Grasping Pattern Estimation Based on Co-occurrence of Object and Hand Shape

Takuya Kawakami*, Tadashi Matsuo, Yoko Ogawa, Nobutaka Shimada
Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga, Japan
Email: *tkawakami@i.ci.ritsumei.ac.jp

***Abstract*—We propose a method of visually recalling appropriate grasping method from the shape of object parts. We quantitatively express the grasping method (i.e. interaction state of hand and object) by introducing its numerical descriptor (interaction descriptor), and recall how to grasp by machine learning of the relationship between partial object shapes and interaction descriptors. Then the relationship between the interaction descriptors and object appearance are learned by CNN regression. Part-based relationship between objects and grasping hand can be modelled and recalled multiple grasping modes for the same object if possible.**

I. Introduction

An object as a tool has its own functionHuman has several *typical grasping types* and switches them according to the tool's function [1]. In *Robot Vision* field, which automatically tries to recognize kinds of objects and their functions, many applications of deep learning methods give significantly good performances mainly for static object shapes.

In order for the robot to properly grasp the object, it is necessary to recognize the class of the object and recall an appropriate hand shape corresponding to it so that its function is successfully activated. Recently the main stream of object recognition research has been changed toward focusing to relationship between hand and object, namely treating object handling scenes. Wu et al.[2] described cooking scenes by modeling relationship between hands, cooking tools and food-staffs with their locational positions, but they did not observe detailed finger works. Cai et al.[3] utilized Deep neural nets to describe relationship between hand appearances and object textures and suceeded a task classifying a few grasping types.

However, appearance based recognition of object function like [4] is a challenging problem because all possible appearance of new objects, which are produced every day, cannot be completely registered. In order to visually recognize the object function, the following two characteristics are available: "1) an object is composed of a combination of typical graspable parts", and 2) "the objects with the same category of function should have their common shape of the parts and the corresponding grasping method".

Therefore, we propose a method of visually recalling appropriate grasping method from the shape of object parts. We quantitatively express the grasping method (i.e. interaction state of hand and object) with a numerical descriptor (interaction descriptor), and recall how to grasp by machine learning of the relationship between partial object shapes and interaction descriptors. Interaction descriptor is made by extracting features by shift invariant sparse auto-encoder (SISAE) [5] from grasping images composed of three channels of depth image, hand mask image, and object mask image. Due to the sparsity nature of SISAE, it is possible to generate a discrete grasping cluster for each object on the interaction descriptor space. In addition, since SISAE is invariant for parallel movement of a pattern in an input image, it can describe grasping images with the same hand shape and slightly different grasping positions as the same descriptor.

The collection of grasping images and object images, which are automatically captured from the scenes where a human grasps an object, are used as training samples for learning the descriptor space. Then the relationship between the interaction descriptors and object images are learned by CNN regression. Part-based relationship between objects and grasping hand can be modelled by training with partial image patches. Multiple grasping modes are clustered based on the similarity in the interaction descriptor space. Finally multiple grasping modes for an object can be estimated by mosaicing recalled grasping image patches with the same grapsing mode into a whole grapsing image. The performances are shown through experimental results.

II. overview

Fig.1 shows the overview of the system. The proposed system learns the relationship between an object shape and a grasping method by observing scenes where a human grasps it. And then, it can recall an appropriate grasping method from a partial shape of an unknown object.

As shown in Fig.2, the system maps each grasping image onto a interaction descriptor space generated by a SISAE, which is trained with grasping images. The trained auto-encoder consists of an encoder E and a decoder D. The former encodes a colored grasping image I_{grasp} as a 30-dimensional descriptor $E\left(I_{grasp}\right)$. The latter restores the grasping image as $D\left(E\left(I_{grasp}\right)\right)$ from the descriptor.

Next, as shown in Fig.3, we train a inference model R so that it can calculate a possible interaction descriptor from an object image. The inference model R is trained with an object image I_{obj} and an interaction descriptor $P_{teacher}$ that is calculated from a grasping image I_{grasp} of the object. By using the trained R, we can recall a possible interaction descriptor from an object, which is not used in the training. Then, we can recall a grasping image of the object by decoding the interaction descriptor as shown in Fig.4.

