Object Grasping By Learning Hand-Object Interaction from Human Behaviors

Masaki Yano*, Hiroya Fukuhara, Tadashi Matsuo, Nobutaka Shimada
Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga, Japan
Email: *yano@i.ci.ritsumei.ac.jp

***Abstract*—We propose a method that enables a robot to grasp an object based on how a human grasps it. By observing a human grasping an object, we model the relationship between a shape of the object and a hand shape of the human. In advance, the model is trained with pairs of images before/after a human grasps an object. By using CAE(Convolutional Auto-Encoder)[1], we can train the model without labeling interaction between a hand and an object. With this model, a robot can infer a hand-object appearance when a human grasps an object from an object shape. The robot can grasp the object by moving its hand to a shape of the object where a human touches for grasp it. By experiments for actual objects, we show availability of proposed technique.**

I. INTRODUCTION

Recently, robots work in the various fields. In the future, robots will support us more at living space. These robots need to manipulating the objects which we usually use. However, there are many kinds of objects we usually use. Each object has its own function and can be categorized according to it. An object category (e.g. 'cup') includes many varieties of shape and size. It is hard to find a method to estimate how to grasp and use by hand for all of the shape varieties.

Yamazaki, et al. proposed a grasp planning method based on object's 3D model[2]. In the method, a stable grasping pose can be found by evaluating each pose with a contact area of a robot hand and a target object and balance of gravity. However, the method requires many images taken from viewpoints around the target object to reconstruct the 3D model.

Each functional category of objects has some typical interactions between human hands and object parts, for example, grasping a handle of a cup, supporting a bottom of a cup etc. The same functional category has a similar grasping hand pose and the object shape. In this study, we describe the hand-object interaction in a numerical form and model a relationship between a shape of an object and how to grasp it (hand-object interaction) by using a machine learning scheme. We build a model to infer an hand-object interaction from an object appearance. A robot manipulator can grasp the object part designated by the inferred hand-object interaction.

II. GRASPING PATTERN INFERENCE

We define a grasping pattern as a point of touch and the hand shape which grasping an object. We infer a grasping pattern from an object image.

Matsuo and Shimada propose 'interaction descriptor space' to describe a hand-object interaction[3]. They train a model which infers hand-object interaction from an object image by using CAE (Convolutional Auto-Encoder) and CNN (Convolutional Neural Network)[4].

For inference hand-object interaction, We train an Auto-Encoder and an estimator. We train the Auto-Encoder with interaction images (Fig. 1). The grayscale texture is hand-object appearance when a human grasps the object. The mask images are made of the grayscale texture.

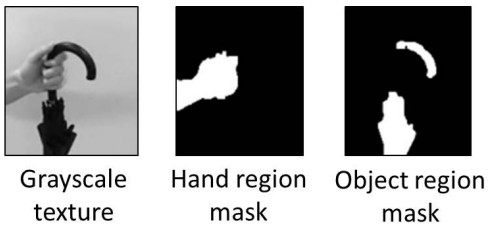

Fig. 1. Interaction images

The Auto-Encoder includes an encoder and a decoder. The encoder (Fig. 2) describes the interaction image as interaction descriptor (a 30-dimensional vector describing hand-object interaction). The decoder (Fig. 2) restores the inferred interaction descriptor to grayscale texture, hand region mask, and object region mask images.

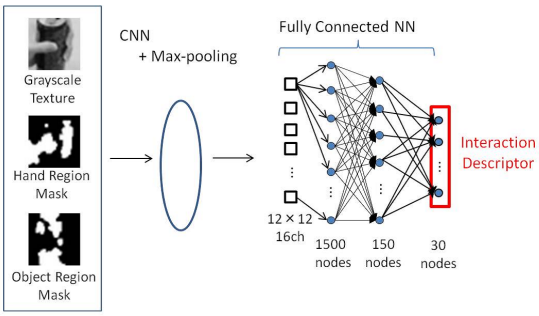

Fig. 2. The structure of encoder

In addition, We train the estimator (Fig. 4) with pairs of an object appearance and an interaction descriptor (output of the encoder) about the object. The estimator infers interaction descriptor and interaction probability from an object appearance.

