Object Recognition Based on Human Description Ontology for Service Robots

Hisato Fukuda
Saitama University
JSPS Research Fellow
255 Shimo-okubo, Sakura-ku,
Saitama, 338-8570, Japan
Email:fukuda@cv.ics.saitama-u.ac.jp

Satoshi Mori
Saitama University
255 Shimo-okubo, Sakura-ku,
Saitama, 338-8570, Japan
Email:tree3mki@cv.ics.saitama-u.ac.jp

Yoshinori Kobayashi
Saitama University,
JST PRESTO
255 Shimo-okubo, Sakura-ku,
Saitama, 338-8570, Japan
Email:yosinori@cv.ics.saitama-u.ac.jp

Yoshinori Kuno
Saitama University
255 Shimo-okubo, Sakura-ku,
Saitama, 338-8570, Japan
Email:kuno@cv.ics.saitama-u.ac.jp

Daisuke Kachi
Saitama University
255 Shimo-okubo, Sakura-ku,
Saitama, 338-8570, Japan
Email:kachi@mail.saitama-u.ac.jp

***Abstract*— We are developing a helper robot able to fetch objects requested by users. This robot tries to recognize objects through verbal interaction with the user concerning the objects that it cannot detect autonomously. We have shown that the system can recognize objects based on an ontology for interaction. In this paper, we extend a human description ontology to link a "human description" to "attributes of objects" for our interactive object recognition framework. We develop an interactive object recognition system based on this ontology. Experimental results confirmed that the system could efficiently recognize objects by utilizing this ontology.**

Keywords*— *object recognition; service robots; ontology; human description;

I. INTORODUCTION

As the number of elderly and handicapped persons in developed countries continues to rise, the potential of service robots to offer assistance has increasingly attracted attention. We have been developing a helper robot able to fetch objects requested by users. Such a robot must recognize the desired object(s) in order to carry out its tasks. However, it is difficult for a system to recognize objects autonomously without fail under various real-world conditions. To address this problem, we are currently working on an interactive object recognition system [1, 2]. In this system, the robot asks the user to verbally provide information (e.g., color and shape) about an object that it is unable to detect autonomously. In our sample scenario (Fig. 1), the user asks the robot to fetch the can in that scene. The robot asks the user about its color, and detects the desired object (can) by using its color (blue).

In order to be effective, an interactive object recognition system (a robot) needs to be able to recognize human descriptions correctly. However, humans describe objects in various ways. In addition, humans use various descriptions depending on the situation. In such a scene shown in Fig. 2, the description "the blue object" may indicate the middle of the three objects. However, if this center object does not exist, the left object (tissue paper box) may be described as "the blue object." In the Japanese language, the adjective "marui" ("round") is used to describe both 2-D circles and 3-D spheres. In this scene, the description "the marui (round) object" may indicate the right object. However, it may also indicate spherical objects such as balls and cylindrical objects such as cans depending on the situation. In the interactive object recognition framework, to interact with users unambiguously, we need to organize relationships between human descriptions and attributes of objects that they indicate. In our previous study, to address the problem, we proposed the basic ontology for interactive object recognition in [1] and elaborated on the ontology of object shapes in [2].

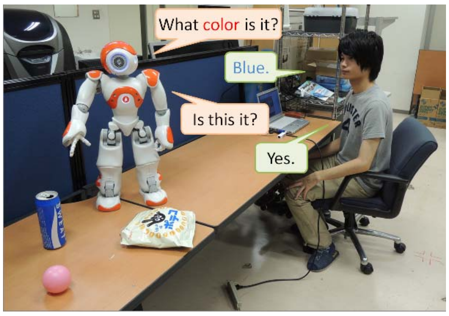


Fig. 1: Interactive object recognition.

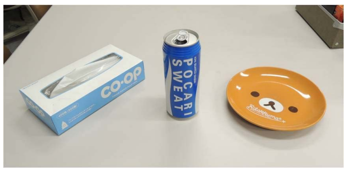


Fig. 2: Example scene.

