2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Recalling Candidates of Grasping Method from an Object Image using Neural Network

Makoto SANADA, Tadashi MATSUO, Nobutaka SHIMADA, and Yoshiaki SHIRAI, *Member, IEEE*

Abstract— **Robots are required to support people's work. In order to alleviate the burden on people, it is desirable that robot can automatically generate and execute complicated motions according to simple directions from people. However, there are multiple grasping methods for one object. In order to select a motion suitable for the direction, it is important to estimate candidates of grasping method for the object. In this research, we purpose to recall candidates of grasping position and hand shape from an object image. In learning, a network that outputs a plurality of grasping method candidates for one object image to each channel of a multi-channel image is used. At this time, a plurality of grasping methods are not learned at same time, learned one by one. The similar grasping method for the similar object shape is automatically clustered to each output channel in the learning process, and a grasping method having a characteristic difference is presented as a candidate. We show the usefulness of this method using experimental examples.**

I. INTRODUCTION

Currently, robots that support work in people's lives are being developed. Even in support of simple work, it is necessary to give many directions to a robot. For example, in the direction of bringing an object, the robot requires directions such as designation of target object, grasping method (grasping position and hand shape), and motion of robot arm in the state of grasping object. To give these detailed directions to the robot is a heavy burden on the director, there are even possibilities that it is easier for people to work on themselves. It is desirable that robot can generate and execute a suitable motion on its own based on a simple direction from a person. Therefore, it is important to automatically detect the target object, recall and select the grasping position and hand shape suitable for the direction and object shape, and generate the robot arm motion according to the direction. There are many researches to detect objects from image, but there are not many researches to recall grasping methods that match the direction from the object shape. Therefore, in this research we will focus on research to recall multiple grasping methods.

In related researches, the robot selects a grasping method based on pre-prepared primitive object shape information and grasping types suitable the object shape. Ekvall et al. prepared primitive object shape information and grasping types from demonstration data that a human grasps an object. The robot searches and execute a grasping type suitable the object shape from the grasp experience database [1]. Nagata et al. prepared primitive object shape information and grasping types suitable for each object shape [2]. In these researches, it is effective only for object information prepared in advance. A lot of advance data is required to grasp with many object shape. In addition to these researches, there are also researches to acquire object shape information from image and determine a grasping type without requiring advance data. For example, there are methods of determining a suitable grasping method from candidates of grasping methods that are estimated from stereoscopic data of the object [3, 4], and method of estimating the object shape from the appearance of the object and determining the suitable grasping method [5, 6].

Objects often have some grasping methods depending on usage. For example, if you grasp to move the cup, you can grasp the upper part of the cup. However, if you grasp to throw away the water in the cup, the grasping method would be best to grasp the sides or handles of the cup. We consider a mechanism to observe the grasping motion of a person and learn it automatically. In one observation, only one grasping method can be obtained for an object. Therefore, it is necessary to self-acquire that multiple object grasping methods exist in a certain class of object shape by clustering in learning from the case. In this research, we propose a method to recall candidates of multiple grasping position and hand shape from object shape while automatically performing clustering of grasping methods.

II. METHOD

In this research, we construct a convolution neural network (CNN) to estimate a plurality of the grasping position and hand shape recalled from one object image. The construction of the suggested network is shown in Fig.1. The Grasping Method Candidates Recall Network consists of two steps. The first step is the estimation of a plurality of grasping position candidates from an object image. The second step is the estimation of grasping hand shape from an object image and a heat map indicating a grasping position. We call the Grasping Position Estimation Network used network in the first step and the Grasping Hand Shape Estimation Network used network in the second step.

*Research supported by ABC Foundation.

Makoto SANADA is with the Ritsumeikan University Graduate School of Information Science and Engineering, Major of Advanced Information Science and Engineering, 1-1-1, Nojihigashi, Kusatsu city, Shiga, Japan (e-mail: gr0320ki@ed.ritsumei.ac.jp).

