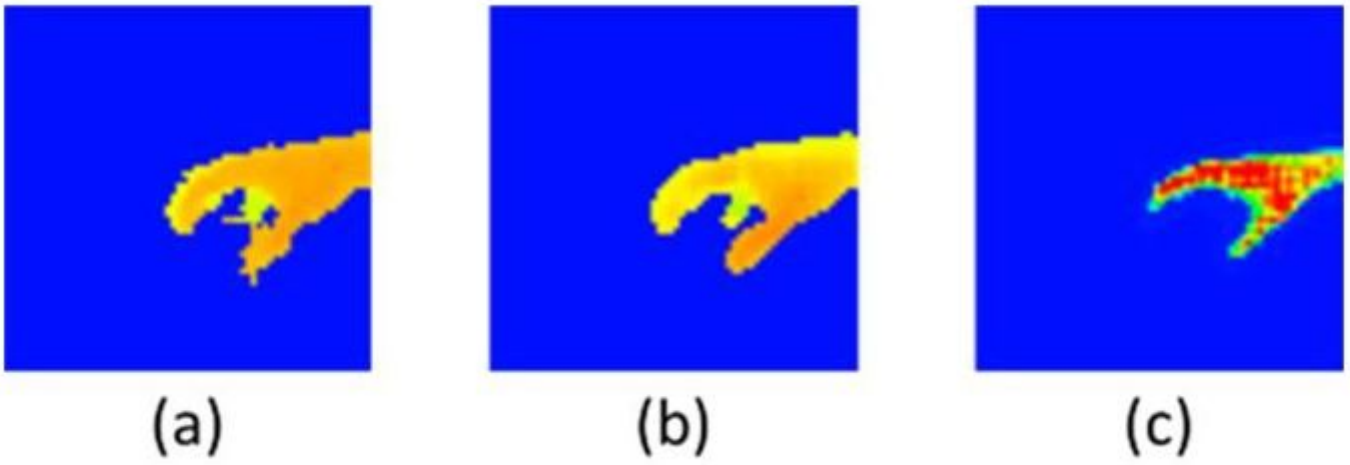



**Figure 4**

Effect of the number of output channels of the grasping hand shape network. (a) the correct hand shape, (b) the estimation result by the two-channel output network, (c) the estimation result by the one channel output network.