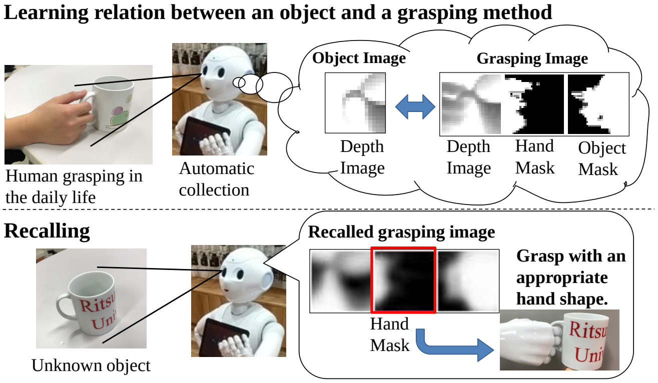


Fig. 1. Overview of the proposed system

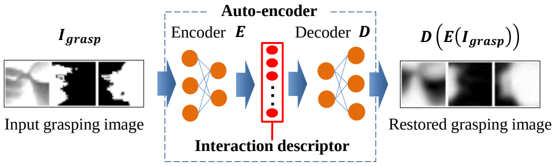


Fig. 2. Interaction descriptor based on auto-encoder

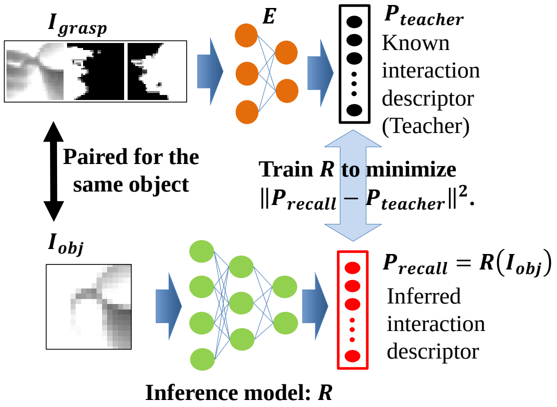

I_{grasp}
E
$P_{teacher}$
Known interaction descriptor (Teacher)
Paired for the same object
Train R to minimize $\|P_{recall} - P_{teacher}\|^2$.
I_{obj}
$P_{recall} = R(I_{obj})$
Inferred interaction descriptor
Inference model: R

Fig. 3. Training of inference model R

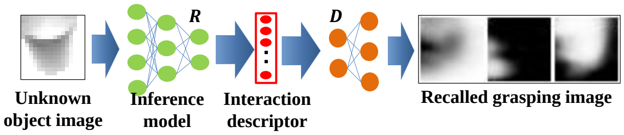


Fig. 4. Recalling grasping image from an unknown object image

To focus on partial shapes related to hand-object interactions, we generated grasping images as 32 × 32 pixel sub-images extracted from an whole scene where a hand grasped an object.

III. Grasping Image

We define a grasping image as a 3-channel image consisting of a depth image, a hand region mask and an object region mask because spatial relation between an object and a hand is important to identify an interaction.

Fig.5 shows the process of collecting grasping images from daily scenes of interactions. First, we take depth images with a depth camera and extract a sequence of depth images from an initial frame, which is just before touching an object, to a frame where the object is held up by a hand. Then, we remove points except the object and the hand by trimming based on depth and excluding a floor plane. Here, remained points consist of points in the object and those in the hand. To generate an object or hand region mask, we need to distinguish between the former and the latter. To find approximate positions of the object in each frame, we obtain the 3-dimensional coordinate system relative to the object by matching the points in the initial frame (which includes an object only) with those in the following frames by the Iterative Closest Point (ICP) algorithm. And then, we match each point in the initial frames with the points in the following frames by the Nearest Neighbor algorithm on the relative coordinate system. We consider a point in the following frames matched with a point in the initial frame as a point in the object. And we consider the other points as a point in the hand. We generate a depth image of an object and a hand, an object or hand region mask from those classified points and we combined them as a 3-channel image. And also, we obtain an object-only depth image from the initial frame.