In this paper, we aim to develop an interactive object recognition system that deals with various human descriptions and recognizes objects efficiently by utilizing their ontology. To do that, we further examine human descriptions of objects through experiments using human participants and extend the ontology to deal with various descriptions. We also add the ability to reassess priorities among physical attributes indicated by the same descriptions through interaction with the user. These priorities help a robot (system) to recognize the desired object efficiently. As mentioned before, humans may use the same descriptions to indicate multiple physical attributes such as "marui" for circles and spheres. Suppose that the user mentions such a multiple meaning expression and the recognition system finds multiple possible objects that may be indicated by the expression in the scene. The system will eventually know what object the user indicated by the expression through interaction. Based on the information obtained in such cases, the system reassesses the priorities so that the system can efficiently recognize target objects through a smaller number of interactions. In experiments, we evaluated the recognition system.

II. RELATED WORK

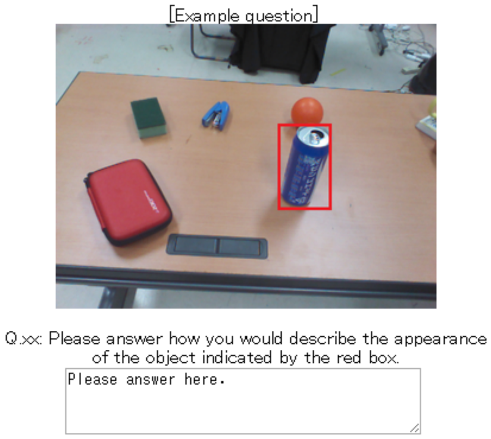
Since Winograd's pioneering work [3], a great deal of research has been conducted on systems able to comprehend a scenario or tasks through interaction with the user [4, 5]. However, such studies have primarily dealt with objects that can be described sufficiently with simple word combinations, such as "blue box." Moreover, these studies have neglected to consider constructing ontologies.

Ontology has recently been increasingly studied in robotics as a means to provide robots with organized knowledge. For example, Kobayashi et al. [6] proposed a robot action ontology and a robot knowledge ontology autonomously constructed from the Japanese version of Wikipedia. Ontologies have also been utilized for the interpretation and categorization of images: Maillot et al. [7] constructed an ontology that describes the visual appearance of objects by texture, color and spatial relation, and then developed an autonomous object recognition. system using this ontology; Dasiopoulou et al. [8] also proposed an ontology representing image features for object recognition; and Holzapfel et al. [9] used an ontology defining object classes hierarchically in their robot system in order to learn unknown objects through dialogue. However, none of these studies considered a key problem posed by human description, namely that humans employ various ways of describing objects.

We therefore set out to construct an ontology to link "human descriptions" to "attributes of objects" for our interactive object recognition framework [1, 2]. In this paper, we extend this ontology and present a recognition system that can recognize objects utilizing descriptions from users.

III. EXPERIMENT TO EXAMINE OBJECT DESCRIPTIONS BY HUMANS

We performed an experiment to examine how a human enables another to understand and retrieve a desired object when the second person is not supplied with the name of the object in question.


[Example question]

Q.xx: Please answer how you would describe the appearance of the object indicated by the red box.

Please answer here.

Fig. 3: Interface used to collect human descriptions.

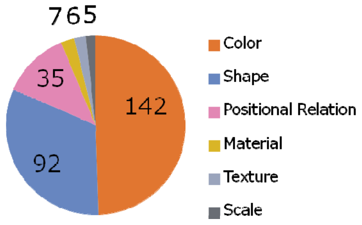


Fig. 4: Classification result of descriptions.

In this experiment, we used 27 participants, all of whom were undergraduate students of our university. We conducted our user study on a questionnaire system to collect human descriptions about appearances of objects when they are not allowed to mention them by their name and function. Participants were asked to describe the object indicated by the red box in a given image when they ask others to get the object (See Fig.3). We assembled 20 daily objects that we may want a robot to bring. We randomly chose five objects from these 20 objects to prepare 15 scene images, thus creating 75 situations (5 objects in 15 images). Each participant dealt with eight cases randomly chosen among them.

In this experiment, we collected 181 descriptions. Fig. 4 depicts the classification result of the descriptions. When any descriptions included multiple types of attributes, such as "red sphere", they were counted in both categories. As shown in Fig. 4, the descriptions for appearance of objects are classified into color, shape, positional relationship, material, texture/pattern and scale. Among 181 descriptions, 91 descriptions consisted of multiple attributes. To clearly distinguish between different objects, combinations of simple descriptions such as "red ball on the left side," are preferred to detailed descriptions about single attribute. We also confirm that humans use various descriptions for the same objects. For instance, "red sphere" and "orange marui (round) object" were used to describe the object shown in Fig.5.

