Human Robot Interaction through Simple Expressions for Object Recognition

Al Mansur, Katsutoshi Sakata, Tajin Rukhsana, Yoshinori Kobayashi and Yoshinori Kuno

Abstract— **Service robots need to be able to recognize and identify objects located within complex backgrounds. Since no single method may work in every situation, several methods need to be combined. However, there are several cases when autonomous recognition methods fail. We propose an interactive recognition method in these cases. To develop a natural Human Robot Interaction (HRI), it is necessary that the robot should unambiguously perceive the description of an object given by human. This paper reports on our experiment in which we examined the expressions humans use in describing ordinary objects. The results show that humans typically describe objects using one of multiple colors. The color is usually either that of the object background or that of the largest object portion. Based on these results, we describe our development of a robot vision system that can recognize objects when a user adopts simple expressions to describe the objects. This research suggests the importance of connecting 'symbolic expressions' with the 'real world' in human-robot interaction.**

I. INTRODUCTION

Service robots have attracted the attention of researchers for their potential use with handicapped and elderly people. We are currently developing a service robot that can bring a specific or a general class of household objects requested by the user. The robot receives instruction through the user's speech, and should be able to carry out two tasks: (1) detect a specific object (e.g., 'coke can'), and (2) detect a class of objects (e.g., 'can'). The robot needs a vision system that is able to recognize various objects in complex backgrounds in order to carry out the two tasks mentioned above. There is no single object recognition method that can work equally well on various types of objects and backgrounds perceived by a service robot. Rather, it must rely on multiple methods and should be able to select the appropriate one depending on the characteristics of the object. An autonomous recognition system for service robot using multiple methods has been initially proposed in [1] and extended in [2]. However, as the recognition rate of autonomous method is not 100%, it is desirable to improve the recognition performance by incorporating user interaction.

Dating back to the work by Winograd in [3], there has been a great deal of research that incorporates Human Robot Interaction (HRI) for scene understanding [4]–[6]. In these studies, objects that can be described by simple word combinations (such as 'blue box' or 'red ball') are considered. However, in our application domain, objects are usually more complex. For example, we may want the robot to bring us a packet of potato chips, in which the package has various colors. In spite of the complexity of the objects, humans usually describe them by simple words as we observed in the experiments described later.

The motivation of our work arises from the aforementioned view of human interface research. Our work is critical within the domain of computer vision research. We are working to develop an interactive object recognition system [7], [8]so that it could be used when the robot is unable to recognize objects by itself alone. In our previous research, we also dealt with simply describable objects. This paper shows the ways the system can be extended to recognize complex objects. Our work is an attempt to connect 'symbolic expressions' with the 'real world' in actual situations and to integrate such a HRI to an autonomous system.

In section 2, we briefly discuss the implementation of the autonomous method. Interactive method is presented in sections 3. Experimental results are given in section 4 and finally we conclude the paper in section 5.

II. AUTONOMOUS OBJECT RECOGNITION

Here we briefly describe the autonomous object recognition method. Details are given in [2].

A. *Object Categorization*

Objects encountered by service robots can be described by their color, shape, and texture. By 'texture' we mean the pattern (not necessarily regular and periodic) within the object contour. For example, in our notation, the label on a bottle is its texture. We used three features for recognition: intensity, Gabor feature, and color. We split the objects into five categories depending on characteristics. Textureless simple-shape objects are named category 1. We need to use shape features to recognize such objects. Kernel PCA (KPCA) in conjunction with Support Vector Machine (SVM) can be used in this case. In category 2, some objects have textures although these textures do not characterize them and the texture contents of different members of the class are not the same. Even some members may have texture-free body. As a result, we need to use information regarding their shapes in order to describe them. Using SIFT, any specific textured object of this category can be recognized. To recognize a texture-free specific object or a class of this category we use KPCA+SVM. Since these objects are shape-based, we should use Gabor feature because it works well on objects with different textures. Category 3 and 4 objects have similar textures and these textures are required for their recognition.