Now we have pairs of a 3-channel image representing an interaction and a depth image including only an object used in the interaction. To train the auto-encoder for generating interaction descriptor, we extract a 32 × 32 pixel image from a 3-channel image. It is called a "grasping image" and it is randomly extracted so that both a hand region and an object region in it are larger than 5% of the extracted image. To train the inference model, we extract a 32 × 32 pixel image from an object-only depth image and it is paired with a grasping image including the corresponding part of the object.

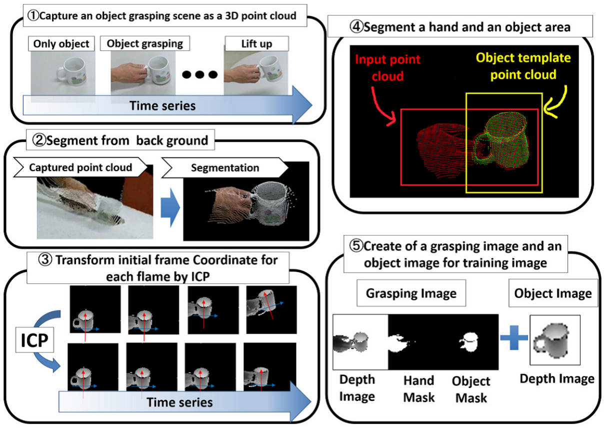

①Capture an object grasping scene as a 3D point cloud
Only object
Object grasping
Lift up
Time series
②Segment from back ground
Captured point cloud
Segmentation
③ Transform initial frame Coordinate for each flame by ICP
ICP
Time series
④Segment a hand and an object area
Input point cloud
Object template point cloud
⑤Create of a grasping image and an object image for training image
Grasping Image
Object Image
Depth Image
Hand Mask
Object Mask
Depth Image

Fig. 5. Process of collecting grasping images from daily scenes of interactions

IV. Shift invariant auto-encoder

We train an auto-encoder as a shift invariant auto-encoder[5] so that an interaction descriptor represents spatial pattern invariant to shift transforms. The shift invariant auto-encoder is trained with a cost function utilizing an input I and its shifted versions $T_i(I)$ as shown in (1), where T_i means a shift operator;

$$(T_i(I))(x,y) = I(x + \Delta x_i, y + \Delta y_i). \qquad (1)$$

The cost function consists of 3 terms, “shift sensitivity”, “restoration error”, and “sparseness cost”.

The shift sensitivity term is defined as

$$C_{var} = \sum_I \sum_i \| D(E(I)) - D(E(T_i(I))) \|_{L2}^2. \qquad (2)$$

By optimizing the encoder E and the decoder D so that they minimize (2), their combination is approximately shift invariant for learned inputs. To evaluate restored patterns without respect to shift transforms, the restoration error term should be small if a restored input matches one of shifted versions of its original input. It is defined as

$$C_{res} = \sum_I \min_i \| T_i(I) - D(E(I)) \|_{L2}^2. \qquad (3)$$

The sparseness term causes that similar inputs are encoded into descriptors close to each other[6]. It is defined as

$$C_{sparse} = \sum_I \frac{\|E(I)\|_{L1}^2}{\|E(I)\|_{L2}^2}. \qquad (4)$$

The total cost function is defined as a weighted sum of the above 3 terms;

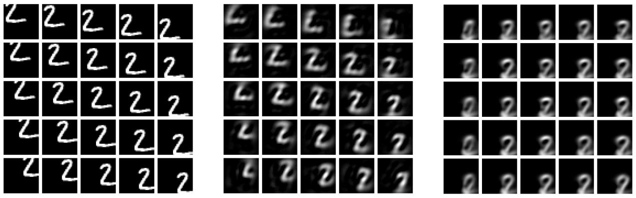
$$\lambda_{var} C_{var} + \lambda_{res} C_{res} + \lambda_{sparse} C_{sparse}. \qquad (5)$$

The auto-encoder, a pair of E and D, is trained so that it minimize the cost function (5).