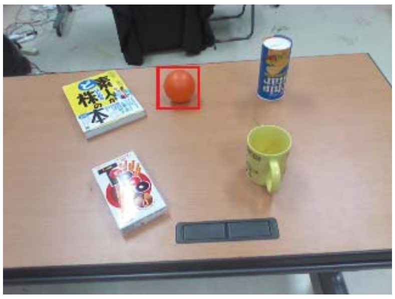


Fig. 5: Example image used in the experiment.

IV. ONTOLOGY FOR INTERACTIVE OBEJCT RECOGNITION SYSTEM

As mentioned in the introduction, the relationships between the descriptions used by humans and the actual compositions of objects are not so simple. To address this problem, we proposed a basic organization of ontology for interactive object recognition to consider the difference of human descriptions depending on situations [1]. In this paper, we extend this ontology to deal with various descriptions by humans. We first summarize the organization of the ontology proposed in our previous work and describe the extended parts.

The basic ontology consists of three sub-ontologies for object recognition utilizing human descriptions: the object sub-ontology, the human description sub-ontology, and the object attribute sub-ontology. Fig. 6 shows the whole basic ontology. Details about these sub-ontologies are found in [1]. Here, we describe the human description sub-ontology, which plays the major part in understanding human descriptions.

This human description sub-ontology organizes concepts of human descriptions about appearances of objects. We found that humans frequently refer to objects by their color and shape. It is essential to consider the two issues. One is what the description indicates. Humans may use the same descriptions to indicate different physical compositions. A description "marui (round)" is used for both a 2-D circle and a 3-D sphere. However, there is an order to the linking between a word and a concept of the composition of objects. We therefore need to consider the multiple meanings of human descriptions as well as the order among them. The other is where the description indicates. Humans may use the same descriptions to indicate objects viewed in different ways. Humans sometimes specify where their descriptions indicate such as "round when viewed from above" for cylindrical objects. However, they often omit such modifiers. The system should consider this in its recognition process. Considering these two issues, we organized the human description sub-ontology.

The category of "Human Description" consists of the "What-Description," corresponding to descriptions of object attributes such as "red" and "marui (round)," and the "Where-Description," corresponding to descriptions specifying perspective aspects such as "viewed from above" (although in general these tend to not be explicitly mentioned). We used Hozo [9] to construct the ontology. In Hozo, a 'part-of' relation and an 'attribute-of' relation are indicated by 'p/o' and 'a/o', respectively, as in Fig. 6.

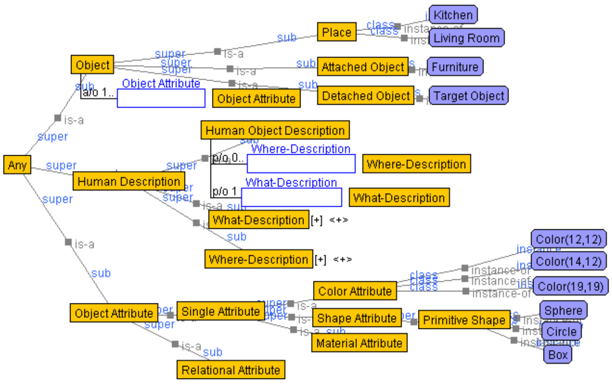


Fig. 6: Whole ontology for interactive object recognition.

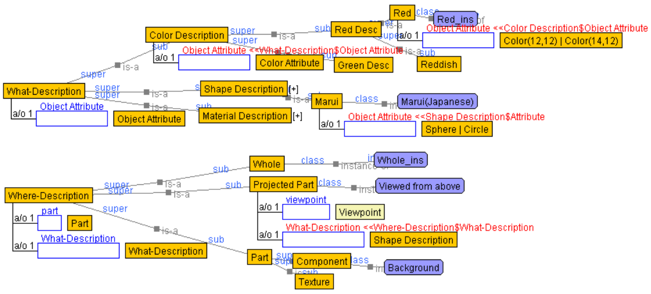


Fig.7: Human description sub-ontology.

