


RESEARCH ARTICLE

Open Access



Recalling of multiple grasping methods from an object image with a convolutional neural network

Makoto Sanada^{1*} , Tadashi Matsuo^{1,2}, Nobutaka Shimada^{1,2} and Yoshiaki Shirai^{1,2}

Abstract

In this study, a method for a robot to recall multiple grasping methods for a given object is proposed. The aim of this study was for robots to learn grasping methods for new objects by observing the grasping activities of humans in daily life without special instructions. For this setting, only one grasping motion was observed for an object at a time, and it was never known whether other grasping methods were possible for the object, although supervised learning generally requires all possible answers for each training input. The proposed method gives a solution for that learning situations by employing a convolutional neural network with automatic clustering of the observed grasping method. In the proposed method, the grasping methods are clustered during the process of learning of the grasping position. The method first recalls grasping positions and the network estimates the multi-channel heatmap such that each channel heatmap indicates one grasping position, then checks the graspability for each estimated position. Finally, the method recalls the hand shapes based on the estimated grasping position and the object's shape. This paper describes the results of recalling multiple grasping methods and demonstrates the effectiveness of the proposed method.

Keywords: Object grasping, Convolutional neural network, Recall, Grasping method, Clustering

Introduction

Recently, several studies have been conducted on robots grasping objects. For a robot to perform a grasping motion, a massive amount of information is needed, which includes object shape, grasping hand shape, and arm motion information. Providing such a large amount of information can be difficult for users. Therefore, it is desirable for the robot to automatically generate the actions of grasping an object. Considering that the grasping method depends on the succeeding manipulation, it is important to generate multiple patterns of grasping methods.

The final goal of this study is to realize a robot that manipulates objects based on the human's object manipulation. When controlling the pose and motion of the robot hand, the goal was for the robot to mimic the human's grasping based on the recalled grasping shape of human. In this paper, the recall of multiple grasping methods as the previous step of it is investigated. The hand shapes will be used as a clue to determine the posture of the robot hand in the next step.

Past studies have explored a variety of approaches 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. To use the prior database of the grasping type, it is necessary to prepare the object shape and grasping type patterns in advance and define the relationship between them. Nagata et al.

*Correspondence: gr0320ki@ed.ritsumeikan.ac.jp

¹ Graduate School of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan
Full list of author information is available at the end of the article

[2] proposed to approximate an object by shape primitives and find multiple grasping methods that are presented to the user. 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]. Additionally, 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 which recalled only one type of grasping method for an object [5, 6]. However, there are several ways to manipulate the object after grasping. For example, when carrying a cup, the upper part of the cup is grasped, and during drinking, the side of the cup is grasped. Due to the grasping method being 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. This 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 the object and hand interaction are 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 of grasping method. However, only one grasping motion can be observed in a single observation at a time, and it is

not possible to obtain other grasping methods from the observed motion. Therefore, there is a need for a method to solve the problem “there are multiple answers for a single input, but only one of them is given at a time for each training sample”. Multiple grasping methods for the similar object shape is automatically learned by clustering the grasping methods during the learning process.

In this study, a model to obtain and learn multiple grasping methods by automatic clustering the grasping methods through the observation of human's grasping motions is proposed. The model was constructed by cascaded convolutional neural networks (CNNs). Since grasping methods are often determined by the shape of the object part that is indicated by the grasping position, they can be accurately clustered based on the grasping position. The proposed method to recall the grasping method consists of two submodules: one for recalling the grasping position and the other for recalling the grasping hand shape. Therefore, the learning of the grasping method is divided into two steps: learning of the grasping position for the given object shape and learning of the grasping hand shape for the pair of the given object shape and the estimated grasping position. 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 different grasping methods, an approach is needed to determine the graspability of the object and 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 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.

“Methods” section 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. “Experiment” section describes the results of recalling multiple grasping methods while proving the usefulness of this study.

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 probability 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.

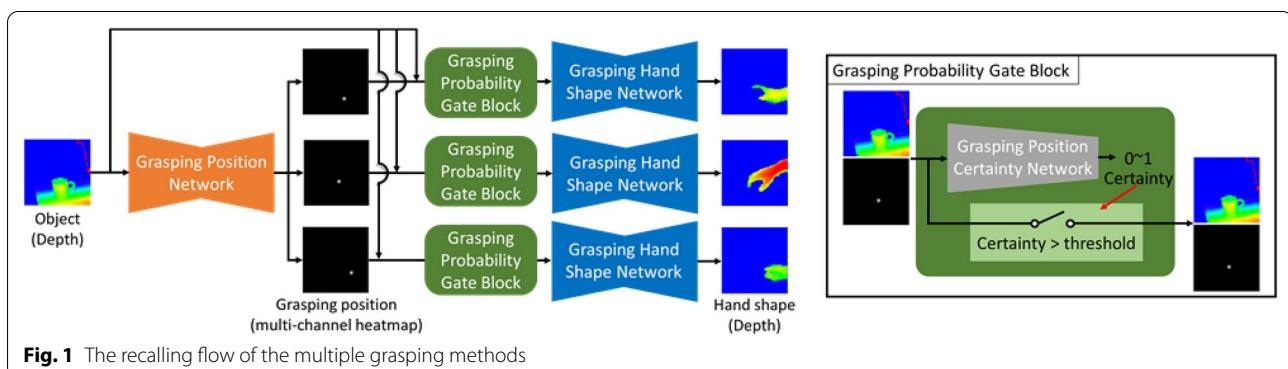
1. The multi-channel heatmap indicates one grasping position for each channel, which is generated by inputting the object depth image into the grasping position network.
2. The certainty for each grasping position candidate is estimated by inputting the combination of the object image and one channel 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 grasping the object 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.