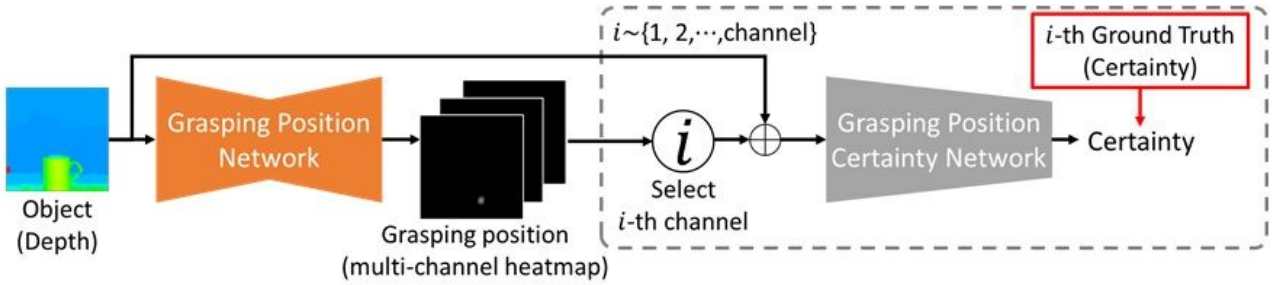



**Figure 5**

Learning flow of the grasping position certainty network