Tadashi MATSUO is with the Ritsumeikan University, Faculty of Information Science and Enginering, 1-1-1, Nojihigashi, Kusatsu city, Shiga, Japan

Nobutaka SHIMADA is with the Ritsumeikan University, Faculty of Information Science and Enginering, 1-1-1, Nojihigashi, Kusatsu city, Shiga, Japan (e-mail: shimada@ci.ritsumei.ac.jp).

Yoshiaki SHIRAI is with the Ritsumeikan University, Faculty of Information Science and Enginering, 1-1-1, Nojihigashi, Kusatsu city, Shiga, Japan.

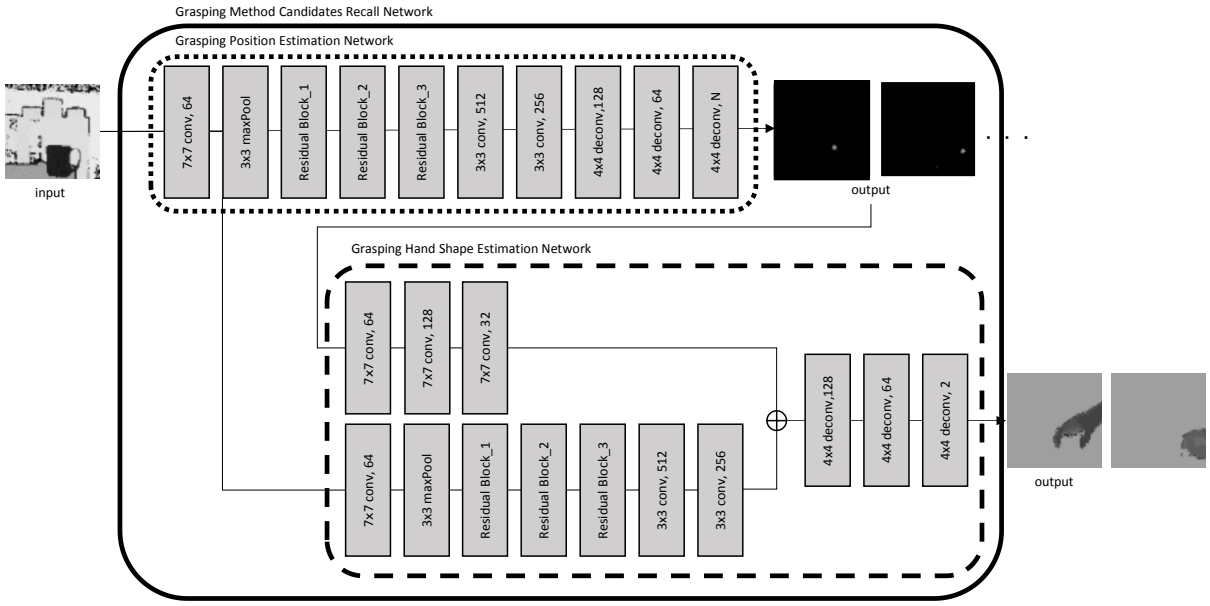

Figure 1 The construction of the suggested network

A. *Grasping Position Estimation Network*

The Grasping Position Estimation Network outputs a multi-channel heat map indicating candidates of grasping position with an object image as input (dotted line frame in Fig.1). The input object image is a 16bit depth image. For each channel, the heat map indicating a grasping position candidate is outputted one by one. In learning, when there are a plurality of grasping position candidates for one object image, we do not give all grasping positions for an object as ground truth, but only one candidate is selected and given as ground truth. With this learning method, it is unnecessary to simultaneously give all possible grasping position candidates at the time of learning, and the grasping position is automatically clustered. As a result, similar grasping position candidates are aggregated, and only completely different grasping position candidates are output to the different channels of heat map. This network learns to minimize the equation (1).