Grasping position network

This network takes an object depth image and outputs the multi-channel heatmap that indicates one grasping position in each channel, as shown in Fig. 2. Each channel represents a typical grasping position cluster for the objects. In the learning setting, the training dataset provides one correct answer for one input. This is because all the training data is 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



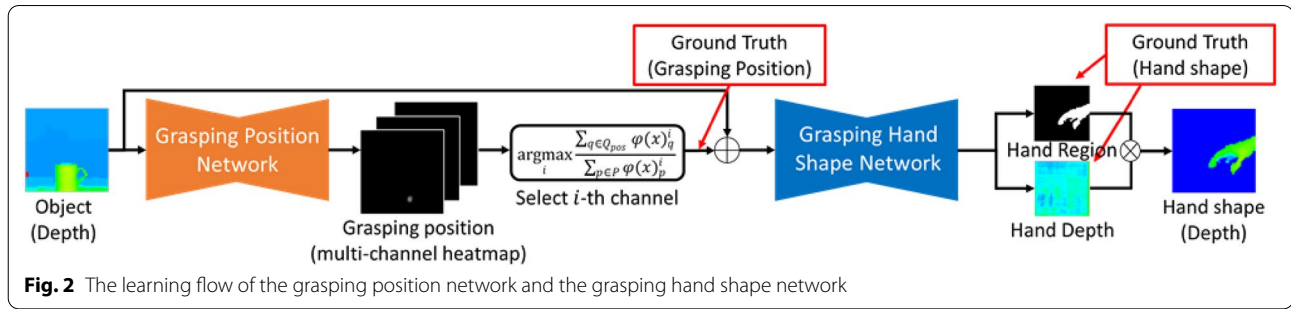


Fig. 2 The learning flow of the grasping position network and the grasping hand shape network

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. Additionally, a constraint is introduced where each channel image for the estimated grasping positions must be as different as possible due to the different types of grasping positions are clustered for each of the channels.

The loss function of this network is presented in Eq. (1).

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

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; and y_{pos} is the ground truth heatmap of the grasping position.

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.

The channel selection method is described in Eq. (2).

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

where: P is a set of all the pixel coordinates for the one-channel heatmap; p and q are coordinate indices; and

$\operatorname{Binary}(\cdot)$ is the function that binarizes the pixel values larger than a threshold to one and it or less to zero.

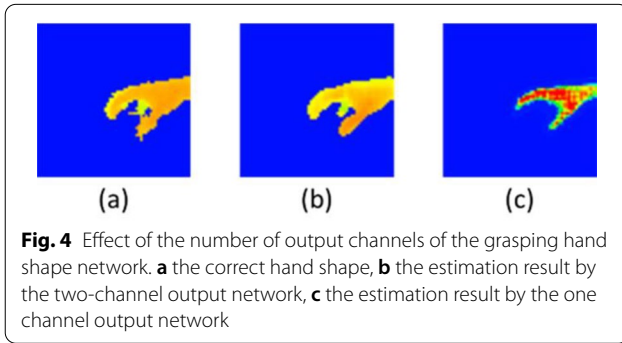
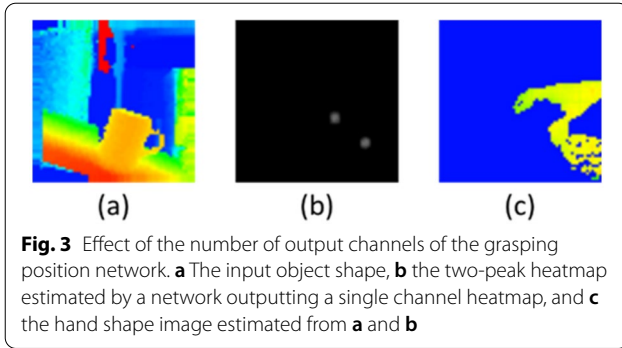
Equation (2) selects the channel that has the largest overlap between the estimated heatmap and the ground truth of grasping position. Once a channel is trained by feeding a ground truth, the channel becomes to generate the heatmap with a single high peak and a small variance for the other similar input object images, whilst other channels still generate a heatmap with low peaks and a large variance. Therefore, when a different grasping position is presented as the ground truth, the other channels tend 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 quickly 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. 3b. 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. 3c. 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.

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. Only the depth values in the correct hand region are given to the corresponding pixels in this channel whilst 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 rather than the hand, and the detail inside the hand tends to be neglected as shown in Fig. 4c. The two-channel representation (depth and mask) can recall the three-dimensional hand shape more accurately as shown in Fig. 4b.

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

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

$Loss_{region}$ is the expression of the cross-entropy loss for the first channel and presented in Eq. (4).

$$Loss_{region} = \frac{1}{n(P)} \sum_{p \in P} \left\{ -y_{handR_p} \log \psi(x, \varphi(x)^k)_{handR_p} - (1 - y_{handR_p}) \log (1 - \psi(x, \varphi(x)^k)_{handR_p}) \right\} \quad (4)$$

where: x is an input object image; $\varphi(x)^k$ is the k th-channel heatmap that is selected by Eq. (2); $\psi(x, \varphi(x)^k)_{handR}$ is the hand region likelihood image that is estimated by the grasping position network $\psi(\cdot)$; y_{handR} is the ground truth of the hand region likelihood; and P is a set of all the pixel coordinates in the hand shape image.