Here, we demonstrate shift invariant property of the proposed method using experiments for digit patterns. As an encoder, we used a neural network consisting of a single CNN with 9×9 filter kernels and 16-channel outputs and a max pooling layer with stride 2 and a three-layer fully connected neural network (NN), where each layer has 1500, 150, 30 outputs respectively. As a decoder, we used a three-layer fully connected NN, where each layer has 150, 1500, 1024 outputs, respectively. In addition, we used a hyperbolic tangent as an activation function, which is placed between each pair of layers. We generated two pairs of encoders and decoders with the same structure. One was trained as an ordinary auto-encoder, and the other was trained as a shift invariant auto-encoder by minimizing (5) for digit images of training images in the MNIST database [7]. The shift invariant auto-encoder was trained with the following shift parameters:

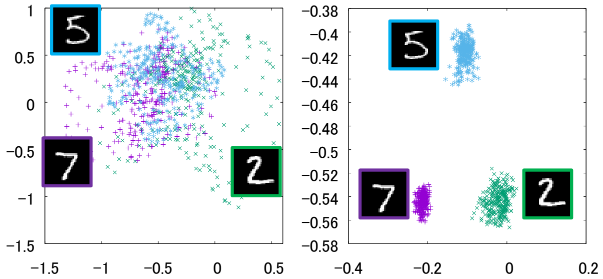
$$\{(\Delta x_i, \Delta y_i)\} = \{-8, -6, -4, -2, 0, 2, 4, 6, 8\}^2. \qquad (6)$$

For the ordinary auto-encoder, we used additional images that were randomly shifted according to the parameters in (6). Both auto-encoders were trained by stochastic gradient descent (SGD) [7] with learning rate 1.0×10^{-3}, and both were updated with every 50 samples that were randomly extracted from the training images (60k samples) in the MNIST database. We used auto-encoders that were updated 100,000 times ($\approx$ 83 epochs).



(a) A digit image and its shifted versions (b) Images restored by an ordinary auto-encoder (c) Images restored by a shift invariant auto-encoder

Fig. 6. Input images and restored images of a digit



(a) By an oridnary auto-encoder (b) By a shift invariant auto-encoder

Fig. 7. Distributions of descriptors of digits

As an example, we encoded and decoded an test image of the digit “2”, which is not used in training auto-encoders. Input images are shown in the left of Fig. 6, where the center image is the original image in the MNIST database and the others are its shifted versions. Images in the center of Fig. 6 are restored from input images by an ordinary auto-encoder. Images restored by a proposed shift invariant auto-encoder are shown in the right of Fig. 6. The images restored by the ordinary auto-encoder are located depending on the shifts in the input images. Conversely, the images restored by the shift invariant auto-encoder are very similar to each other and they are close to a typical shape of the digit “2”.

In addition, we calculated the distributions of the descriptors from the shifted images. We encoded the digit images corresponding to “2”, “5”, and “7” and their shifted versions using the two auto-encoders. The left figure in Fig. 7 shows the distributions from the ordinary auto-encoder, and the right one shows those from the shift invariant auto-encoder. In these figures, 30 dimensional descriptors are projected onto a two-dimensional space spanned by the three mean vectors of the descriptors for digits “2”, “5”, and “7”. By comparing these figures, we see that descriptors generated by the shift invariant auto-encoder are obviously concentrated for each digit. This means that a descriptor generated by a shift invariant auto-encoder represents the spatial subpattern. In addition, descriptors in Fig. 7 make clusters corresponding to digits, even though we have entered no digit information when training the shift invariant auto-encoder.

V. Experiments of Recalling Grasping Method

In our experiments, we use 4 categories of objects that are a mug, a cup, a ball, and a spray can as shown in Fig. 8. Each category includes 4 instances of objects.