$$Loss_{posi} = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\sum_p \left(\phi(x)_{i,p} + \phi(x)_{j,p}\right)^2}{\sum_p \left(\phi(x)_{i,p} - \phi(x)_{j,p}\right)^2 + \varepsilon} + \frac{\sum_p \left(\phi(x)_{m,p} - y_p\right)^2}{\sum_p \left(\phi(x)_{m,p} + y_p\right)^2} \quad (1)$$

$$m = \arg\min_k \frac{1}{p}\sum_p \left(\phi(x)_{k,p} - y_p\right)^2$$

x is the object image as input, $\varphi(x)$ is the multi-channel heat map as output, y is the heat map indicating correct grasping position as ground truth, i, j and k are the channel indices of multi-channel heat map, p is the pixel index of a heat map, and N is the number of output channel. The first term is an equation that minimizes the inverse of the difference between each channel of the multi-channel heat map to maximize the difference. The inverse is normalized with the sum between channels to calculate the relative difference between the channels. This term restricts the multi-channel heat map to output different grasping position for each channel. If the output heat map gathers evenly on each channel as shown in Fig.2 a), and similar results are output on all channels, clustering of similar grasping positions cannot be expected. We expect to output the estimated grasping positions separately for each channel with the multi-channel heat map as shown in Fig.2 b). The second term is the expression for minimizing the squared error between the ground truth and the selected channel of the multi-heat map. It is normalized by the sum between channels to minimize the relative error. It works to bring the channel closest to the ground truth in the output heat map closer to the ground truth.

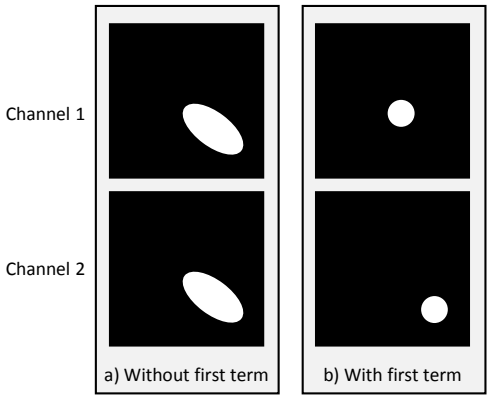

Figure 2 The effect of first term in the equation (1)

In this research, the output of the Grasping Position Estimation Network was set as the two channel heat map.

B. *Grasping Hand Shape Estimation Network*

The Grasping Hand Shape Estimation Network outputs the grasping hand shape image by inputting an object image and a one channel heat map indicating the grasping position for the input object image (broken line frame in Fig.1). The input object image of this network is common to the input of the Grasping Position Estimation Network, and the input heat map is the output of the Grasping Position Estimation Network. The output hand shape image is a 16bit depth image. This network learns to minimize the equation (2).

$$Loss_{shape} = \frac{\sum_p \left(\psi\left(x, I_{posi}\right)_p - y_p \right)^2}{\sum_p \left(\psi\left(x, I_{posi}\right)_p + y_p \right)^2} \quad (2)$$

x is the object image as input, I_{posi} is the one channel heat map as input, $\psi(x, I_{posi})$ is the grasping hand shape image as output, y is the grasping hand shape image as ground truth, and p is the pixel index of a grasping hand shape image. $Loss_{shape}$ is the expression for minimizing the squared error between the ground truth and the estimated hand shape image. This equation is normalized as in the equation (1).

C. *Grasping Method Candidates Recall Network*

The Grasping Method Candidates Recall Network is a network connected the Grasping Position Estimation Network and Grasping Hand Shape Estimation Network. The input of the Grasping Position Estimation Network part is an object image, and the input of the Grasping Hand Shape Estimation Network part is the object image and the grasping position heat map. The object image as input is common with the Grasping Position Estimation Network. The grasping position heat map is one channel of multi-channel heat map which is output of the Grasping Position Estimation Network. At learning time, one channel heat map closest to the ground truth among the multi-channel heat map is input to the Grasping Hand Shape Estimation Network. Equation (3) shows the loss function of the Grasping Method Candidates Recall Network.

$$Loss = Loss_{posi} + Loss_{shape} \quad (3)$$

D. *Dataset*

The dataset used for learning consists of object images, heat maps showing the grasping position of the object, and hand shape images when the object is grasped at the position of the heat map. Object images and hand shape images are a 16bit depth image acquired using Kinect for Windows depth camera. Heat maps are an 8bit Gaussian image. In this dataset, 6,400 pairs of images are prepared, one set consisting of an object image, a heat map showing the grasping position of the object, and a hand shape image grasping the object at grasping position shown by the heat map. When there is a plurality of grasping positions for one object image, a pair created in which another heat map and hand shape image are combined with respect to the same object image.