**Figure 6**

Example of object used in this experiment

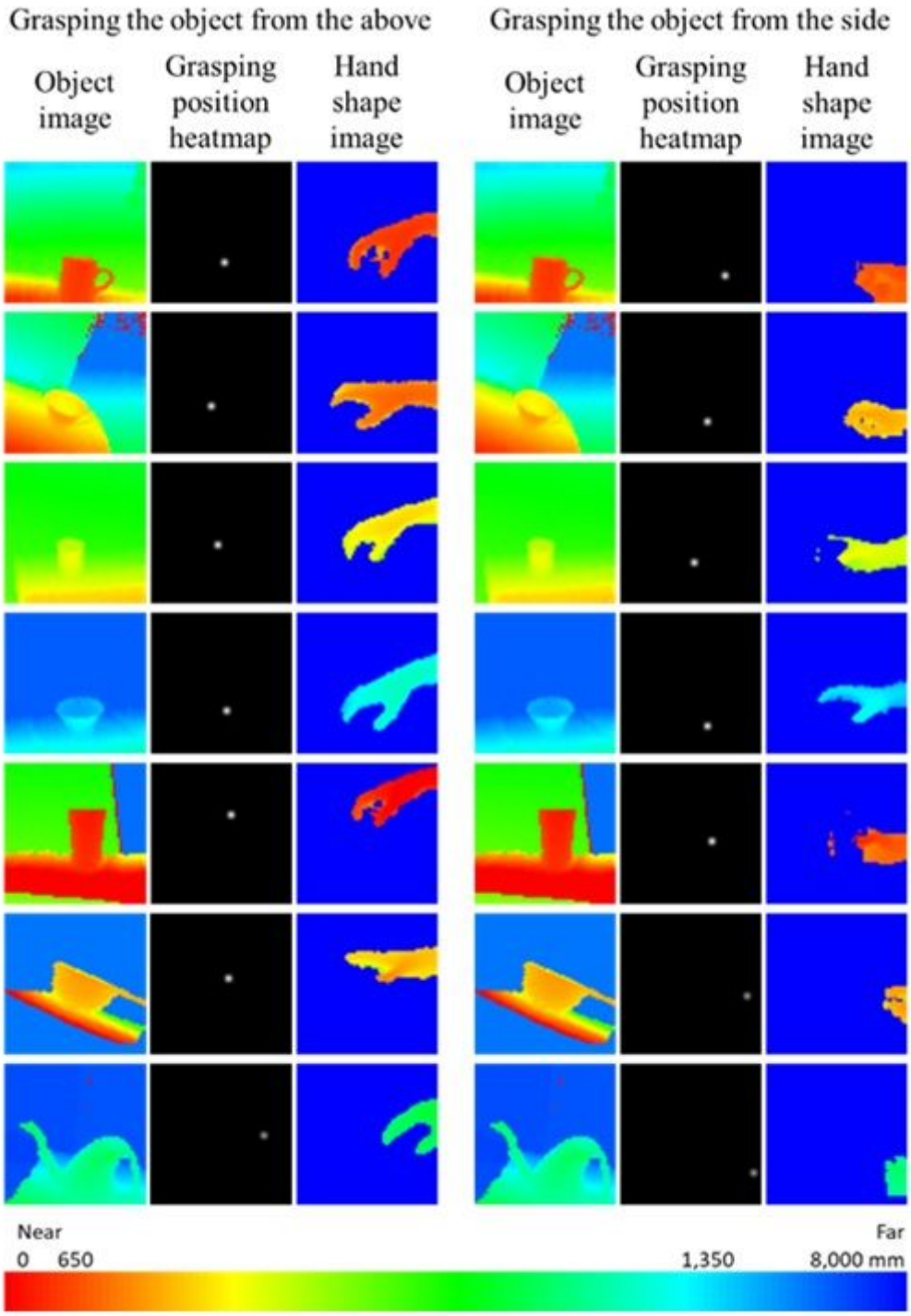
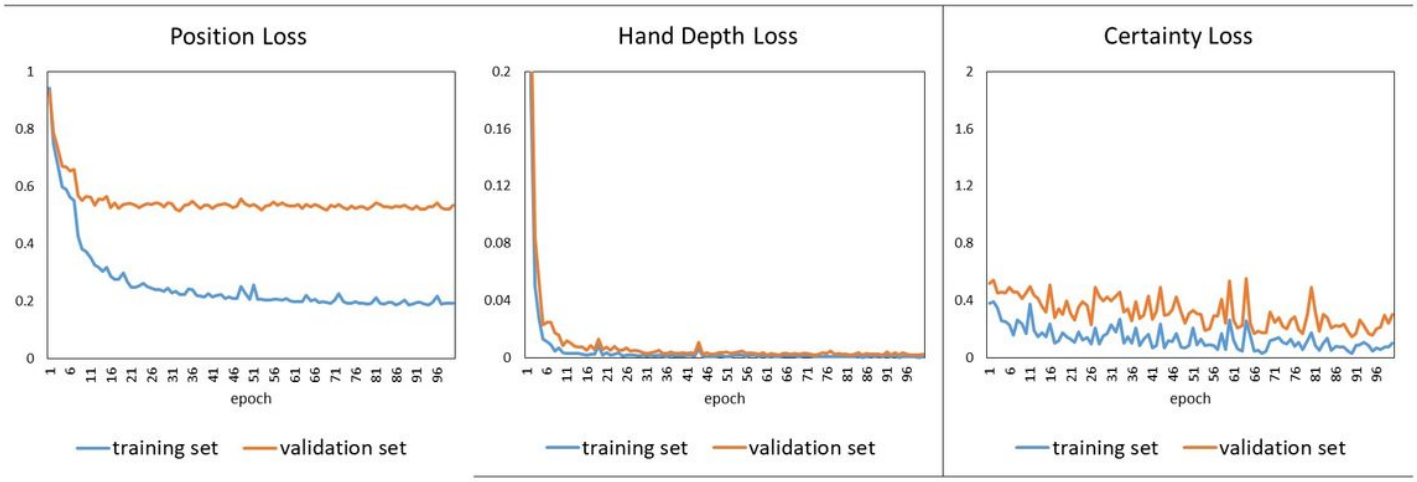