We construct the inference model by the estimator and the decoder. The model infers interaction descriptor and interaction

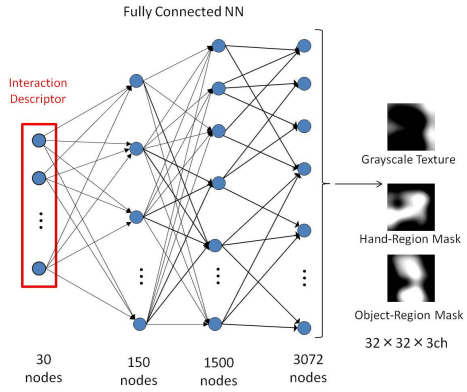

Fig. 3. The structure of decoder

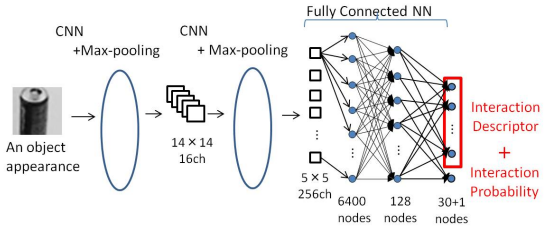

Fig. 4. The structure of estimator

probability from a patch image which is a 32 × 32[pixel] window located in an input image. By using raster scanning, the model infers them of all input image areas. The output of the model is which calculated the sum of the local inference weighted by the inferred probability.

III. TRAINING INFERENCE MODEL AND EVALUATION

Figure 5 shows sets an object and hand-object interaction which are used training. We train the models with 2160 images which include 18 types of interactions.

Figure 6 and Figure 7 show an Interaction Descriptor Space. 10 types interactions are plotted on there. Similar interactions are plotted on neighbor and different interactions are plotted on apart in the Interaction Descriptor Space. In Figure 6, interactions other than 'cutter', 'scissors' and 'mug type1' are plotted on apart. 'Cutter', 'scissors' and 'mug type1' are plotted on overlapping in Figure 6, but they are plotted on different positions in Figure 7. Accordingly, the model can infer different grasping pattern about each category.

Figure 8 shows the result of inference by using another can (Fig. 8(a)). Figure 8(c) shows that left side of the can is hand region. In addition, Figure 8(d) shows that all partial of the can is object region. The inferred images (Fig. 8(b),(c),(d),(e)) show that a hand should approach the can from the left side when a human grasps it. It is similar to the training image (Fig. 8(g)). The interaction probability image (Fig. 8(f)) shows high interaction likelihood around the can.

Accordingly, it is successful in inference regarding an unknown object.