In this dataset, sixteen kinds of object data of eight type cups and eight type mugs are prepared. Each object is taken from eight viewpoints, and 25 patterns of images for each viewpoint are prepared by shifting the object position in the image. In addition, a rotation of -20 to 20 degrees was randomly given to each object image. The background of the object image is randomly synthesized from a prepared background image. Furthermore, two kinds of grasping position heat maps and grasping hand shape images are prepared for each object image. A part of the dataset is shown in Fig.3. The first and fourth columns are the object image, the second and fifth columns are the heat map indicating the grasping position, and the third and sixth columns are the hand shape image grasping the object at the directed grasping position for each object image. The data pair in the left column shows the grasping position and the grasping hand shape when

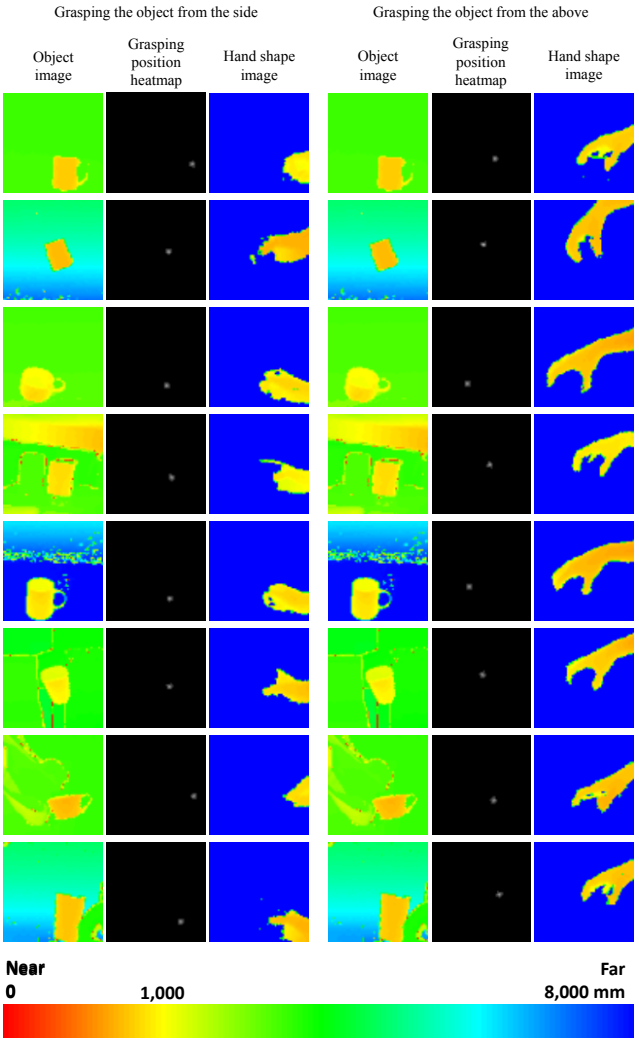

Figure 3 A part of the dataset and the color scale for depth value

grasping the object from the side and the data pair in the right column shows the grasping position and the grasping hand shape when grasping the object from above. Although the object image and the hand shape image are originally 16bit gray scale images, they are displayed as color map in order to make the depth change visually easier to understand. The scale of color with respect to the depth value is shown in the lower part of Fig.3. The data pairs of the first row in Fig.3 are a pair of data having the same object image and different grasping positions and hand shapes. In sixteen types of object data, 800 pair of each one piece of cup and mug object image data are used as evaluation data and the remaining 5,600 pair data are used as training data.