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 and presented in Eq. (5).

$$Loss_{depth} = \frac{1}{n(Q_{hand})} \sum_{q \in Q_{hand}} \frac{\left(\psi(x, \varphi(x)^k)_{handD_q} - y_{handD_q} \right)^2}{\left(y_{handD_q} \right)^2} \quad (5)$$

where: $\psi(x, \varphi(x)^k)_{handD}$ is the hand region depth image that is estimated by the grasping position network $\psi(\cdot)$; y_{handD} is the ground truth of the hand region depth; and Q_{hand} is the set of pixel coordinates in the hand region for the correct hand shape image.

By minimizing this expression, this network learns to estimate the depth values that are closer to the ground truth in the hand region.

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, the learned grasping position network was utilized to learn this network. The training data was classified 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 was also selected in advance. When training this network, a probability of 1 is assigned as

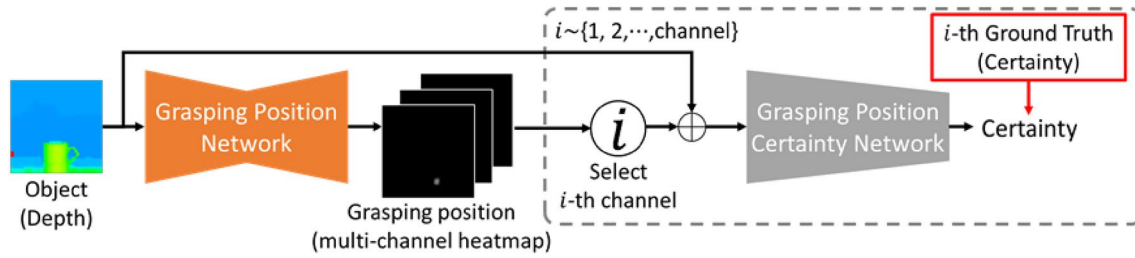


Fig. 5 Learning flow of the grasping position certainty network

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. (6).

$$Loss_{certainty} = \frac{1}{c} \sum_{i=1}^c \left\{ -y_{cert}^i \log \Phi(x, \varphi(x)^i) - (1 - y_{cert}^i) \log (1 - \Phi(x, \varphi(x)^i)) \right\} \quad (6)$$

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.

Experiment

To prove the usefulness of the proposed method, multiple grasping methods were recalled by using networks that learned the grasping methods as described in Sect. 2. In this experiment, the number of output channels of the grasping position network was set to three and the weight parameter, w , in Eq. (1) to one. Given the time it takes to observe the human's motion and to collect the data of the grasping method, the artificial data of the various grasping methods were used for learning.

Structure of networks

This method uses three networks, and each network was designed with referent to the lightweight model of Resnet

Table 1 Details of the structure of each network

Network name	Grasping Position Network	Grasping Hand Shape Network		Grasping Position Certainty Network	
input data	object depth image	object depth image	grasping position heatmap	object depth image	grasping position heatmap
layers	Conv (7x7, 64, 1) Max Pooling (3x3, 2) Residual Block_A (64, 128, 1) Residual Block_B (64, 128) Residual Block_B (64, 128) Residual Block_A (128, 256, 2) Residual Block_B (64, 256) Residual Block_B (64, 256) Residual Block_A (256, 512, 2) Residual Block_B (128, 512) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Conv (3x3, 512, 1) Conv (3x3, 256, 1) Deconv (4x4, 128, 2) Deconv (4x4, 64, 2) Deconv (4x4, 3, 2)	Conv (7x7, 64, 1) Max Pooling (3x3, 2) Residual Block_A (64, 256, 1) Residual Block_B (64, 256) Residual Block_B (64, 256) Residual Block_A (256, 512, 2) Residual Block_B (128, 512) Residual Block_B (128, 512) Residual Block_A (512, 1024, 2) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Conv (3x3, 1024, 1) Conv (3x3, 512, 1)	Conv (7x7, 64, 2) Conv (7x7, 128, 2) Conv (7x7, 32, 1)	Conv (7x7, 64, 1) Max Pooling (3x3, 2) Residual Block_A (64, 256, 1) Residual Block_B (64, 256) Residual Block_B (64, 256) Residual Block_A (256, 512, 2) Residual Block_B (128, 512) Residual Block_B (128, 512) Residual Block_A (512, 1024, 2) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Residual Block_B (256, 1024) Conv (3x3, 1024, 1) Conv (3x3, 512, 1)	Conv (7x7, 64, 2) Conv (7x7, 128, 2) Conv (7x7, 32, 1)
		Concatenation		Concatenation	
		Deconv (4x4, 512, 2)		Fully Connected (2000)	
		Deconv (4x4, 256, 2)		Fully Connected (1)	
		Deconv (4x4, 2, 2)			

[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 points, an object depth image, and a grasping position heatmap, and estimate a hand shape image or a certainty.

Dataset

The dataset consists of the object depth images as the input, 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 at any given time. 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, each training sample was defined 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 required time-consuming efforts, artificial training samples, in which different grasping methods are associated with the same synthesized object images, were employed.