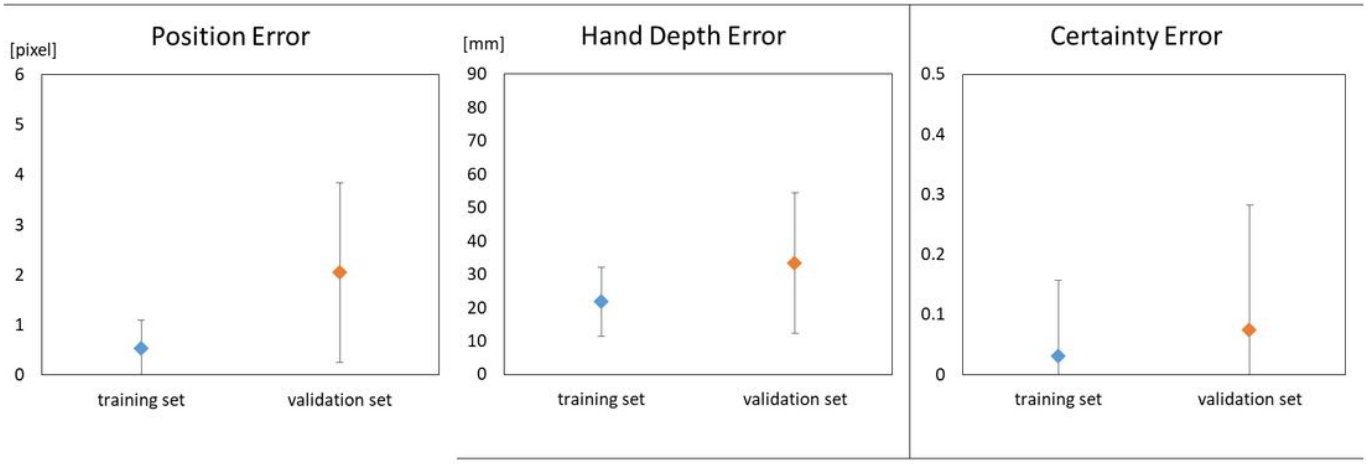

**Figure 7**

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.