Authors are with the Department of Information and Computer Sciences, Saitama University, Saitama 338-8570, Japan. {mansur,sakata,tajin,yosinori,kuno}@cv.ics.saitama-u.ac.jp

Examples of these two categories include fruit (e.g. pineapple) and computer keyboards. KPCA+SVM based method works well on this type of objects. In our experiments, we found that intensity feature works better than or the same as Gabor feature for some objects of this type. They are named as category 3 objects in this paper. For other objects of this type, Gabor feature obtains better recognition rate and they are designated as category 4 objects. Many of the texture classification methods [9]–[11] use Gabor filters for feature extraction. Robust feature extraction using Gabor filters requires a large set of Gabor filters of various scales and orientation. This makes the computation huge. In this respect, intensity feature is desirable due to its simplicity and speed. Categorization helps us to avoid time consuming feature extraction process wherever possible. Category 5 objects have similar color histograms. We use a combination of color and intensity features for their recognition.

B. Recognition Methods

Four different methods have been integrated for the autonomous recognition system. As shown below, different methods are employed for different object categories according to object characteristics.

Method 1: Used for recognition of specific object from category 3, category 4 and category 2 (textured).
Method 2: Used for recognition of specific object and class from category 1, specific (texture-free) object and class from category 2 and class from category 4.
Method 3: Used for class recognition of category 3 objects.
Method 4: Used for class recognition of category 5 objects.

We use SIFT following [12] in method 1. In method 2, we apply a battery of Gabor filters to each of the training and test images (grayscale) to extract the edges oriented in different directions. Dimensionality of these Gabor feature vectors are reduced by KPCA [13]–[15] and are used to train a SVM classifier. In method 3, KPCA features are derived from the intensity images and then a SVM classifier is trained. In method 4, an SVM classifier is build using color features. Here, another intensity based SVM classifier is trained (as in method 3) and used to reduce the false positive results of the first classifier. Details of these four methods and the algorithm for selecting one of these methods automatically are given in [2]. When only one object per class is available, the robot uses method 1 if the object is textured, or color histogram if the object is texture-free.

III. INTERACTIVE OBJECT RECOGNITION

We are implementing our algorithms on our experimental robot Robovie-R Ver.2 [16]. This 57 kg robot is equipped with three cameras (2 pan-tilt and one omnidirectional), wireless LAN, various sensors, and two 2.8 GHz Pentium 4 processors. Our service robot has access to a few variants of a certain class of objects and its training set is usually small. In spite of a small training set we achieved a reasonable recognition rate. However, the recognition methods are not 100% accurate. It is necessary to improve the recognition performance by any feasible way. In our application the robot user is assumed to be a physically disabled person with speaking capability. The robot is designed to help him or her bring an object upon request. When the robot fails to find the object it may ask the user to assist it using some short, user-friendly conversation. We have already developed some interactive object-recognition methods [7], [8] for the recognition of simple single-color objects in plain background. Here we extend these works for complex objects in complex backgrounds.

A. Grammar and Sentence Pattern

In interactive object recognition, robots have to understand and analyze the user's instruction. Instructions are grouped into ten categories (Table I). In order to build a sentence pattern, words or phrases must be selected from the predefined vocabulary list. We limit the vocabulary list to avoid ambiguity during speech recognition. The user must follow the sentence structure (Table I) and choose the words from the registered word list (Table II) for the corresponding vocabulary type to communicate successfully with the robot. Optional words, though not required, provide more natural speech. For example, the user can say, "Get me a noodle." This satisfies the grammar of 'Object ordering: class' and it uses the vocabulary from Phrase 1 and Object Name. Likewise, the user could also say, "May I have the Nescafe (brand name) Coffee jar?". The vocabularies are listed in Table II. Language processing presented here is not state of the art. We developed it for checking the effectiveness of the interactive object-recognition technique. At present, user instruction is given through a keyboard and the robot response is generated by text to speech. We will use the techniques developed by researchers on natural language understanding in the future.

B. Object Description by Human: An Experiment

In order to carry out our experiment, we assembled ordinary objects that we may want a service robot to bring, such as food and drink (Figure 1). Humans can usually recognize such objects when the object name (e.g. potato chips) is mentioned. We examined how humans describe objects when they were not allowed to mention the object by name.