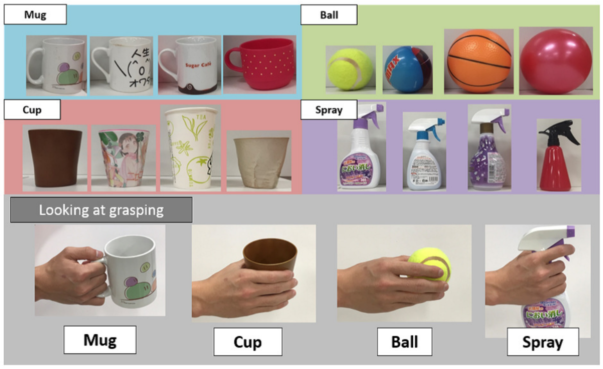


Fig. 8. Objects used in the experiments

A. Structure of auto-encoder

We used an auto-encoder with the following structure to generate numeric interaction descriptors. As an encoder, we use a neural network consisting of a single CNN with 9×9 filter kernels and 16-channel outputs, a L2 pooling layer with stride 2, and a three-layer fully connected neural network, where each layer has 1500, 150, 30 outputs respectively. As a decoder, we use a three-layer fully connected NN, where each layer has 150, 1500, $3072 = 3 \times 32 \times 32$ outputs, respectively. In addition, we used a hyperbolic tangent as an activation function, which is placed between each pair of layers.

B. Interaction descriptor

We prepared 4 objects for each category (Fig. 8) and we selected one object for each category as a test sample and used others as training samples. We collected 100 images of interactions for each object and extracted grasping images from them. Then, we trained an auto-encoder for generating interaction descriptors with those training samples.

To see the information extracted by the auto-encoder, we encoded some training samples of grasping images and then decoded their descriptors. Fig 9 shows input images and corresponding restored images. These figures show that grasping images can be approximately restored from descriptors and fingers are restored so precisely as to be distinguished from each other. In the case of a mug in Fig. 9, a grasping image is restored at a position different from the input. This is because the shift invariant auto-encoder extracts shape itself without respect to shifts.

In Fig. 10, we show the disribution of interaction descriptors for each object. In the figure, interaction descriptors are projected onto two dimensional space spanned by the first and the second principal compnants, which are calculated from interaction descriptors of the training grasping images. Descriptors from the same object are drawn with the same color. The boxes with the red border mean test samples that are not used in training. The figure shows that descriptors of a specific object are near to each other on the descriptor space. Also, descriptors of an object and another object are distributed near to each other if the objects belong to the same category. This means that similar interactions are mapped as descriptors near to each other and interaction descriptors successfully reflect similarity of interactions.