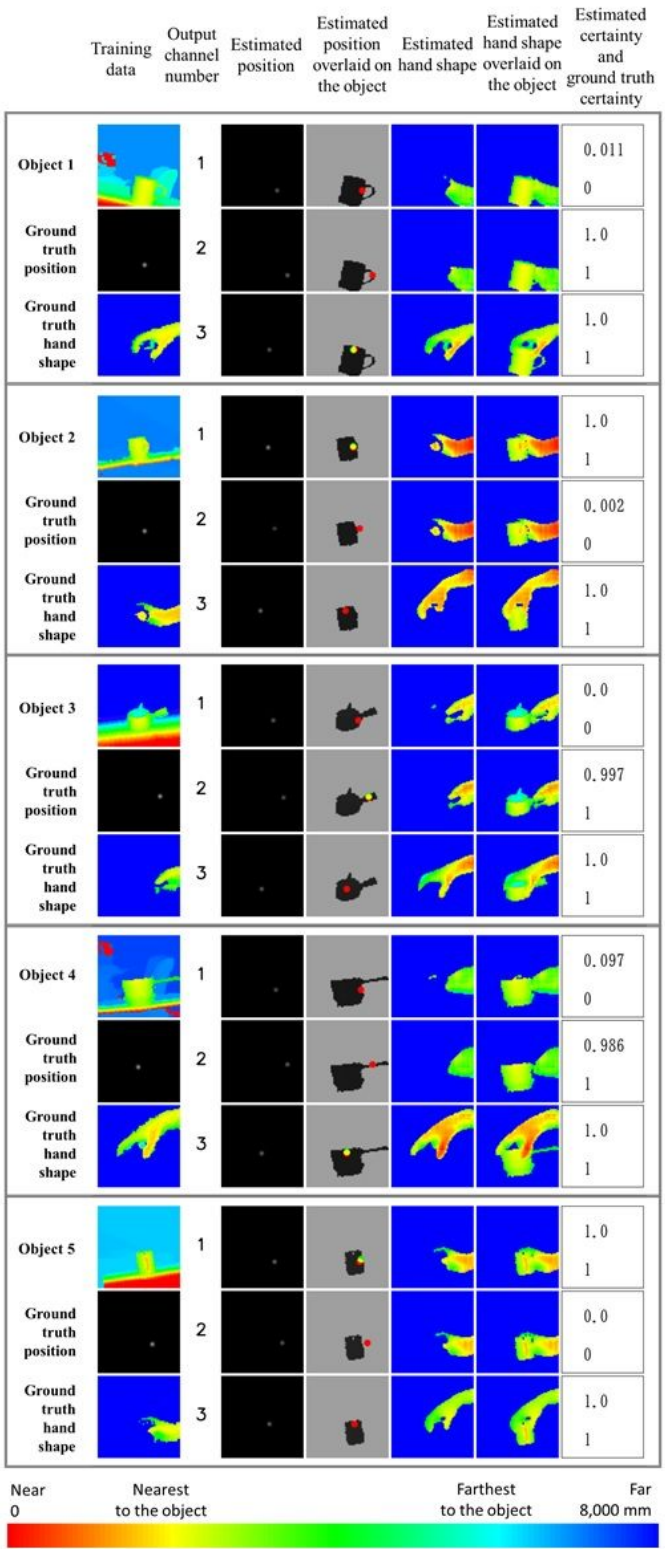

## Figure 8

Loss of the grasping position, the hand shape, and the certainty in each epoch




## Figure 9

Mean error and standard deviation per a sample of the grasping position, the hand depth, and the certainty




## Figure 10

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 in red and green, and the intersection pixels in 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 and the hand shape images encodes the depth values by the “red(near)-blur(far)” colormap.