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

The dataset was prepared by capturing 19 types of objects: cups with and without a handle, watering cans, teapots, and containers. These objects had different bottom depths, different sizes, and different shapes such as cylinders or inverted truncated cones and were captured from eight different viewing angles. To increase the variety of object images, objects were randomly translated by shifting them in 25 patterns, rotating them in the range of -20° to 20° , and scaling them in the range of 0.94–1.06. A grasping position heatmap and a hand shape image are also processed by the same transformation as an object image for consistency.

The dataset was divided into a training set, which includes 15 types of objects, and a validation set of the four object types with 640,000 training data points and 1600 validation data points being prepared in total.



Fig. 6 Example of object used in this experiment

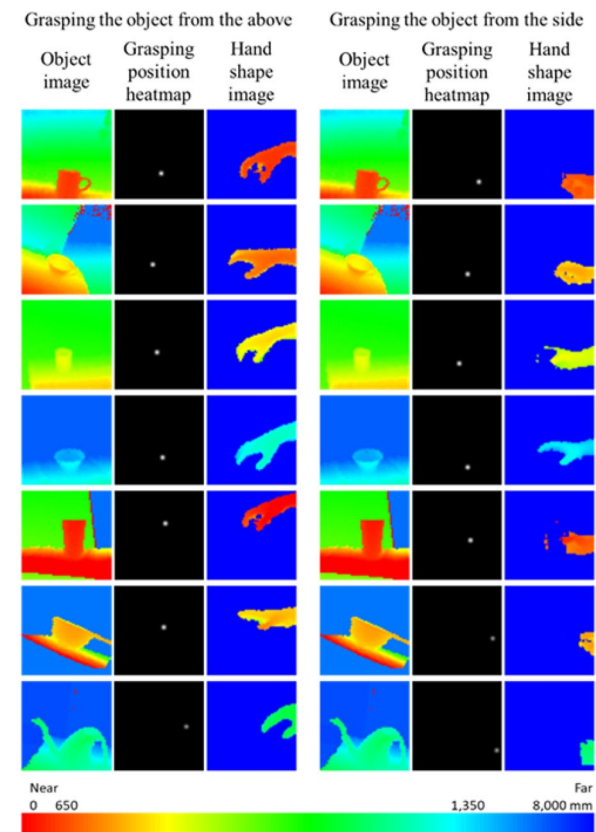
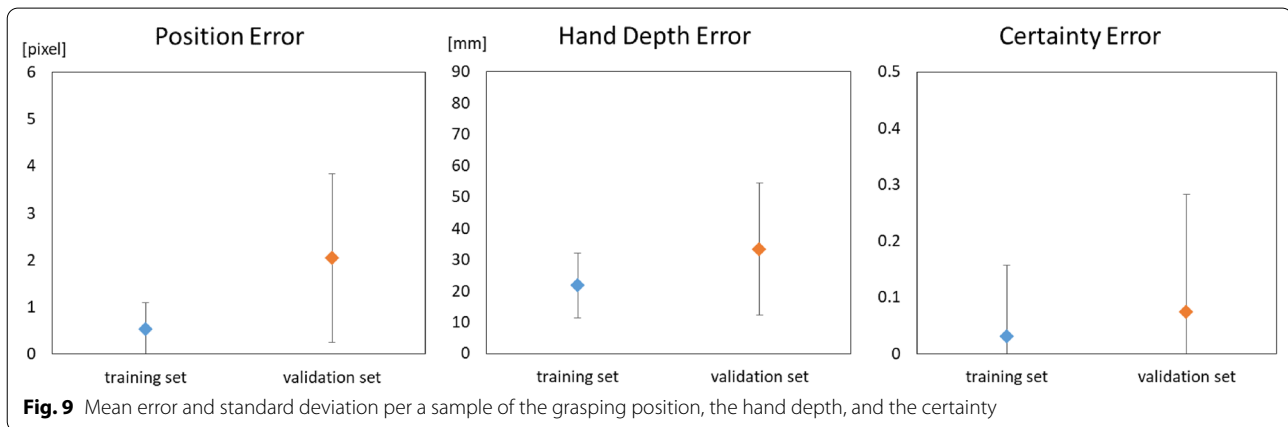
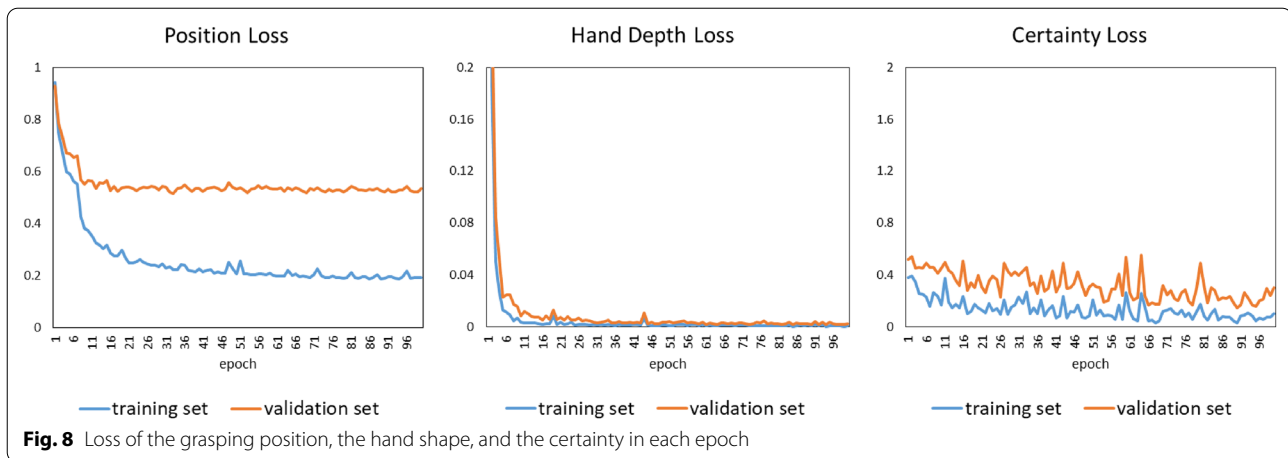