Fig. 7 shows the details of the two concept categories of the "What-Description" and "Where-Description." The "What-Description" category stores concepts pertaining to the composition of objects. It has an "Object Attribute" slot as the attribute-of-relation (Note that "attribute-of" is the term used in Hozo to specify the "attribute" of the class (in Hozo's use of the term, such a slot specifies a "role" that the class plays). We use the expression "Object Attribute" to represent attributes of the object such as color and shape to avoid confusion). The "Object Attribute" slot specifies the possible range of values or entities of the object attribute. In Fig. 7, for instance, the description "Color(12,12) |Color(14,12)" means that the object attribute for the description "Red" should be either "Color(12,12)" or "Color(14,12)." The initial possible range can be changed through a history of interaction with the user. If multiple candidate objects are detected in the scene, the system must select the target object by considering the priority order among the possible object attributes. To do this, we align the object attribute values or entities in the order of priority. For example, in the case of the description "Color(12,12) |Color(14,12)" in Fig. 7, "Color(12,12)" is the first choice for the what-description "Red" unless any other information is given. How to estimate this priority is mentioned in the Section VI.

"Where-Description" stores the concepts pertaining to the perspective that humans mention with regards to viewing the object. It has a "Part" slot as the 'attribute-of-relation' slot. Usually, the object attribute is that of the whole object. However, a description such as "viewed from above" indicates that the object attribute mentioned concerns a projection of the 3D object viewed from the perspective in question, which is specified in the "Viewpoint" a/o slot.

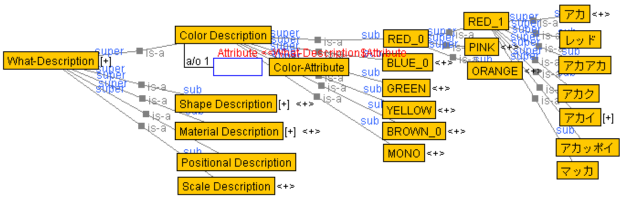


Fig. 8: What description sub-ontology.

In this paper, to deal with various descriptions, we extend the what-description sub-ontology by adding the descriptions that humans may use to existing basic ones such as “red”, “marui (round)” and “made of wood”. We added 140 descriptions as sub-classes of what-description. Some were collected in the experiment described in Section III. We collected the rest in a more systematic manner. We manually examined a dictionary for speech analysis [14] to extract words that can be used to describe object appearances. Fig. 8 shows the part of the what-description ontology.

In Fig.8, the rightmost entries indicate actual descriptions in Japanese used by humans. We hierarchically organize the color-description sub-ontology by grouping manually depending on the similarity of color. This grouping enables the system to refer to other descriptions similar to a given description. We organize the shape-description as proposed in [2].

If a given description is not included in the ontology, the system cannot recognize objects using the description. In order to respond any user’s request, it is desirable that the system (ontology) includes as many descriptions as possible. However, manual extension of the ontology is labor intensive. Therefore an important future project is developing an automated extension of the ontology through interaction with users.

V. RECOGNITION SYSTEM BASED ON HUMAN OBJECT DESCRIPTION ONTOLOGY

We developed a system that recognizes objects based on our ontology. The system mainly consists of three modules: the language analysis module, the image processing module, and the recognition module. The language analysis module converts the user’s natural language inputs into the commands for the robot system using our ontology and the Japanese dependency structure analyzer Cabocha [14]. The image processing module provides functions for autonomous object recognition, image segmentation, and attributes detection. The module processes RGB-D images acquired by a Kinect sensor. The recognition module recognizes the desired object by using the results from the language analysis module and the image processing module. Below we summarize the main processes of the image processing module and the recognition module.

Image processing module: It is relevant here to briefly summarize our previous work on the attribute detection of objects [1], in the context of discussing the attributes covered. Fig.9 shows an example of attribute detection results.

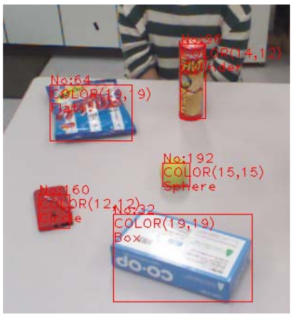


Fig. 9: An example of attribute detection.