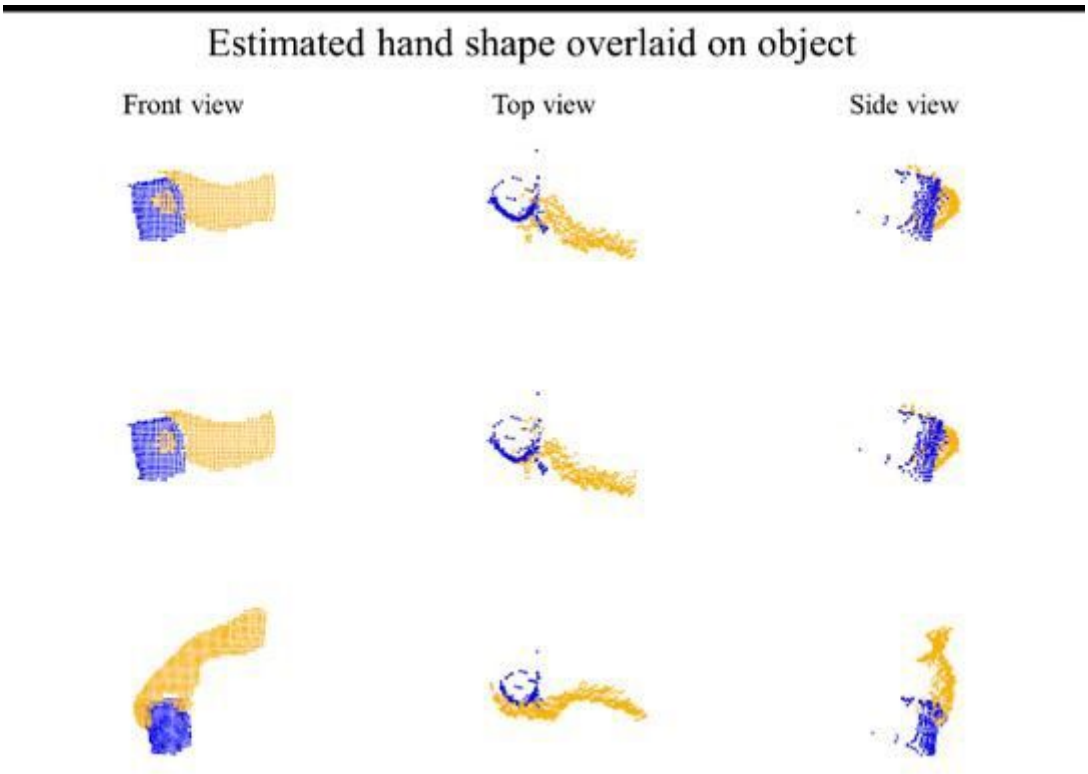
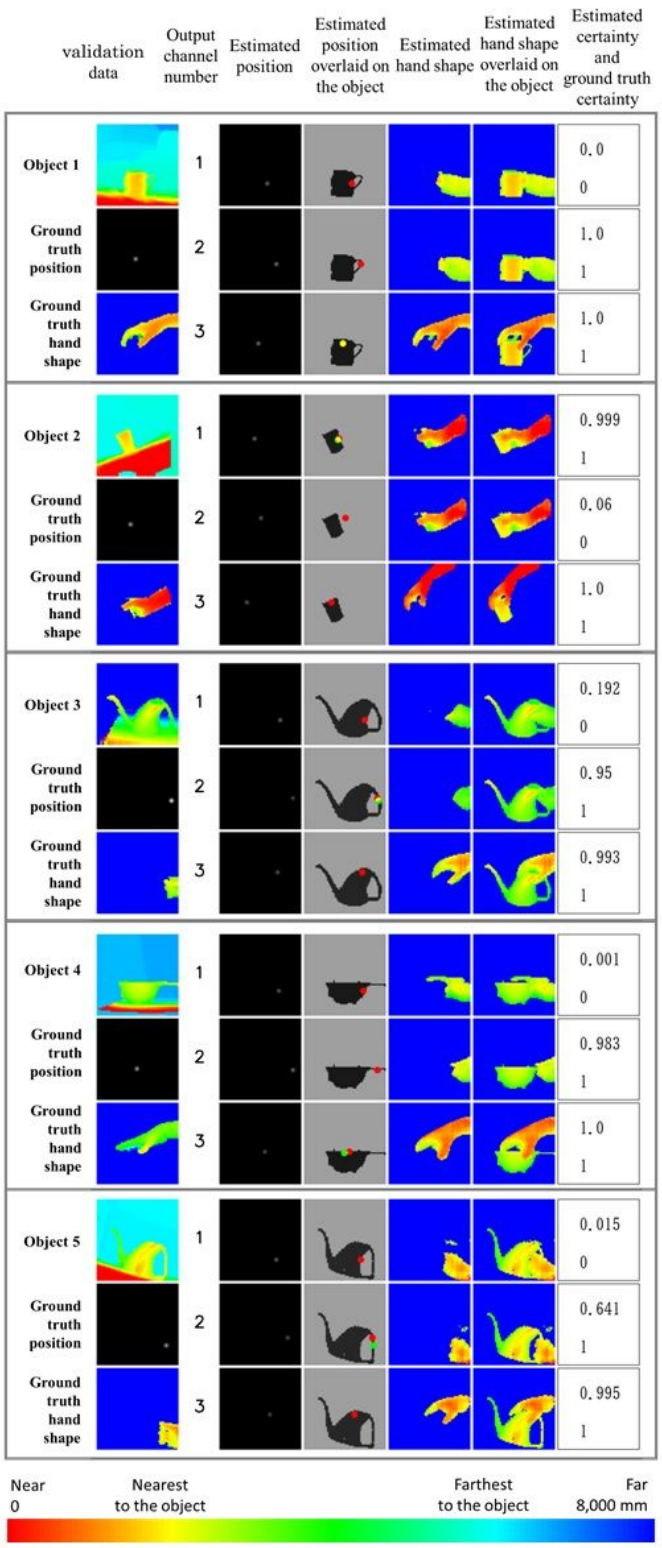