Fig. 7 Examples of the dataset consisting of object depth images, grasping position heatmaps and depth images of the grasping hand shape. Each row demonstrates examples of the different grasping methods for the same object image. The object and hand shape image are displayed by a colormap for the visibility of the depth

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

Results and discussion

Figure 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 losses converge to a constant value as the learning



progresses whilst in Fig. 9, there was no large difference between the mean error of the training set and the validation set.

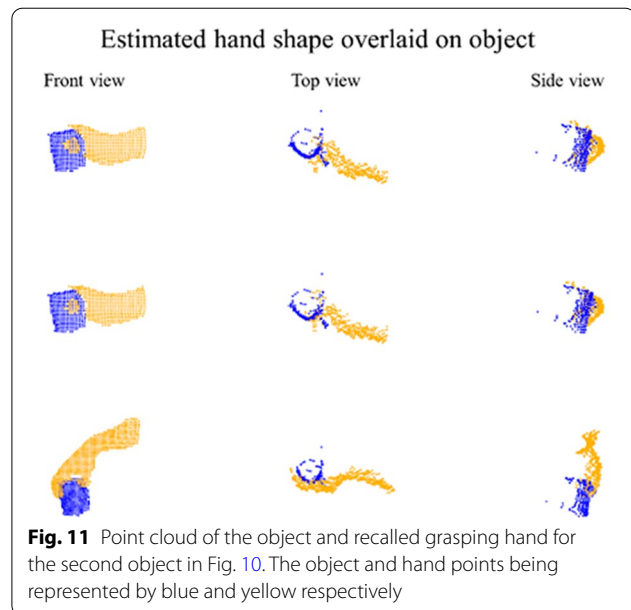
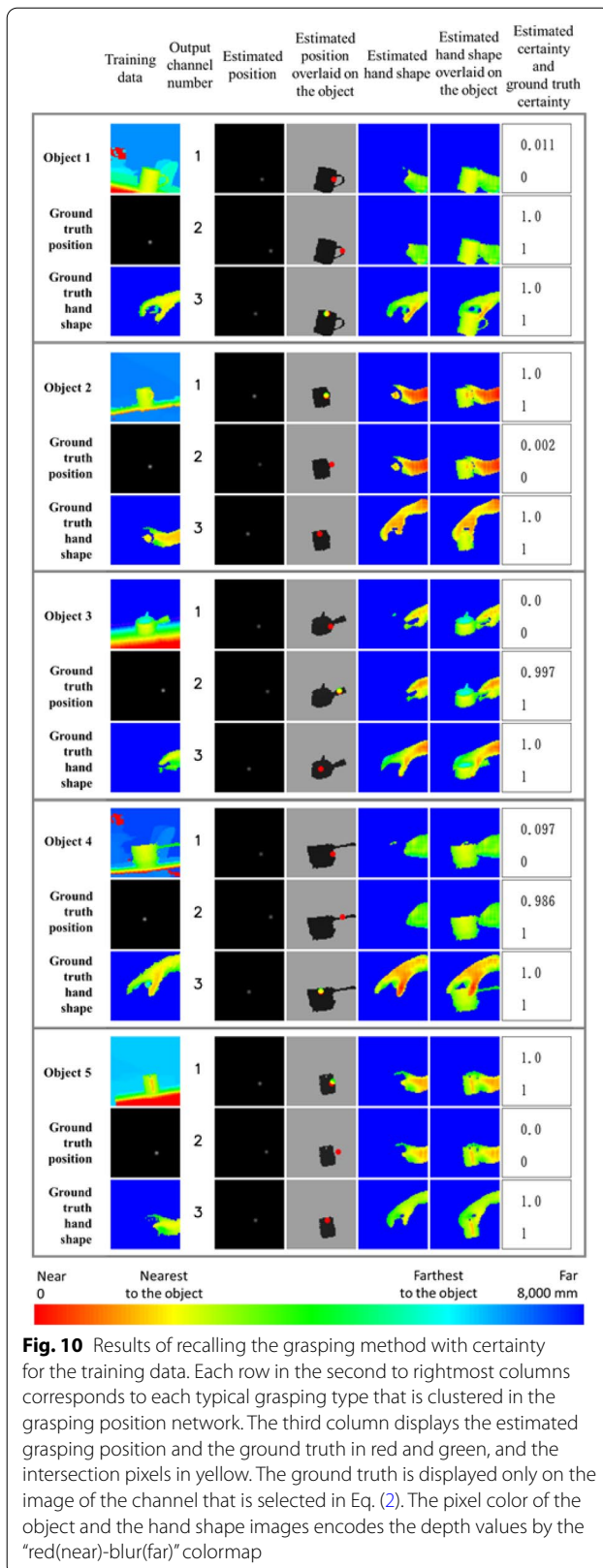
Figure 10 shows the estimated results of the grasping method and its certainty for the training data, which includes five different object shapes. While it may be enough that the hand shape only for the grasping position channel judged as “graspable” (i.e., with high certainty) is obtained, for analyzing the estimation process, the recalled hand shapes for any other grasping position candidates are presented here. Figure 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 was 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 were observed in the training dataset. The grasping positions that were 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, were also recalled by one of the channels because the partial shapes were similar to each other. These grasping positions were judged as having low grasping certainty and then rejected.