- **Color attribute detection:** In our previous study, we found that humans often describe multicolor objects in terms of only one color, usually that of the background or that occupying the largest surface area of the object. We implemented a module to detect these colors. The main algorithm is based on color segmentation and its convex hull. We separate YCbCr space into 20 classes (colors). For each of the 20 colors, the pixels occurring with values in the ranges covered are extracted as the area of the color. We next obtain the convex hulls. We determine the color of the largest area and the background by tallying the color instances in those areas. The example of Color(14, 12) in Fig. 9 indicates that the largest-area color is color 14 and the background color is color 12.
- **Shape attribute detection:** We prepared several processes to detect basic primitive shapes such as sphere, cylinder, and box, by using a model fitting method using 3D data acquired from the Kinect Sensor. Since the system possesses 3D data about the objects, it is able to respond to descriptions such as “viewed from above.”

Recognition module: The recognition module recognizes the desired object by using detected objects, their attributes and user’s descriptions. The system first tries to recognize the desired object requested by the user. If this fails, the system begins interactive object recognition.

If any description is given, such as “red”, the system searches the concept of the given description on the ontology and retrieves its “Object Attribute” slot. The system detects any objects that have an object attribute included in the “Object Attribute” slot of a given description. If the system recognizes that the current object attribute can be included in the range through interaction, then the system adds it to the range list and updates the ontology accordingly. If the system finds multiple objects, the object which has the value of the highest priority is selected as the target object.

If any candidate is not detected by using the given description, the system searches the ontology for any related descriptions. Since we organize the human description sub-ontology by grouping human descriptions by their similarities, we utilize this hierarchal structure. First, the system refers to the sub concepts of the original description concept. If this fails to detect the target object, the next choice is to refer to the super concept of the original description. This referring process continues recursively until the system detects any candidate or the trace back process reaches the root concept.

VI. UPDATING ONTOLOGY THROUGH INTERACTION WITH USERS

Humans may use the same descriptions to indicate multiple physical properties such as “marui” for circles and spheres. We have found that there exist some priorities among them [2]. For example, “marui” may indicate spherical objects if both spherical and circular objects exist in the scene. In the human description sub-ontology, each human description has a slot for specifying its physical properties. The priorities among different physical properties are indicated in this slot for multiple meaning descriptions. In [2], we examined the priority for the Japanese expression “marui” by experiments using human participants. However, examining priorities for numerous expressions through experiments is labor intensive. Thus we add to the system, the ability to learn such priorities through a history of interaction with users.

We adopt Scheffe’s method of paired comparisons [14] to determine such priorities. In the example scene shown in Fig. 2, a description “marui (rounded) object” may indicate both the middle “cylindrical shape” object and the right “circular shape” object. In such a case, if the user describes the target object as “marui” and the system finds through interaction with the user that the target object is the right circular object, the system knows that circles have a higher priority than cylinders. The system accumulates such paired information for each concept of descriptions through interaction with the user.

The actual process of priority assignment is as follows. When comparing attribute A and attribute B, if an instance is observed to indicate “A>B,” a score of 1-point is given. On the other hand, if an instance is observed to indicate “A<B,” a score of -1-point is given. The system calculates the scores of all attributes that the description may indicate by using Scheffe’s method of paired comparison.

Fig. 10 shows an example of the description “marui (round)”. In this scene (Fig.10 (a)), there are three objects that the description “marui (round)” may indicate. If the user describes the target object as “marui” and the system recognizes through interaction with the user that the ball is the user’s desired object, the system updates the priority for description “marui.” Here, the system understands that “SPHERE” is felt more “marui (rounder)” than “CIRCLE” and “CYLINDER”. This result gives two instances for the description “marui”, “SPHERE>CYLINDER” and “SPHERE>CIRCLE”. If available instances are only these two, the priority scale for “marui” is calculated as Fig. 10 (b). The priorities are normalized into [0.0 1.0] in this system. From these data alone, the system cannot tell which is more “marui” “CIRCLE” or “CYLINDER”. However, if the system is used for some period and accumulates more paired comparison results, it can obtain more precise priority scale by using Scheffe’s method of paired comparison. Fig. 10 (c) shows the priority scale for “marui” estimated from the experimental data reported in [2].