III. RESULTS AND DISCUSSIONS

A part of the estimation result of the Grasping Method Candidates Recall Network using training data is shown in Fig.4. It shows the results of the eight types input images. The first column shows the input object image, the second column shows the estimation heat map of grasping position and the third column shows the ground truth of grasping position. The fourth column is the synthesized image of the input object image excluding the background, the estimated heat map of

grasping position and the ground truth of grasping position. In the synthesized image, the estimated grasping position is displayed in blue and the correct grasping position is displayed in red. In addition, pixels where the estimated grasping position and the correct grasping position overlap are displayed in yellow. The fifth column shows the grasping hand shape image estimated from the object image of the first column and the estimated grasping position heat map of the second column. The sixth column shows the ground truth of grasping hand shape, the seventh column shows the color depth map which synthesizes the estimated hand shape image and the object image excluding the background, the eighth column shows the color depth map which synthesizes the ground truth of grasping hand shape and the object image excluding the background. In the synthesized depth map, the low pixel value is the closer to red and the high pixel value is the closer to blue. The row with the object image and the row below it are paired, the upper row is the result of channel 1 of output multi-channel heat map in the Grasping Position Estimation Network, and the lower row is the result of channel 2.

Looking at the estimated grasping position heat map in the second column, heat maps close to the correct of the third column are output in all results. Looking at the synthesized results of the grasping position in the fourth column, although there is a shift of several pixels between the correct grasping position and the estimated grasping position, yellow pixels indicating overlapping between the two grasping positions are displayed with all results. These results were also confirmed in all training data of 5,600 pairs. The mean and standard deviation of the distance between the maximum value coordination in ground truth and the maximum value coordination in estimated heat map are 0.21 ± 0.71 pixel. It is possible to estimate the grasping position close to the correct grasping position with high accuracy. In addition, it can be seen that it is possible to estimate the grasping position without being affected by the difference of viewpoint and object shape from the estimation results of various input images such as the first column.

Looking at the channel 1 heat map and the channel 2 heat map, the two-channel estimated heat map of each input object image shows different grasping positions for each channel heat map. It is considered that this is the result of similar grasping positions being learned on the same channel and grasping positions with distinctive differences being learned on different channel. From these results, it can be seen that the clustering of the grasping positions is automatically performed in the learning process.

Looking at the hand shape estimation result in the fifth column, grasping hand shape images close to the ground truth shown in the sixth column are obtained with many results. Also, even if a hand shape image different from the ground truth such as the seventh row is estimated, it can be seen that the hand shape grasping the object is estimated by looking at the seventh column. In the estimation results of the object image taken from the obliquely above viewpoint, the hand shape grasping the object from above may be estimated to be a hand shape close to the ground truth such as the fifteenth row, but there is a tendency to estimate the grasping position slightly above the object such as the three and seventh rows.