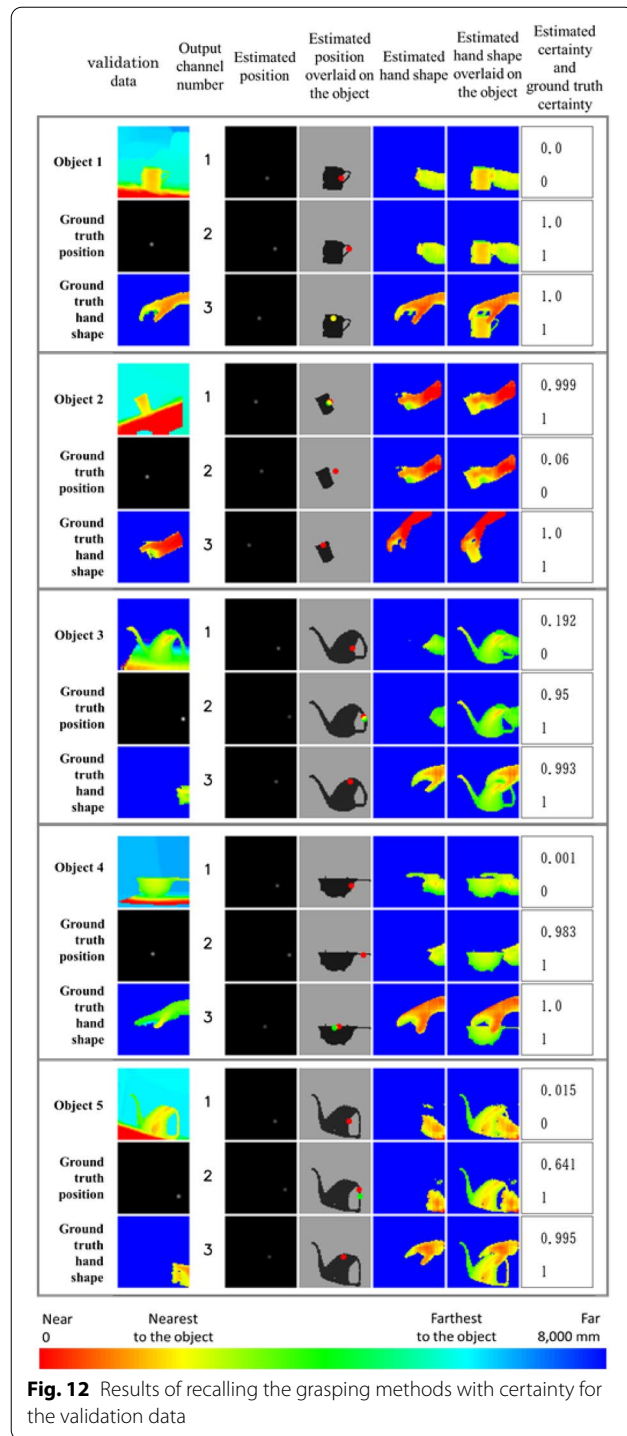
As shown in Fig. 9, the mean error and standard deviation of the estimated grasping position for all the training data were 0.53 pixel and 0.55 pixel. The average object height for the training data was approximately 8 cm and the size projected on the image was approximately 20 pixels, corresponding to 0.2 cm in 3-D space scale. The estimated positions were regarded as near the ground truth with the largest error in the training data being approximately 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, comparable to the size of a fingertip.