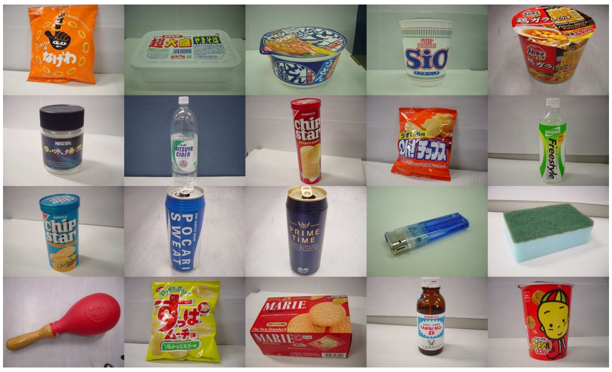


Fig. 1. Object Examples

TABLE I
GRAMMAR

Purpose	Sentence structure	Example
Feedback	Feedback	Yes/No
Object Ordering: class	Phrase 1+a/an+ Object Name	Get an apple.
Object Ordering: specific	Phrase 1 + (the) + (Specifier/color) +Object Name (at least one 'the' or 'specifier/color' is required	Get my cup.
Positional information 1	Verb 1 + Positional adjective/ Preposition 1 + (Article) + Specifier/color + Object Name	Look at the left of Seafood noodle.
Positional information 2	Verb 1 + Positional adjective/ Preposition 1 + that + (Object Name)	Look behind that.
Positional information 3	Verb 1 + Preposition 2 + (Article) + (Specifier/color) + Object Name + (and) + (Article) + (Specifier) + Object Name	Look between Pepsi can and tea bottle.
Positional information 4	Phrase 1 + Positional adjective/ Preposition 1 + (Object Name/object/one)	Get the left one.
Instruction to point	(Phrase 2) + Verb 2	Please show me.
Instruction to find	(Phrase 3) + Verb 3 + (Article) + Specifier/color + Object Name	Can you find the wooron tea bottle?
Object description with single color	Color	Red

Figure 2 is a representation of the experimental setting. Ten pairs of participants took part in the experiment. A board was used to separate them. We placed about 20 objects (among those shown in Figure 1) on top of the table on participants B's side. We also placed one of the same objects on participant A's side. We asked participant A to describe the object without naming the object, and asked participant B to choose the correct object. Participant A continued with the description until participant B selected the correct object. We videotaped the experiments and examined the descriptions. We examined 227 utterances that described an object.

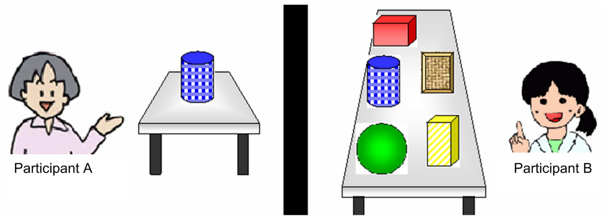


Fig. 2. Experimental Setting

TABLE II
VOCABULARY

Type	Registered words
Feedback	Yes, No
Phrase 1	May I have, Can I have, Can I get, (Please) get (me), (Please) Bring, I'd like, I would like, Give (me)
Phrase 2	Please, Could you (please), Can you (please)
Phrase 3	Could you, Can you
Verb 1	(Please) look (at/to), (Please) check
Verb 2	Show (me), Point
Verb 3	Find, See
Specifier	My, Coke, [brand name], etc.
Positional adjective	Left, Right
Preposition 1	Front, Behind, Top, Bottom
Preposition 2	Between
Object name	Noodles, Cup, Jar, Bottle, Coffee jar, etc.
Color	Red, Green, Yellow, etc.

We first classified the utterances into vision-based and knowledge-based. The former are descriptions that can be obtained through vision, such as 'red', and 'round'. The later are descriptions in which prior knowledge is needed, such as 'food' and 'juice'. Figure 3 shows the percentages of vision-based and knowledge-based utterances. If any utterances included both types of description, they were counted in both categories. As shown in this figure, participants more often used vision-based than knowledge-based descriptions. Since it is still difficult for computer vision to recognize objects based on knowledge-based descriptions, we concentrated further analysis on vision-based descriptions.