VII. RECOGNITION EXPERIMENT

In this paper, we extend the ontology to deal with various descriptions. In addition, we hierarchically organize similar descriptions (concepts) so that the system can effectively use related descriptions in the event that the system cannot detect the desired object from the user's original description. We also modify the system to update the priorities among the different attributes that can be described in the same way. We performed experiments to examine the effectiveness of these extensions. We concentrated on whether or not the system could recognize the objects for given descriptions and thus did not use a robot in the experiment.

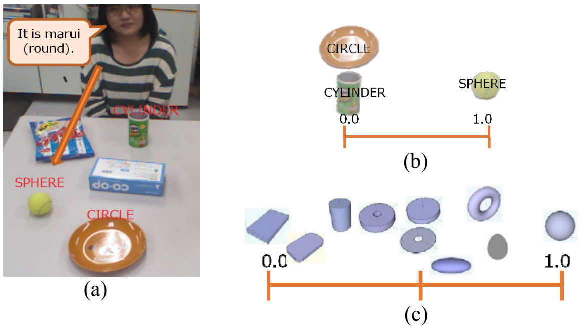
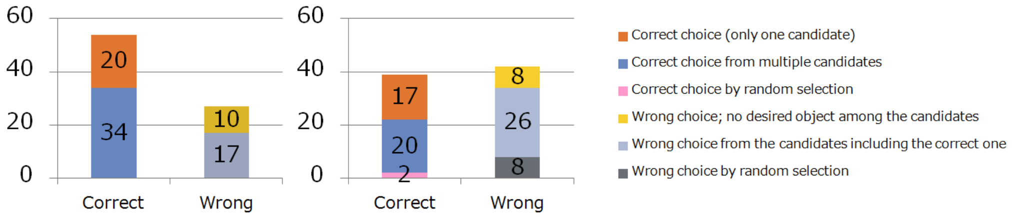


Fig. 10: An example of estimation of priority.

We use the data set acquired in the experiment in Section III. We divided the data into training and test data sets. In the training step, we updated the human description sub-ontology, such as the correspondences of human descriptions to attributes of objects and their priorities by using the training data. Since the data set contained a set of descriptions and the attributes of the desired and other objects, the system was able to reassign the priorities by the method described in the previous section. In this experiment, we considered only color and shape attributes, which are the two most often mentioned attributes.

In the test step, using the test data set, we examined whether or not the system could detect the desired objects from given descriptions. In this experiment, the system did not use multiple interactions. It used only the description obtained in the experiment in Section III in each case. We compared the proposed system and the conventional system. The conventional system had only the information about the correspondences between the human descriptions and the attributes of objects. Since it did not have the priority information, if it detected multiple objects as candidates, the system selected randomly one of the candidates. Also since the correspondences were not organized hierarchically as in the ontology, when the system could not detect candidates, it randomly selected an object from all detected objects.

Fig. 11 depicts the result of the experiment. We tested 81 cases. As shown in Fig. 11(a), the proposed system recognized the desired objects correctly in 54 cases / 81 cases (66%). In 20 cases among them, the system uniquely recognized the desired objects, in the rest of the 34 cases, it recognized the desired objects from multiple candidates based on the prioritized attributes. In 27 cases, the recognition failed. However, in 17 cases among them, the system detected the desired objects as the candidates.



(a) Proposed Method (b) Conventional Method

Fig.11 Result of recognition experiment.

Comparing the proposed system and the conventional system, since the proposed system referred to the sub/super concepts of the original description, the number of times that the system detected the desired object as candidates increased (53 → 61). Moreover, since the proposed system had information about the priorities of attributes in each description concept, the rate of the system to select the desired object from multiple candidates increased (20/46→34/51). These results confirm that the proposed system can recognize desired objects more efficiently compared with the conventional system.

In this experiment, we evaluate only the recognition system based on our ontology, not an entire robotic system. In order to evaluate the total system quantitatively, we need to consider all errors generated by speech recognition, segmentation (object detection), attribute detection and human robot interaction. We explore these issues in future work.