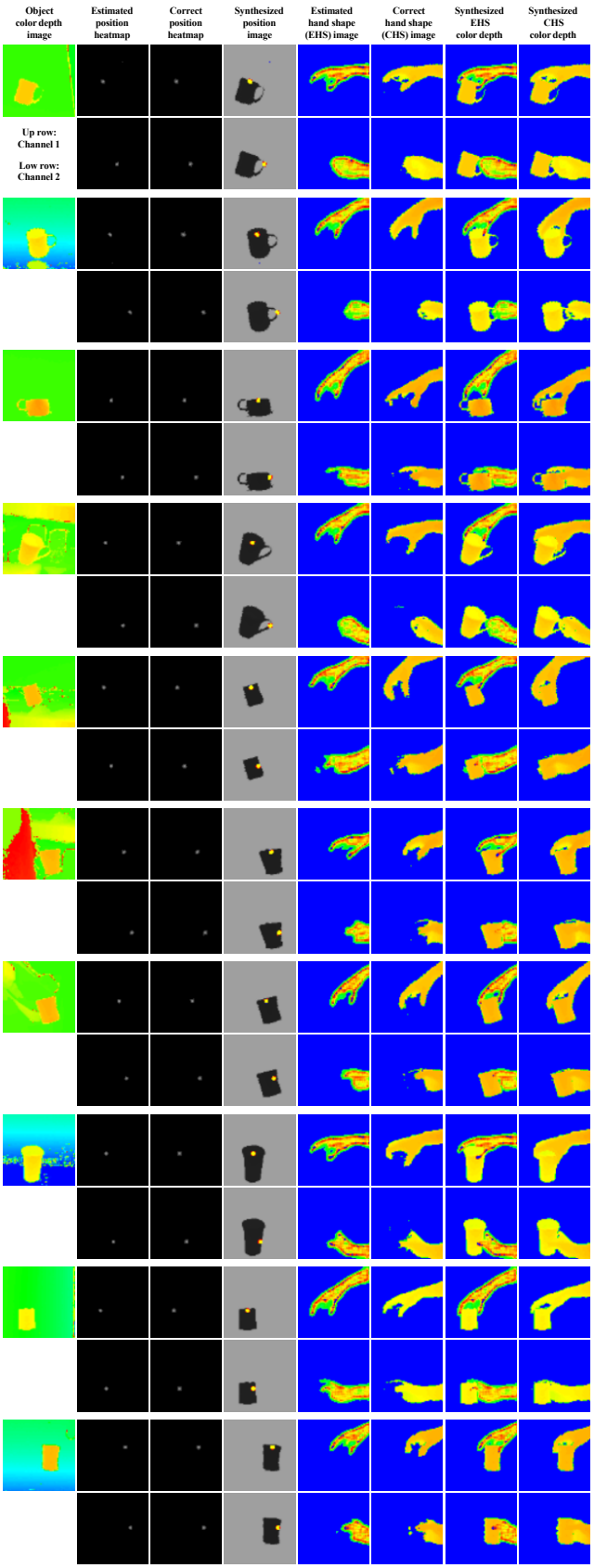

Figure 4 The estimation results of the Grasping Method Candidates Recall Network using training data

The mean and standard deviation of the error between the depth value of the hand region coordinates in the ground truth and the depth value of the same region coordinates in the estimated hand shape image are 127.3±26.3 mm/pixel.

Looking at the synthesized depth map of object and hand shape in the seventh and eighth columns, the pixel values indicating the depth of the synthesized depth map of the estimated hand shape have variation, as compared with the synthesized depth map of the correct hand shape. However, in the synthesized depth map of the second and fourth rows, the place indicating the low pixel value like thumb in the correct synthesized heat map is represented by a color close to red in the estimated synthesized depth map and the place indicating the high pixel value like palm in the correct synthesized depth map is represented by a color close to green in the estimated synthesized depth map. Although there are variations in pixel-wise, the overall trend of the estimated hand shape image is estimated to be close to the correct hand shape image.

A part of the estimation result of the Grasping Method Candidates Recall Network using evaluation data is shown in Fig.5. Fig.5 shows in the same arrangement of data as Fig.4.

Looking at the estimated grasping position heat map in the second column, most of the results show the grasping position similar to the ground truth as heat map of the third column. However, a heat map containing tow grasping positions such as the sixth row was sometimes output. The mean and standard deviation of the distance are 1.61±1.25 pixels. This network can estimate the high accuracy heat map.

The synthesized results of the grasping position in the fourth column display yellow pixels indicating overlap between the correct grasping position and the estimated grasping position with about 96% of all evaluation data. About 3% of the result of estimating the grasping position different from the ground truth estimates the heat map that the ground truth and the estimated grasping position are next to each other like the eighth row. However, the estimated grasping position different from the ground truth also indicates the side of the object, which is not considered to be a problem. As for the result of the remaining about 1%, the grasping positions indicating the handle portion and the side portion are mixed like the sixth row. The reason why the mixed heat map was estimated is that the dataset used for learning included two grasping method: 1) grasping from the top and 2) grasping from the side. In grasping from the side, when the handle is on the right side of object, the handle portion is the correct grasping position, and when the handle is not on the right side of object, the side portion is the correct grasping position. Therefore, large variations occurred in the grasping position when grasping from the side. In other words, clustering in two clusters was applied to dataset having substantially three kinds of grasping position candidates. Because of this, it is considered that a heat map in which mixed two grasping positions in one channel of output heat map was estimated. As a countermeasure, it is considered that a network is needed to increase the number of clusters or to automatically perform clustering while estimating the appropriate number of clusters.

Even in the evaluation data, the difference of the object shape and the viewpoint did not affect the grasping position estimation.