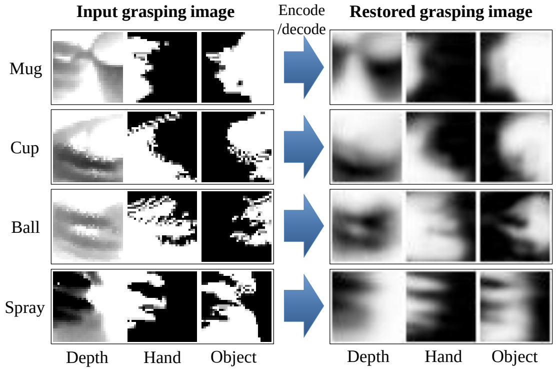


Fig. 9. Grasping images restored by the auto-encoder

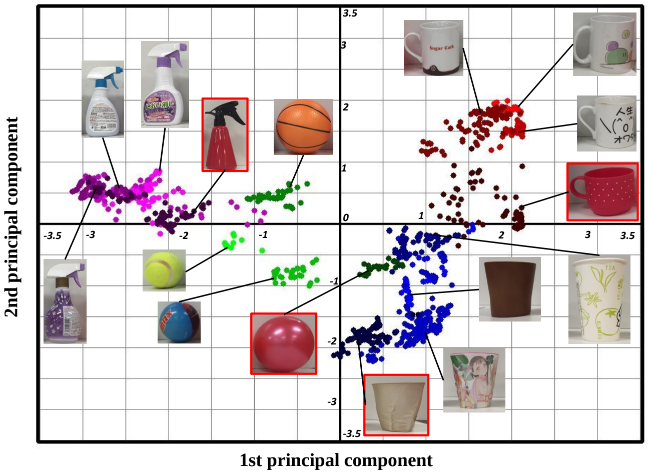


Fig. 10. Distribution of interaction descriptors

C. Structure of inference model

As a inference model, we used a neural network with the following structure, which calculates a possible interaction descriptor from an object-only depth image. The inference model consists of sequences of layers of a CNN layer with 16-channel 9×9 filter kernels, a L2 pooling layer with stride 2, a subtractive normalization layer, a CNN layer with 64-channel 5×5 filter kernels, a L2 pooling layer with stride 2, a subtractive normalization layer, a fully connected layer with 1500-dimensional outputs. and a that with 30-dimensional outputs.

As explained in the section III, we have pairs of a grasping image and a corresponding object-only depth image. The inference model is trained so that it outputs an interaction descriptor of a grasping image from the corresponding object-only depth image.

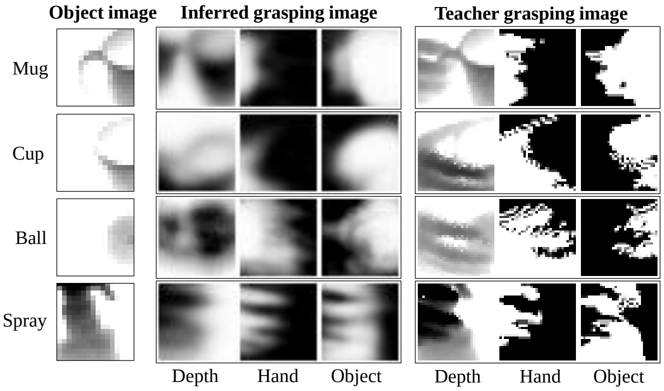


Fig. 11. Inferred grasping images of training samples

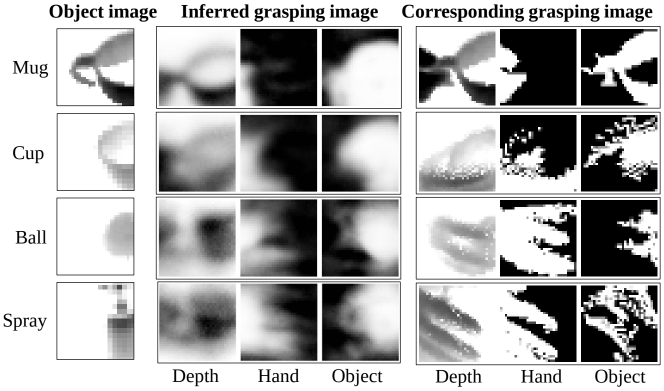


Fig. 12. Inferred grasping images of test samples

D. *Recalled grasping image*

In Fig. 11, we show some grasping images decoded from interaction descriptors inferred from object-only depth images of training samples by the inference model. In the case of a mug, the inferred object region mask correctly had a grip-like part and the hand region mask was also correctly recalled. In the case of a spray can, the inference model also successfully recalled an interaction descriptor appropriately reflecting a corresponding interaction. In the grasping images recalled for a cup and a ball, the hand region masks were roughly recalled though the object region masks had an additional small grip-like part.

In Fig. 12, we show examples of test samples. In the case of a mug, the inferred hand region mask had a shape for holding the grip. In the case of a cup and a ball, the inference model roughly recalled hand region masks for holding the objects. In the case of a spray can, shape of fingers were approximately recalled but the object region mask differed from the input.