VIII. CONCULUSION

It is difficult to recognize objects autonomously and accurately under various conditions, a problem that we have been seeking to address through our interactive object recognition system. In our system, the robot asks the user to verbally provide information about an object that it cannot detect autonomously, thereby enabling it to identify the item and fetch it. However, an obstacle was presented by the fact that humans tend to describe different objects in various ways. In order to address this problem, we proposed constructing an ontology for interactive object recognition. In this paper, we extended the ontology to deal with various forms of human description, and developed an object recognition system based on this ontology. Experimental results confirmed that the system could efficiently recognize objects by utilizing this ontology.

An ongoing challenge is to utilize the ontology for generating utterances of the system (robot). Utilizing the ontology for interaction, we will improve the ability of the total system. In the interactive recognition framework, if the target object can be segmented out in the scene, then the system can recognize it without fail. The segmentation process is therefore vitally important. In addition to improving the segmentation process through using a set of color and depth images, we are also planning to make use of interaction in correcting segmentation results.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number 26240038.

REFERENCES

[1] H. Fukuda, S. Mori, Y. Kobayashi, Y. Kuno, D. Kachi, "Object recognition for service robots through verbal interaction based on ontology," *9th International Symposium on Visual Computing (ISVC2013),* Lecture Notes in Computer Science, vol.8033, pp.395-406, 2013.

[2] S. Mori, H. Fukuda, Y. Kobayashi, Y. Kuno, D. Kachi, "Recognizing Objects with Indicated Shapes Based on Object Shape Ontology", *IIEEJ Trans. Image Electronics and Visual Computing*, vol.42, pp.477-485, 2013. (In Japanese.)

[3] T. Winograd, "Understanding Natural Language," Academic press, 1972.

[4] P. McGuire, J. Fritsch, J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, H. Ritter, "Multi-modal human machine communication for instruction robot grasping tasks," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1082–1089, 2002.

[5] M. Takizawa, Y. Makihara, N. Shimada, J. Miura, Y. Shirai, "A service robot with interactive vision - object recognition using dialog with user," *Proc. Int. Workshop Language Understanding and Agents for Real World Interaction*, pp. 16–23, 2003.

[6] S. Kobayashi, S. Tamagawa, T. Morita, T. Yamaguchi, "Intelligent humanoid robot with Japanese wikipedia ontology and robot action ontology," *Proc. The 6^{th} ACM/IEEE Int. Conf. on Human Robot Intetaction*, pp. 417–424, 2011.

[7] N. E. Maillot, M. Thonnat, "Ontology based complex object recognition," *Image and Vision Computing*, vol. 26(1), pp. 102–113, 2008.

[8] S. Dasiopoulou, V. Mezaris, I. Kompatsiaris, V. K. Papastathis, M. G. Strintzis, "Knowledge-assisted semantic video object detection," *IEEE Trans. Circuits Systems Video Tech*, vol. 15(10), pp. 1210–1224, 2005.

[9] H. Holzapfel, D. Neubig, A. Waibel, "A dialogue approach to learning object descriptions and semantic categories," *Robotics and Autonomous Systems*, vol. 56(11), pp. 1004–1013, 2008.

[10] K. Kozaki, Y. Kitamura, M. Ikeda, R. Mizoguchi, "Hozo: An Environment for Building/Using Ontologies Based on a Fundamental Consideration of "Roleh" and "Relationship"," *Proc. of the 13th International Conference Knowledge Engineering and Knowledge Management (EKAW 2002)*, pp. 213–218, 2002.

[11] J. J. Gibson, "The Ecological Approach to Visual Perceptionm," Routledge, 1986.

[12] L. Cao, Y. Kobayashi, Y. Kuno, "Spatial-Based Feature for Locating Objects," *Proc. ICIC 2012. Lecture Note Computer Science*, vol. 7390, pp. 128–137, 2012.

[13] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, et al. "Mechatronic design of nao humanoid," *Proc. IEEE Int. Conf. on Robotics and Autonation,* pp. 769–774, 2009.

[14] T. Kudo, Y. Matsumoto, "Japanese Dependency Analysis Using Cascaded Chunking," Proc. the 6th conference on Natural language learning (CoLLL-02), pp. 63-69, 2002.

[15] H. Scheffe, "An Analysis of Variance for Paired Comparison",Journal of the American Statistical Association, vol. 47, issue 259, pp. 381–400,1952.