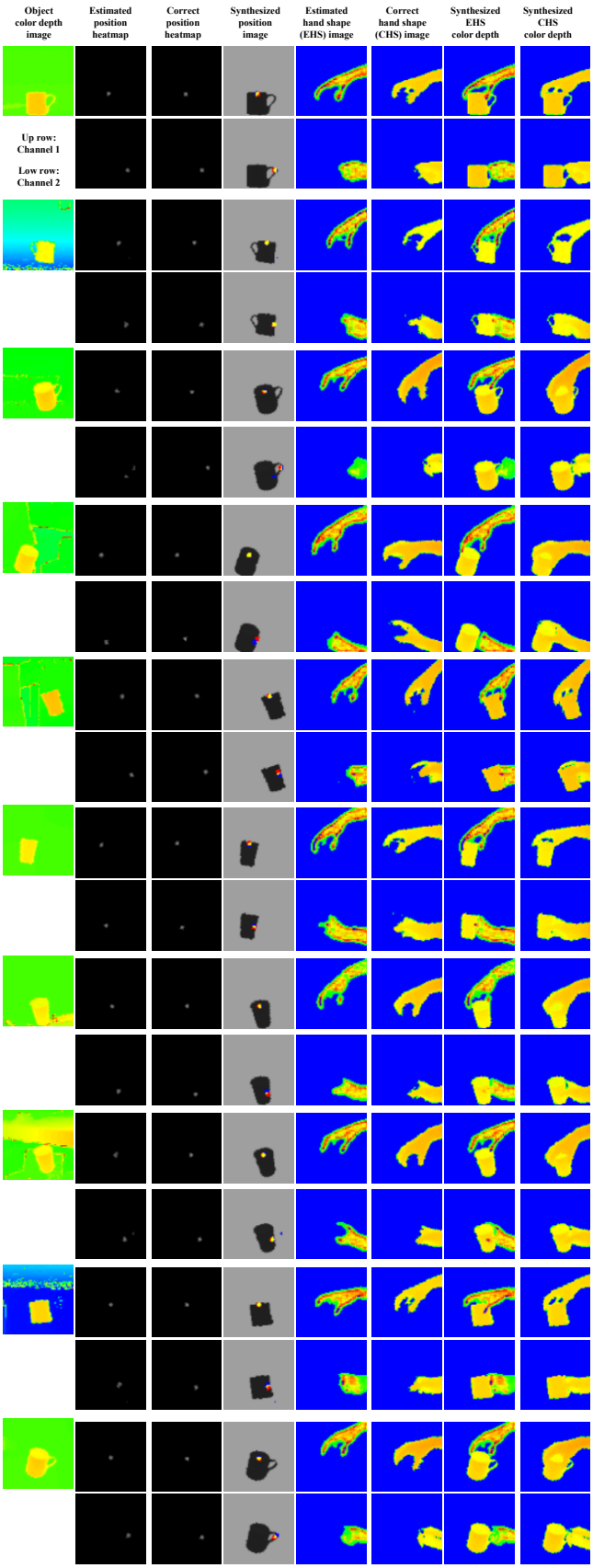

Figure 5 The estimation results of the Grasping Method Candidates Recall Network using evaluation data

Looking at the hand shape estimation results of the fifth column, the grasping hand shape close to the ground truth as the sixth column is estimated in most estimated grasping hand shape images. It can be seen that it is possible to estimate the grasping hand shape suitable for the given grasping position even for the unknown object image. Similar to the estimation result of the training data, the hand shape that grasps the object is estimated even if it is different from the ground truth as in the thirteen row. In addition, as in the fifth and seventh rows, the tendency that the estimation result of the hand shape grasped from above is estimated at the position shifted upward with respect to the object was the same as the estimation result of the training data. The mean and standard deviation of the error are 129.9±31.1 mm/pixel.

In the estimated hand shape image of the sixth row, blurred image is estimated as compared with other estimated images. It is considered that this is because the heat map given as an input was a heat map in which handle and side grasping positions were mixed. This hand shape is close to the hand shape grasping the side portion than the handle portion. It is considered that the hand shape grasped at the grasping position with the larger pixel value in the input heat map was estimated. In the estimated hand shape image of the eight row, the grasping position is shifted slightly downward as compared with the ground truth. It is considered that this is because the position where the grasping position indicated by the input grasping position heat map is slightly displaced downward compared to the grasping position of ground truth is estimated. In other words, it is considered that the hand shape grasping the object at the directed grasping position for the input object image is estimated.