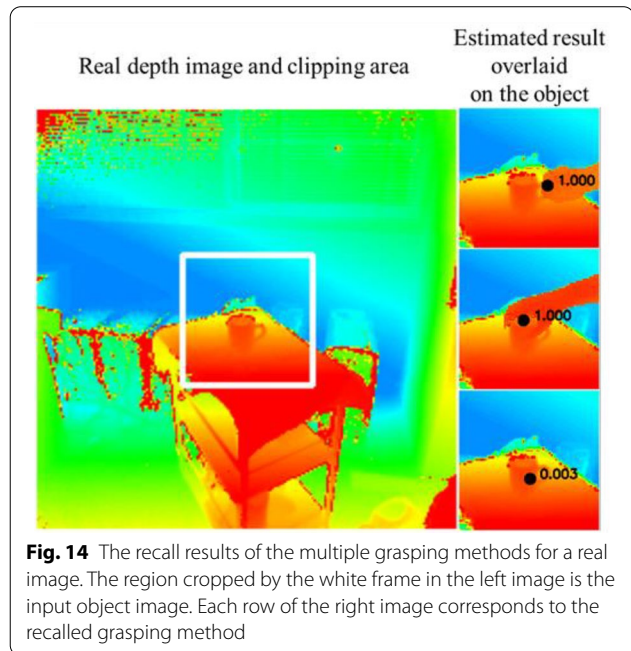
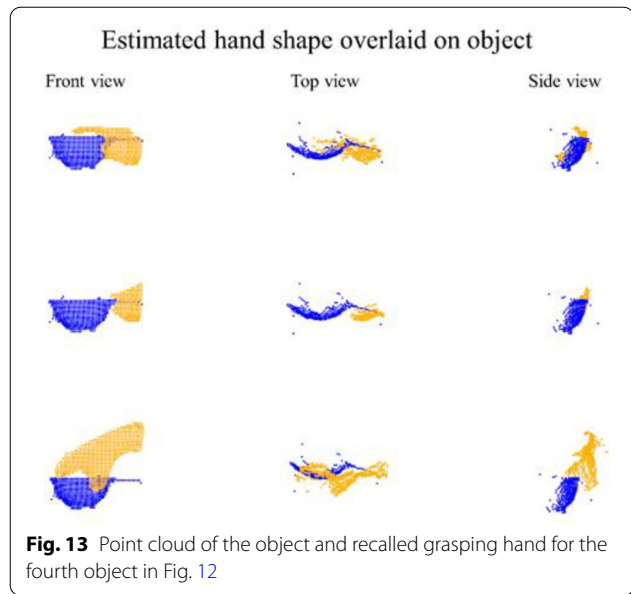
As shown in the fourth column in Fig. 10, the estimated hand shapes for each grasping position that is included in the dataset, such as the second and third rows of the first object, were close to the ground truth. For the grasping positions that were not seen in the dataset, such as the first row of the first object and the second row of the fifth object, the recalled hand shapes were plausible for grasping. However, that grasping is impossible due to the hand being apart from the object or interfering with the object. These ungraspable methods were 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 was 21.7 mm, approximately 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 presented a 30 mm error on average for all the pixels in the hand region, but the pixels around the fingertip had 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, the trained model was



clustered as different grasping patterns when there was a handle or not, and the certainty was estimated appropriately for the object shape and grasping position. As shown in Fig. 9, the grasping position certainty for all the training data was accurately predicted with a mean estimation error of 0.03 and its standard deviation of 0.13. In this study, the grasping is possible for the input object was possible if the certainty is over the threshold value of