**Figure 11**

Point cloud of the object and recalled grasping hand for the second object in Fig. 10. The object and hand points are blue and yellow, respectively.




## Figure 12

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

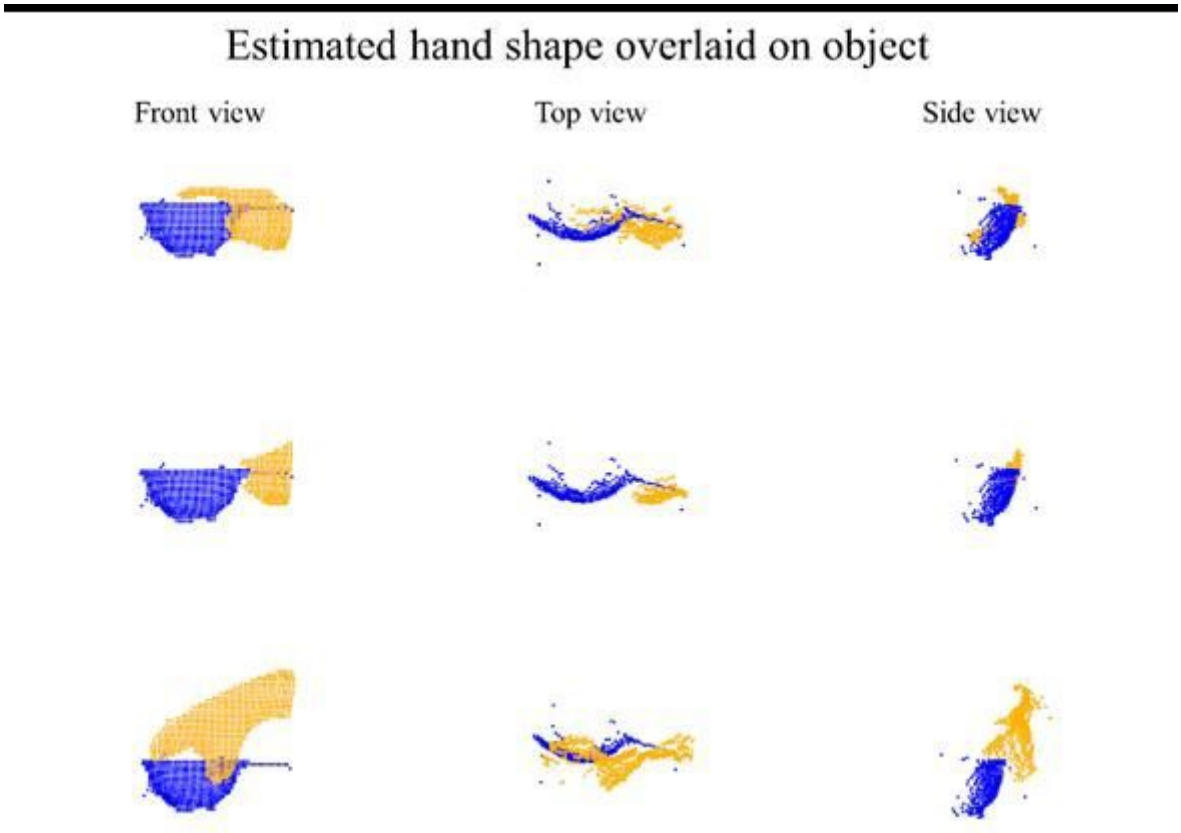



**Figure 13**

Point cloud of the object and recalled grasping hand for the fourth object in Fig. 12

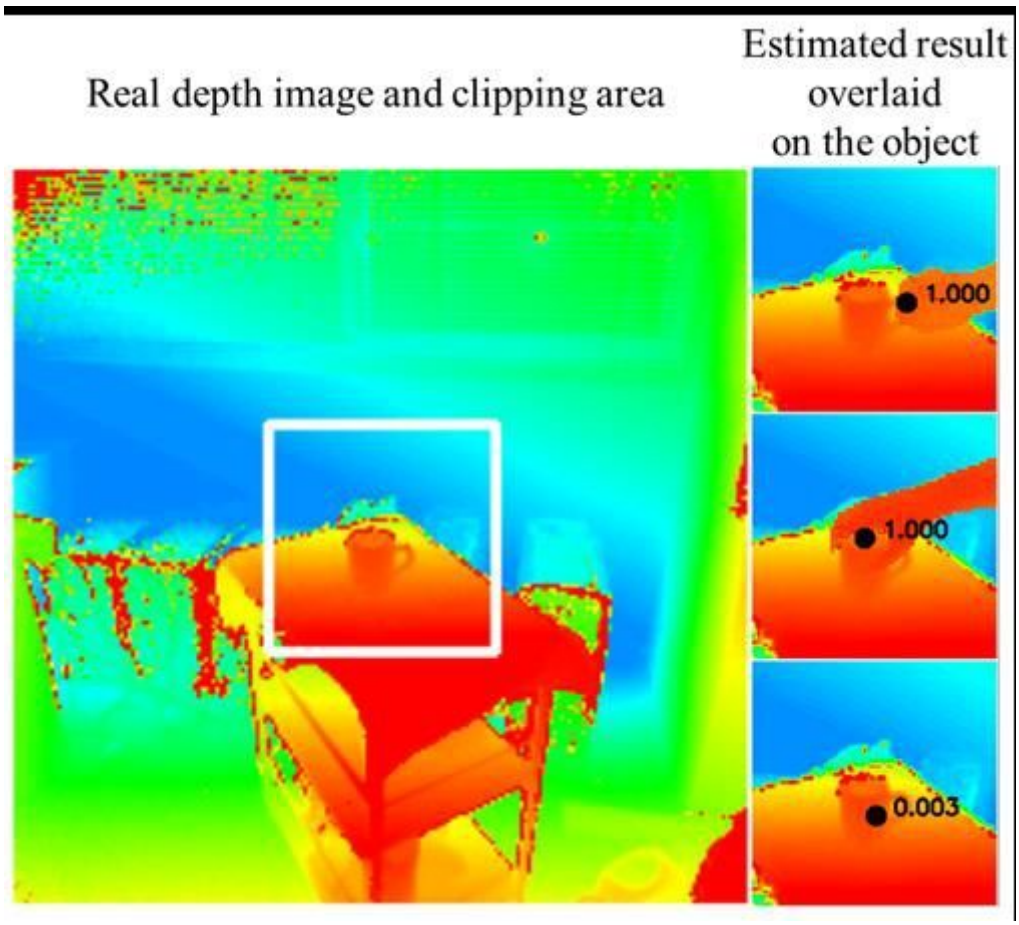



**Figure 14**

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.