The results reveal that participants used vision-based descriptions including color, shape, texture/pattern, size, attachment, and material. The most frequent was color (Figure 4).

Based on these results, we then more closely examined how participants used color. Most of the objects shown in Figure 1 have multiple colors. Participants, however, most frequently used one color (Figure 5).

Further, among descriptions using only one color, our next question concerns how the color is determined. Figure 6 shows this result. Participants use two criteria: (1) background color of the object, (2) color that shares the largest area of the object. Participants used the color satisfying both (36.3 %), the background color but not the color with the largest area (37.5%), and the color with the largest area but not the background color (8.7%).

These results can be summarized as follows. Humans often describe an object by one of the colors of the object. This color is typically either the background color or the color with the largest area. These results imply that robots (vision systems) should be designed to locate objects based upon the background color or the color with the largest area. In this paper, this color is denoted as 'base color' .

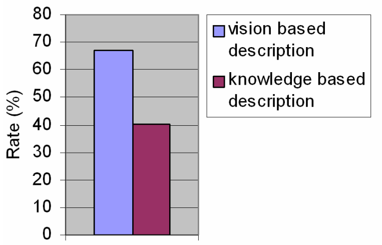


Fig. 3. Utterance Classification into Vision based and Knowledge Based Descriptions

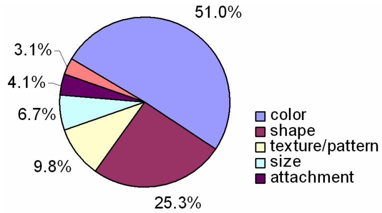


Fig. 4. Classification of Vision based Descriptions

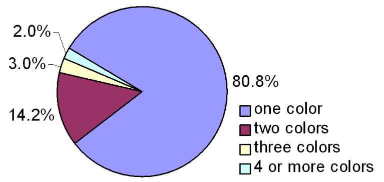


Fig. 5. Number of Colors Mentioned in an Utterance

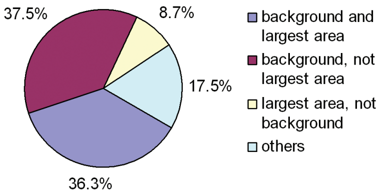


Fig. 6. Color Used to Describe Multi-color Objects

C. Finding Base Color

Ordinary objects usually have multiple colors. Yet, we usually describe such objects with one color as suggested by the results described in the previous section. Robots (interactive object recognition systems) must be able to detect objects by identifying the color.

We have developed an image processing method to detect objects when a color is mentioned in the ways reported in the previous section. The processing steps are as follows.

1. Extract object regions by background elimination.
2. Apply color segmentation and calculate each region's area.
3. If there is only one large color region, memorize the color and go to 8. Otherwise, go to 4.
4. Extract outlines of the regions.
5. Detect feature (corner) points on the extracted outlines.
6. Obtain convex hulls for the detected feature points.
7. Memorize the color with the largest convex hull.
8. Output the region of the memorized color.

In this algorithm, if there is a distinctively large color area in the image, the part is outputted as the target object (Steps 2-3). If there are multiple large regions, convex hulls for these regions are obtained and the object for the largest convex hull is outputted (Steps 4-7).

Figures 7(a)- 7(c) show an example. In this case, the yellow region is far larger than other color regions, and the object is treated as a yellow object. The yellow part is the largest area region and is the base color region. In the human experiments, all participants described this object as 'yellow'.

Figures 7(d)- 7(f) show another example. In this case, the red region area is 6,712 pixels and the yellow region area is 6,661 pixels. Since both are large, convex hulls for these regions are examined. Figure 8 shows the processing results for the red and yellow regions. From these results, 'red' is determined as the base color. The system considers this object as a red object and returns this object if a user requests a red object. If a user requests a yellow object, the system first examines other yellow objects. If there are no other objects, the system considers this object as a candidate of the target object. In the description experiment, nine participants described this object as 'red'. One participant described this as 'red and yellow'.