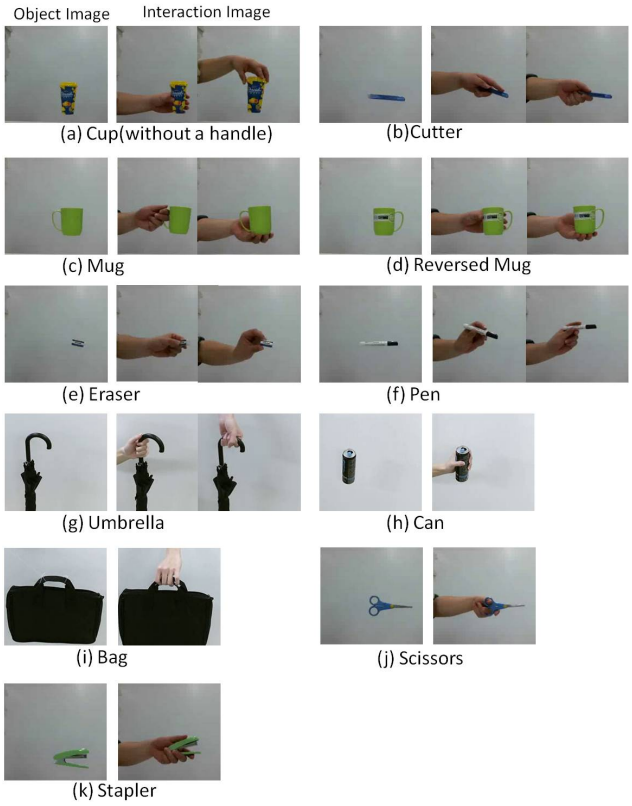

Fig. 5. Samples of training the model

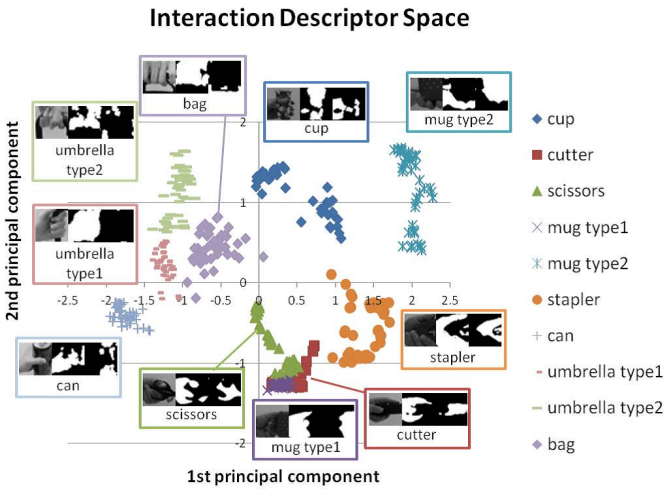

Fig. 6. Interaction Descriptor Space(1st, 2nd principal component)

IV. IMPLEMENTATION OF GRASPING MOVEMENT BASED ON INFERRED HAND-OBJECT INTERACTION

A. *calculate positions of a hand goal and an object center*

For grasping an object, we have to get a goal of position and pose of a hand. For getting them, we calculate a goal of a hand and a center of an object in an object image.

We get a goal of a hand by inferred hand region mask image and get a center of an object by inferred object region mask image. We get their points in the following steps.

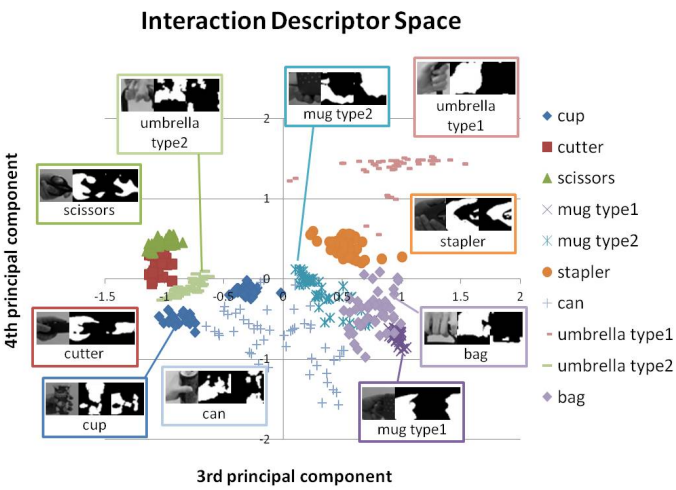

Fig. 7. Interaction Descriptor Space(3rd, 4th principal component)

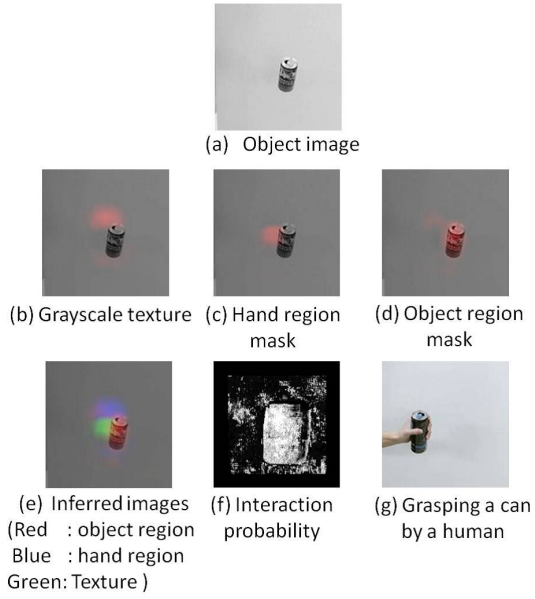

Fig. 8. Inference of an unknown can

1) To infer interaction descriptor from an object image
2) To restore the interaction descriptor to hand region mask and object region mask image
3) To binarize the images based on threshold value
4) To Calculate the center of gravity in max area region