Looking at the synthesized depth map in the seventh and eight columns, as in the learning result, the synthesized depth map of the estimated hand shape has variations in the pixel value as compared with the synthesized depth map of the ground truth. However, in the image of the tenth row, the position of the thumb indicated by the small depth value in the synthesized depth map of the correct hand shape is displayed by a color close to red in the synthesized depth map of the estimated hand shape, and the position of the palm and little finger indicated by a large depth value is displayed by a color close to green. Also in the result of the twelfth row, the number of pixels close to red is large in the arm part indicated by the small depth value, and the number of pixels close to green is large in the fingertip part indicated by the large depth value. Similarly to the training data result, it can be seen that it is possible to estimate grasping hand shape image with the tendency of the pixel value close to the ground truth.

The estimation result of the proposed grasping method candidate recall network shows that it is possible to estimate the grasping position and the grasping hand shape close to the ground truth of each input object image in most of the evaluation data. Among the results of estimating the grasping position and the hand shape different from the ground truth, most of the results show the grasping position neighboring ground truth and a similar hand shape grasping the object. Therefore, it can be considered that the essence of this research that is proposing of multiple grasping candidates for one input could be learned.

IV. CONCLUSIONS

In this research, we propose a method to recall candidates of multiple grasping methods from object shape while automatically performing clustering of grasping methods in learning process.

The grasping position estimation network part with one object image as input can estimate a plurality of grasping positions considering the object shape and the use of the object. In addition, different grasping positions were estimated for each channel of output heat map by learning on a multi-channel output network using dataset composed of one-to-one input and ground truth. As a result, it was seen that many types of grasping positions are automatically clustered on the learning process, and it is possible to estimate a plurality of grasping methods with more distinctive differences.

In this research, two similar shaped objects, cup and mug, are used. However, in general, there are many objects that are greatly different from these object shapes. If the object shape is different, the types of grasping methods and the number of grasping methods are different. Therefore, a mechanism is required to numerically express the certainty factor of the estimated grasping method. In addition, only the final grasping hand shape estimation is performed, however support of human work requires the research of a grasping method selection according to directions given by a human and the generation of motion after grasping object. Therefore, as a future task, it is possible to learn and estimate the finger and hand approach as the trajectory and procedure when grasping object, and the motion after grasping object. Also, when grasping an object using a robot hand, it is necessary that generation of a motor command considering the physicality difference between robot and human.

REFERENCES

[1] S. Ekvall, and D. Kragic, "Learning and Evaluation of the Approach Vector for Automatic Grasp Generation and Planning," Proceedings of IEEE International Conference on Robotics and Automation, pp.1824–1829, 2003.

[2] K. NAGATA, T. MIYASAKA, Y. KANAMIYA, N.YAMANOBE, K. MARUYAMA, S. KAWABATA, and Y. KAWAI, "Grasping an Indicated Object in a Complex Environment," Transactions of the Japan Society of Mechanical Engineers, Series C, vol. 79, no.797 , pp. 27–42, January. 2013.

[3] K. Huebner, and D. Kragic, "Selection of Robot Pre-Grasping using Box-Basecd Shape Approximation," Proceedings of IEEE/RSJ International Conference on Inteligent Robots and Systems, pp.1765–1770, 2008.

[4] K.Yamazaki, M. Tomono, T. Tsubouchi, and S. Yuta, "A Grasp Planning for Picking up an Unknown Object for a Mobile Manipulator," Proceedings of IEEE International Conference on Robotics and Automation, pp.2143–2149, 2006.

[5] M. Cai, K. M. Kitani, and Y. Sato, "Understanding Hand-Object Manipulation with Grasp Types and Object Attributes," Proceedings of Robotics: Science and Systems Conference, 2016.

[6] F. Chu, R. Xu, and P. A. Vela, "Real-world Multi-object, Multi-grasp Detection," IEEE International Conference on Robotics and Automation, 2018.