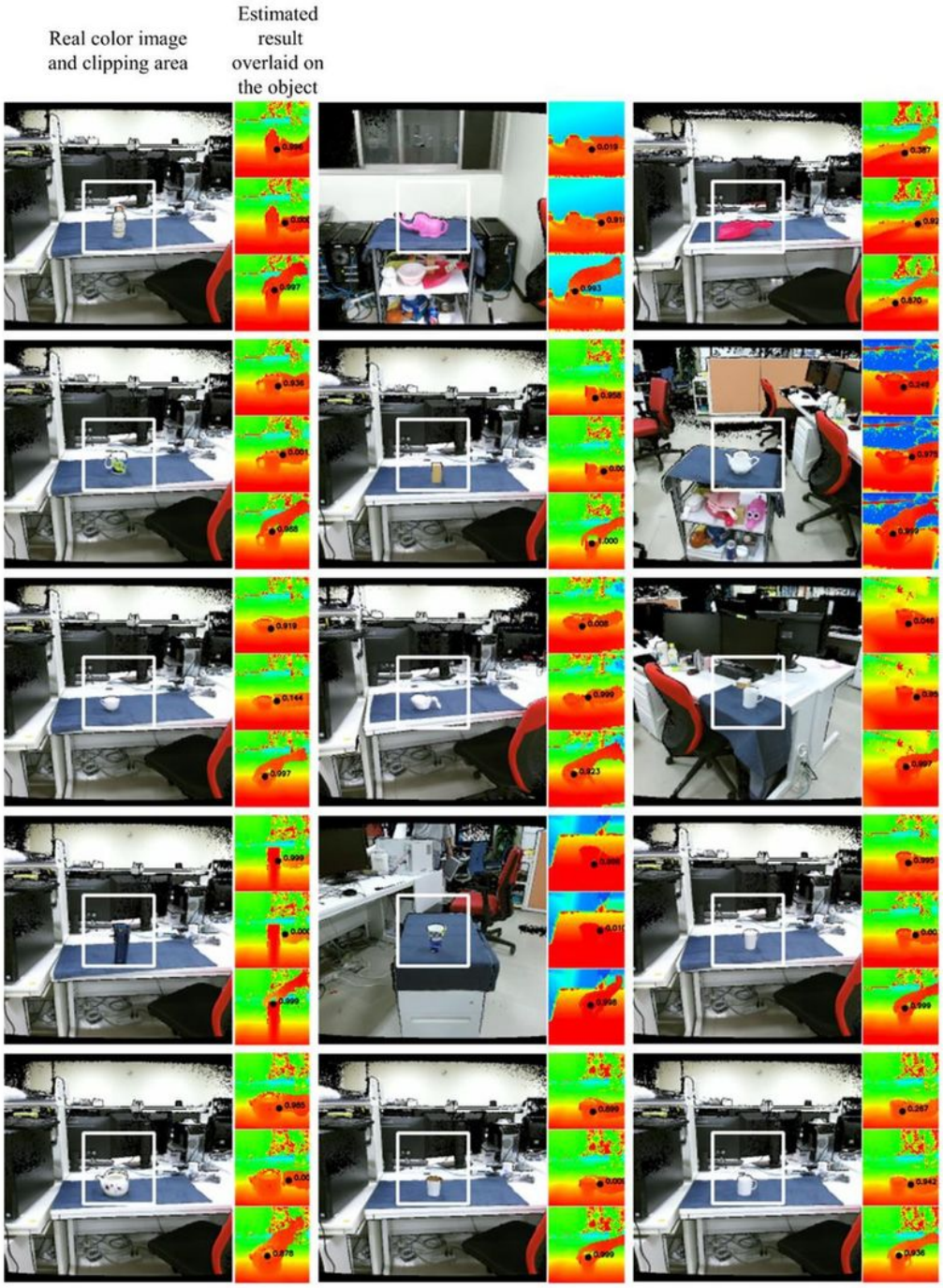



**Figure 15**

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.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.pdf