Finally, we convert the points from an image coordinate system to a robot coordinate system.

B. object grasping

We use the calculated hand position as the goal position. We get a goal pose from calculated 2 points in IV-A. The pose when a normal vector of the robot's palm faces the object is the goal.

We record a hand shape of grasping a can (Fig. 9).After the hand moves to goal, robot replays the hand shape and grasps an object.

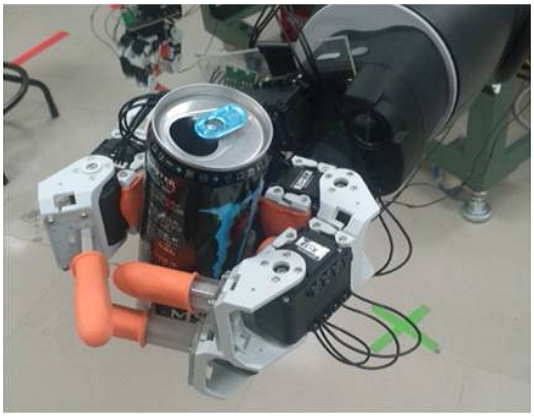

Fig. 9. Hand shape of recorded

V. EXPERIMENT OF OBJECT GRASPING

Figure 10(a) shows the object which a robot grasps in the experiment. A robot grasps an umbrella which was used training. Because of getting the depth easily, we wound a cloth onto the umbrella handle.

Figure 10(b) shows experiment situation. The umbrella puts into an umbrella stand. The height of the umbrella handle is 80cm and distance from the robot to the umbrella is 70cm.

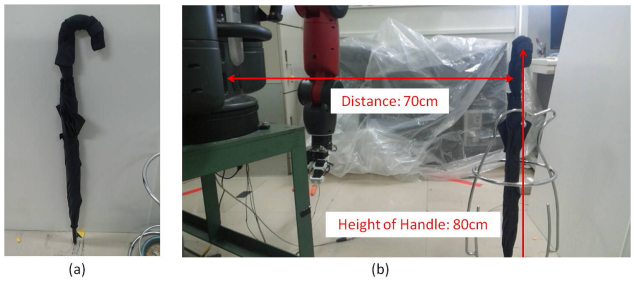

Fig. 10. An object and situation of experiment

Figure 11 shows an input image of the inference model. It is a grayscale image around the umbrella. The image was trimmed off far distance pixels (over 1.5 meters) by using depth information.

Fig. 11. The input of the inferred model

Figure 12 is an inference image (putting on the input). The green mask means hand region and the red mask means object region. It shows that left side and over of the umbrella handle is hand region. In addition, around the umbrella handle is object region.

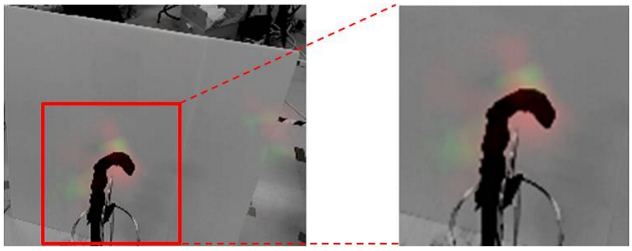

Fig. 12. Result of inference (Red: object region, Green: hand region)

Figure 13 is a binarized image of the inference image (Figure 12). The largest region label of hand region is over the umbrella handle and object region is on the umbrella handle. They show hand position goal is over the handle and pose is facing the umbrella (direction of the floor).