The inference model infers a grasping image from a local part of an object given as an input. So, even if input depth images originates from the same object, inferred interaction descriptors should differ according to the parts included in the inputs. To confirm it, we inferred descriptors from two depth image of a mug, where one included a handle and the other did not include the handle. Fig. 13 shows the inferred grasping images. From the input image with the handle, the inference model inferred a hand shape for holding a handle. On the other hand, from that without the handle, it inferred a hand shape for supporting a bottom of a cup and the shape was close to that inferred from a cup. This means that the inference model learns importance of local characteristic parts such as a handle automatically without special annotation. In addition, the inference model successfully recalled an appropriate interaction descriptor for a side part of a mug even though such side parts of mugs were not used in training. This means that the inference model can recall a descriptor from an object not used in training if its local shape and interaction with it are common to other types of objects used in training.

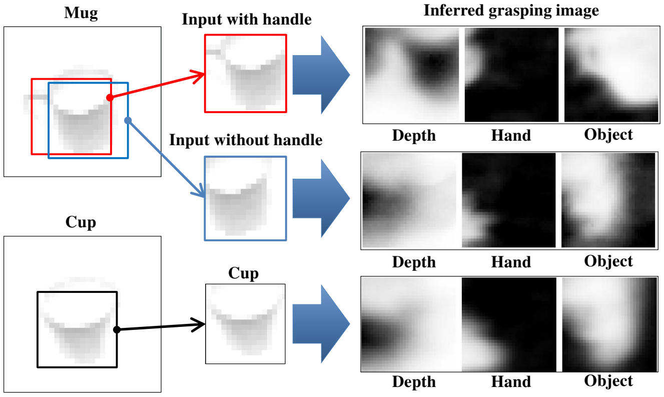


Fig. 13. Recalled grasping images from images with/without handle

VI. Recall of Multilple Grasping Methods for An Object

The CNN-based inference model of the grasping method introduced in the previous sections can estimate an interaction descriptor for each small partial image patch including an object part, then the descriptor is expanded to three-channel image patch which consists of depth, hand region, and object region. For each channel, the inferred image patches are merged into a whole image by mozaicing. For hand region channel, an image map is obtained in which each pixel value means the probability that the pixel corresponds to hand region. Based on the probability map the depth image of grasping hand is cropped. The same process is done for the object region channel.

Before the mosaicing process, the inferred image patches are classified into a few clusters based on the similarity of the interaction descriptor by K-mean clustering. For each cluster, which corresponds to an identical grasping mode, the mosaicing process is applied (see Fig. 14). Fig. 15 shows that three clusters of descriptor are seen, two of them correspond to apparent grasping modes (gripping the handle and grasping the body) and the rest one has no significant grasping type.

Fig. 16 shows the obtained probability map for hand region (left-hand images) and the regions with high probability displayed over the object depth images (right-hand images). The upper images are the result corresponding to class 1, which

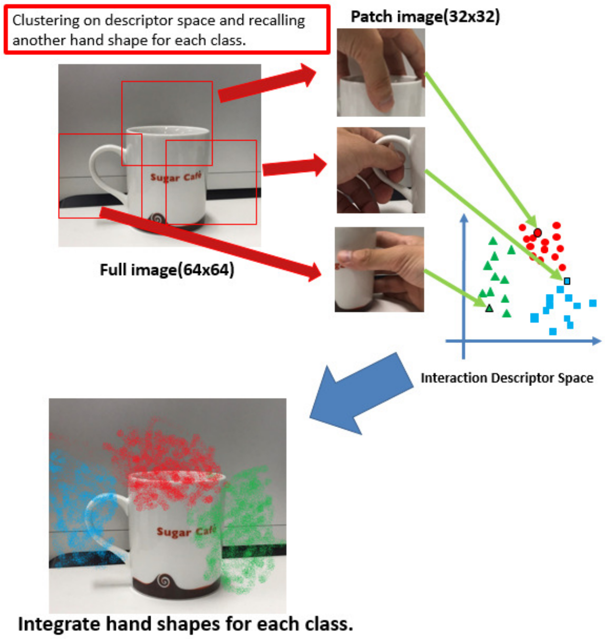

Clustering on descriptor space and recalling another hand shape for each class.
Patch image(32x32)
Full image(64x64)
Interaction Descriptor Space
Integrate hand shapes for each class.