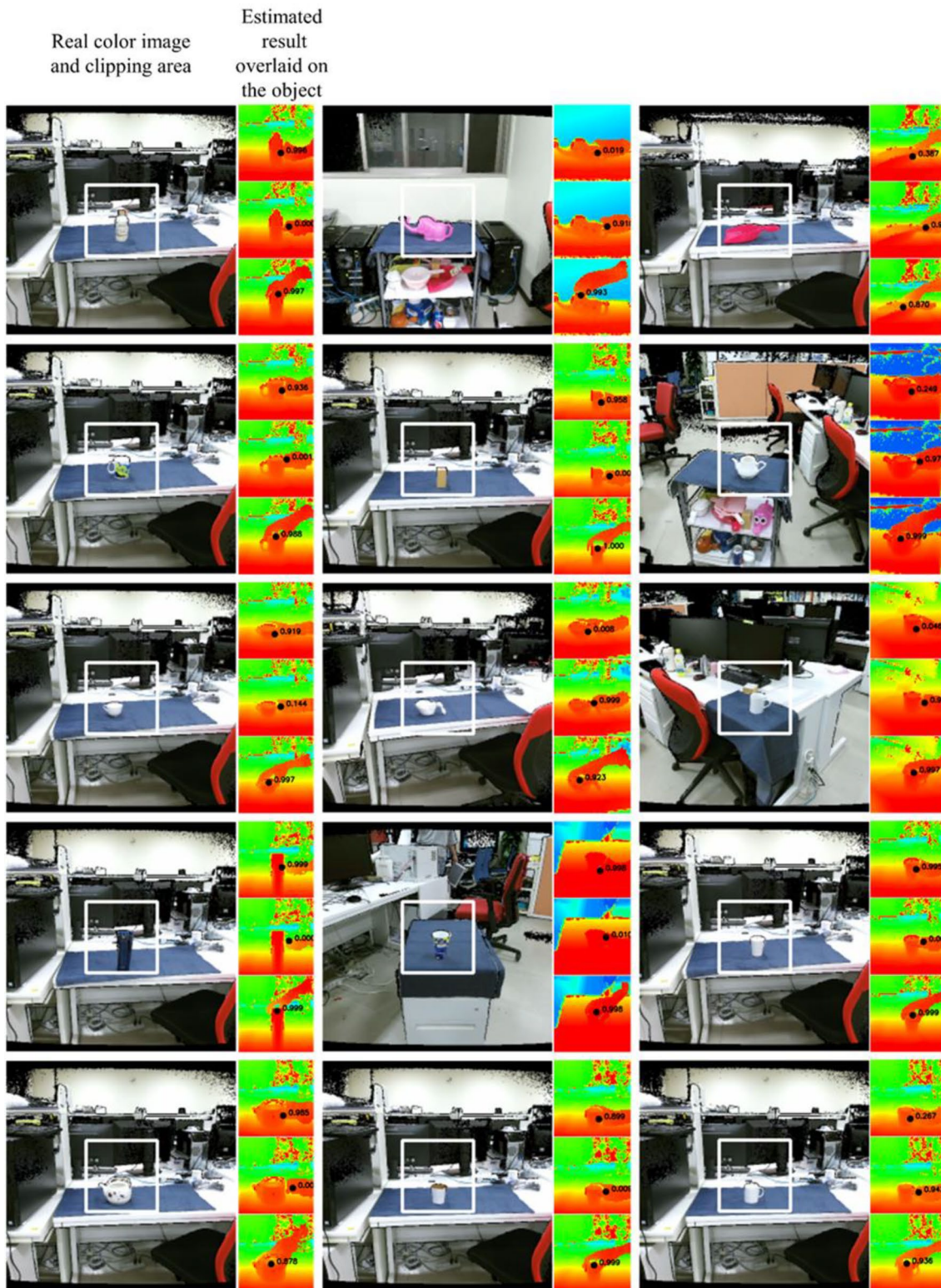


Fig. 15 The recalling results for a variety of real images. The results that are determined to be ungraspable are based on the estimated certainty, which are displayed by the dark images

0.5 for the highest F-value, which gives an accuracy rate of 97.8% and precision of 99.1%.

Figure 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. Figure 13 shows the 3-D point clouds of the object and the recalled grasping hand shapes for the fourth object in Fig. 12.

As shown in the second column in Fig. 12, multiple grasping positions were estimated for the unknown object images and in Fig. 9, the mean error, and its standard deviation of the estimated grasping position for all of the validation data were 2.04 pixel and 1.80 pixel. The average object height for the validation data was approximately 8 cm and the pixel size that was projected on the image was around 20 pixels (corresponding to 0.9 cm in 3-D space scale, which is smaller than the finger width), the estimated positions were considered to be near the ground truth enough. Although there was another estimated grasping with a large position error, as shown in 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 were 33.3 mm 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 occurrence. The mean error and standard deviation of the recalled hand depth except those bad cases was 24.9 mm and 10.5 mm, which was close to the values for the training data. As shown in Fig. 13, since the fingertips were in contact with the object, it was 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 was accurately predicted with a mean estimation error of 0.07 and a standard deviation of 0.20. The grasping position certainty network estimated the certainty for the unknown object images with an accuracy rate of 93.8% and a precision of 96.5% under the condition of threshold of 0.5 at which the F-value is maximized.

Figures 14 and 15 show examples of recall results of the multiple recalled 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 s 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 was overlaid. Figure 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 was recalled for the inside of the handle. The grasping methods were recalled in response to similar local shape features, even for an object having large different shape from the learned object.

Conclusions

This study proposed a method to recall grasping methods for objects having multiple graspable positions and grasping hand shapes. This technique trained 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 were automatically clustered into one of the multi-channel heatmap during learning. In addition, the CNN generated the grasping positions corresponding to the learned typical grasping methods. The plausible grasping methods for the input object were 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 and their certainties were successfully recalled.

Future research should focus on the following points to expand this method:

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.

Acknowledgements

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

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.

Funding

This work was supported by Ritsumeikan Global Innovation Research Organization (R-GIRO) and JSPS KAKENHI Grant Number 18H03313 and JP21K12080.

Availability of data and material

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

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Graduate School of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan. ²College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan.

Received: 19 December 2020 Accepted: 24 June 2021

Published online: 06 July 2021

References

- Ekvall S, Kragic D (2007) Learning and evaluation of the approach vector for automatic grasp generation and planning. *IEEE Int Conf Robot Automat*. <https://doi.org/10.1109/ROBOT.2007.364205>
- Nagata K et al (2010) Picking up an indicated object in a complex environment. *IEEE/RSJ Int Conf Intellig Robot Syst*. <https://doi.org/10.1109/IROS.2010.5651257>
- Chu F, Xu R, Vela PA (2018) Real-world multiobject, multigrasp detection. *IEEE Robot Automat Lett* 3(4):3355–3362. <https://doi.org/10.1109/LRA.2018.2852777>
- Zhang H et al (2019) ROI-based robotic grasp detection for object overlapping scene. *IEEE/RSJ Int Conf Intellig Robot Syst*. <https://doi.org/10.1109/IROS40897.2019.8967869>
- Asif U, Bennamoun M, Sohel FA (2017) RGB-D object recognition and grasp detection using hierarchical cascaded forests. *IEEE Trans Rob* 33(3):547–564. <https://doi.org/10.1109/TRO.2016.2638453>
- Zhang J et al (2020) Robotic grasp detection based on image processing and random forest. *Multimedia Tools Appl* 79:2427–2446. <https://doi.org/10.1007/s11042-019-08302-9>
- Huebner K, Kragic D (2008) Selection of robot pre-grasps using box-based shape approximation. *IEEE/RSJ Int Conf Intellig Robot Syst*. <https://doi.org/10.1109/IROS.2008.4650722>
- Korkmaz S (2018) Training a robotic hand to grasp using reinforcement learning. *ResearchGate*.
- Mueller F et al (2017) Real-time hand tracking under occlusion from an egocentric RGB-D sensor. *IEEE ICCVW*. <https://doi.org/10.1109/ICCVW.2017.82>
- Cao Z et al (2017) Realtime multi-person 2D pose estimation using part affinity fields. *IEEE CVPR*. <https://doi.org/10.1109/CVPR.2017.143>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)