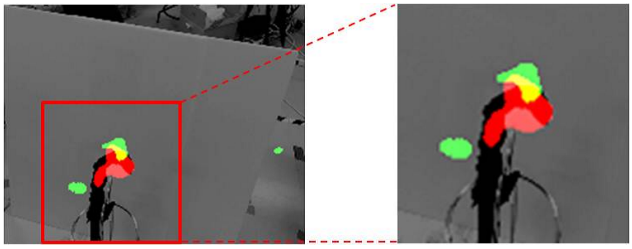

Fig. 13. Result of binarization (Red: object region, Green: hand region)

Figure 14 shows grasping movement by the robot. After 9.9 seconds, the hand reached the goal and started grasping. After grasping the umbrella, we commanded the robot to lift up the umbrella. Lifting up for 11 seconds, robot kept grasping the umbrella.

VI. CONCLUSION

We proposed the method that enables a robot to grasp an object based on the object appearance. We train an Auto-Encoder and an estimator with interaction images and object appearance images. The Auto-Encoder describes the interaction image as interaction descriptor. The estimator infers an interaction descriptor and its probability from an object appearance. The robot grasps an object based on the inference. We show availability of the method by the object grasping experiment. The robot can lift up an umbrella without dropping in the experiment.

In future work, we will infer a hand shape when a human grasps an object. Robot grasps an object by replaying a recorded hand shape in this study. If we can infer the hand shape, don't have to configure a hand shape each object.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers JP24500224, JP15H02764.

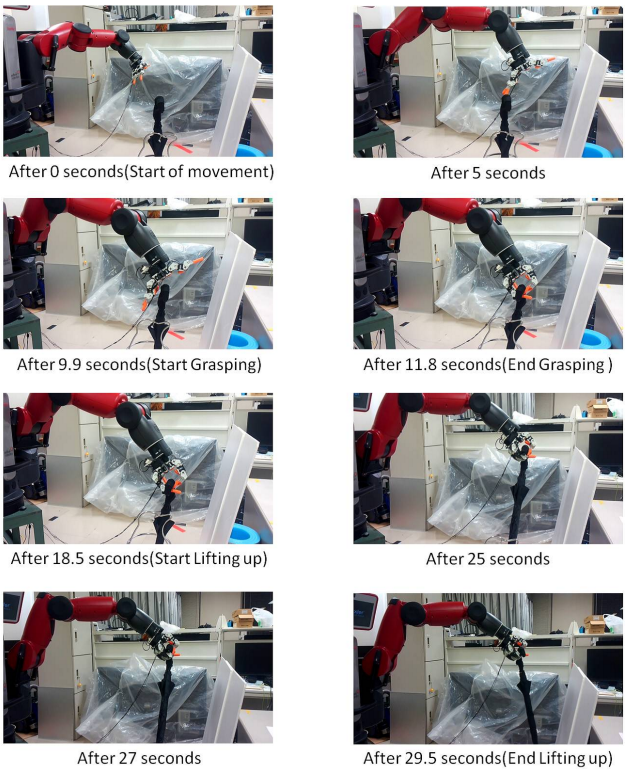

Fig. 14. Movement of grasping an umbrella

REFERENCES

[1] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber: "Stacked convolutional auto-encoders for hierarchical feature extraction.", Artificial Neural Networks and Machine Learning ICANN 2011. Springer Berlin Heidelberg, 2011. pp. 52-59.

[2] Kimitoshi Yamazaki, Masahiro Tomono, Takashi Tsubouchi and Shin' ichi Yuta : "A grasp planning for picking up an unknown object for a mobile manipulator.", Proceedings 2006 IEEE International Conference on Robotics and Automation, ICRA 2006. 2006, pp. 2143-2149

[3] Tadashi Matsuo, Nobutaka Shimada: "Construction of Latent Descriptor Space of Hand-Object Interaction", The 22nd Joint Workshop on Frontiers of Computer Vision (FCV2016), 2016, pp. 117-122,

[4] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner: "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11, 1998: pp. 2278-2324.