Fig. 14. Recall of grasping methods by merging part-based inference results

represents gripping the handle of the mag. The lower ones are that corresponding to class 2, which represents grasping the top of the mag. The results shows that the proposed method can infer two typical grasping methods associated to 3-D shape of different parts for an object.

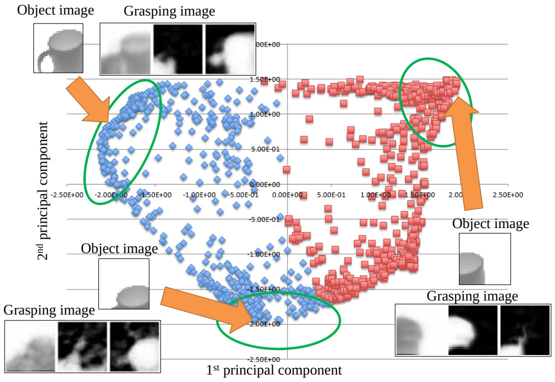


Fig. 15. Clusters of interaction descriptor for a mug

VII. CONCLUSION

We proposed a method of visually recalling appropriate grasping method from the shape of object parts. We quantitatively express the grasping method by introducing interaction descriptor, and recalled how to grasp by machine learning of the relationship between partial object shapes and interaction descriptors. The relationship between the interaction descriptors and object part appearance are learned by CNN regression. Multiple grasping modes for one object can be estimated if the object has them. The experimental results showed that the proposed method can recall the appropriate grasping method for the first-seen object if the similar partial shape and grapsing has been learned.

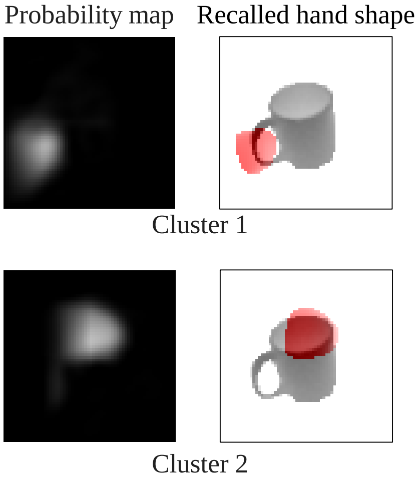


Fig. 16. Inferred grasping methods for a mag: Right) probability map for hand region, Left) recalled hand shape, Top) inference for cluster 1, Bottom) that for cluster 2

Current method can use only appearances of textures and depth information. For more precise description and descrimination of grasping objects, and further applications including their functional operations by autonomous robots, more detailed fingering activities should be modelled into the hand-object interaction description. Since the detailed fingering observations from images are recently available like OpenPose toolkit[8], the extentions of the proposed method toward utilizing such information are to be future works.

REFERENCES

[1] N. Kamakura, *Shape of hand and Hand motion*. Ishiyaku Publishers, 1989. [Online]. Available: https://www.ishiyaku.co.jp/search/details.aspx?bookcode=211970

[2] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[3] M. Cai, K. M. Kitani, and Y. Sato, "An ego-vision system for hand grasp analysis," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 524–535, Aug 2017.

[4] T. Kitahashi, M. Higuchi, A. Kojima, and K. Fukunaga, "Cooperative recognition of human movements and objects and its modeling," *CVIM*, vol. 2005, no. 18, pp. 109–116, mar 2005. [Online]. Available: https://ci.nii.ac.jp/naid/110002694818/

[5] T. Matsuo, H. Fukuhara, and N. Shimada, "Transform invariant auto-encoder," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*. IEEE, 2017, pp. 2359–2364. [Online]. Available: https://doi.org/10.1109/IROS.2017.8206047

[6] T. MATSUO and N. SHIMADA, "Construction of latent descriptor space and inference model of hand-object interactions," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 6, pp. 1350–1359, 2017.

[7] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[8] T. Simon, H. Joo, I. A. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," *CoRR*, vol. abs/1704.07809, 2017. [Online]. Available: http://arxiv.org/abs/1704.07809