D. Use of Spatial Information

In some cases, color information alone is not enough to make the robot recognize a target object. For example, if multiple same color objects exist in the scene, color information alone cannot produce successful recognition and the user has to provide other attributes. We use spatial information of the target object as another attribute. The key idea is to use the position of the target object with respect to an object known to both the user and the robot or position of the target object among multiple candidates. Example instructions include "Look at the left of coffee jar", "Look at the right of white cup", "get the left one". We discuss this type of interaction in the experiments section.

E. Integration of Autonomous and Interactive Methods

When the robot receives an instruction to bring an object, at first, it uses the appropriate autonomous technique to identify that object. If it fails to identify the object correctly, it uses the interactive method using color information. Details of the integration of the two methods is shown in Figure 9. However, in some cases, both methods fail. In such cases, robot asks the user to provide additional attributes of the object such as positional relationship with respect to a known object.

IV. EXPERIMENTS

A. Experiments using Autonomous Methods

First, we evaluate the autonomous object recognition techniques using objects from the Caltech database (available at www.vision.caltech.edu). We obtain satisfactory recognition performance for different categories when appropriate

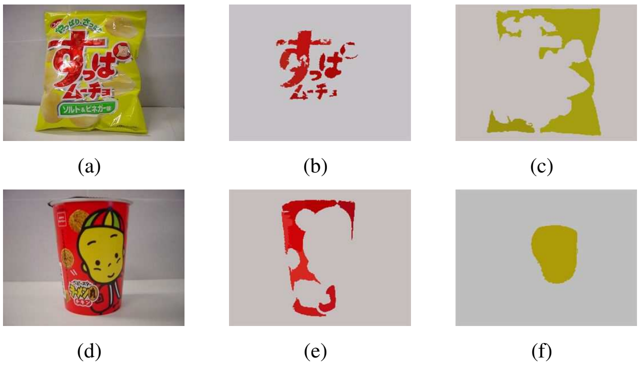


Fig. 7. (a) Object 1 (b) Red Region (c) Yellow Region (d) Object 2 (e) Red Region (f) Yellow Region

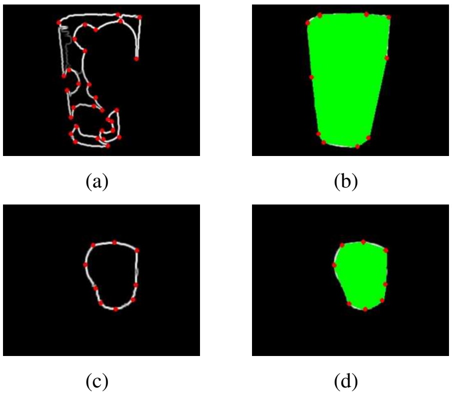


Fig. 8. Computation of the Area of Red Region ((a)-(b)) and Yellow Region ((c)-(d))

methods were used. These results have been shown and class recognition performances of three methods have been compared in [2].

Next, we perform experiments with daily objects (see Figure 1) placed in home scenes. These results confirm that our methods can recognize objects in our application domain with reasonable success rates.

B. Experiments using Integrated System

We have integrated our autonomous object recognition system and the method described in the previous section to develop a robust vision system for a service robot. The system first tries to recognize an object when a user mentions an object name. If the autonomous system fails, the robot asks the user to supply visual or spatial attributes of the object. An example of interaction is given below. Here, the robot detects a 'ramen (noodle) snack' as shown in Figure 11.

User: Get the 'ramen snack'.
Robot: I don't know 'ramen snack'. What color is it?
User: Red.
Robot: I found one.
User: Show me it.
Robot: (points at the object) Is this correct?
User: Yes.

In another experiment (Figure 12), the user asks the robot to get the 'Maxim' coffee jar. However, the user mentioned only 'coffee jar'. The robot recognizes two objects as 'coffee jar' and needs confirmation from the user. User mentions the color of the desired coffee jar. Conversation is as follows:

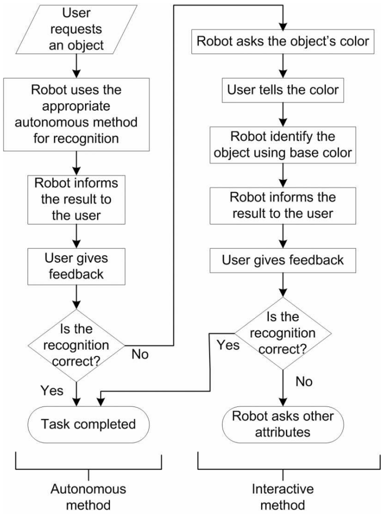

User requests an object
Robot uses the appropriate autonomous method for recognition
Robot informs the result to the user
User gives feedback
Is the recognition correct?
No
Yes
Task completed
Robot asks the object's color
User tells the color
Robot identify the object using base color
Robot informs the result to the user
User gives feedback
Is the recognition correct?
Yes
No
Robot asks other attributes
Autonomous method
Interactive method

Fig. 9. Integration of Autonomous and Interactive Methods

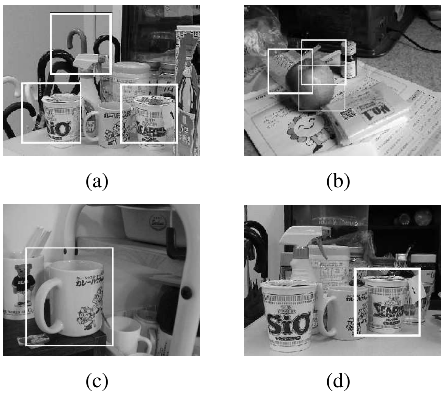


Fig. 10. (a)-(b) Class Recognition Results: (a) Cup Noodles (b) Apple (c)-(d) Specific Object Recognition results: (c) Cup (d) Cup Noodle

User: Get the coffee jar.
Robot: I found two. What color is it?
User: Red.
Robot: I found one.
User: Show me it.
Robot: (points at the red coffee jar) Is this correct?
User: Yes.

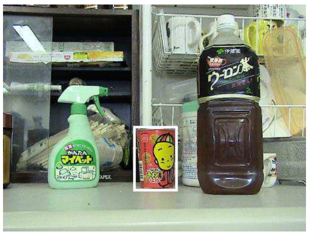
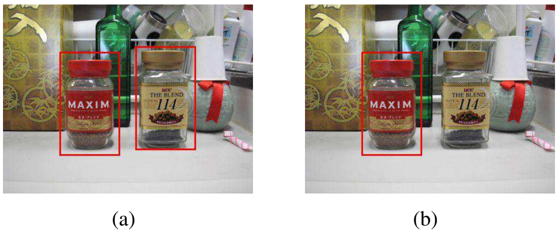


Fig. 11. Example of Interactive Object Recognition



(a) (b)

Fig. 12. (a) Robot Detects Two Coffee Jars (b) Robot Selects the Desired One through Interaction

We experimented with the other objects shown in Figure 1 to evaluate the effectiveness of the interactive method using color information. In the most of these experiments, the robot successfully recognized the requested objects. These results justify that real world multicolor objects can be properly represented by single colors in a HRI system.

In another experiment (Figure 13) the user asks for the object 'ramen snack'. Being unable to detect 'ramen snack', the robot asks for color information. The user designates the color as 'red' but the robot finds multiple red objects. The relative position of the two red objects then serves as a clue to find the target. The conversation is given below:

User: Get the 'ramen snack'.
Robot: I don't know 'ramen snack'. What color is it?
User: Red.
Robot: I found two.
User: Show me.
(Robot points the found objects.)
User: Get the right one.
(Robot finds the desired object.)

V. CONCLUSION

To make a service robot's vision system work well in various situations, we have integrated interactive recognition methods with the autonomous ones. To build a natural HRI, we investigated how humans typically describe an object without naming that object. Through experiments, we found that simple expressions are used to describe complex objects. We then developed a vision system to detect objects requested by a user through such simple expressions. This research reveals the importance of connecting 'symbolic expressions' with the 'real world' in human-robot interaction.

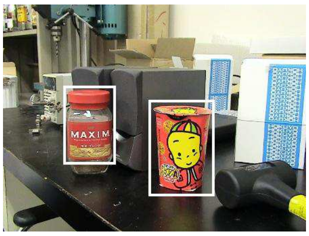


Fig. 13. Interactive Object Recognition using Spatial Information

We have elaborated only descriptions using color, as these were frequently used. However, we will further investigate other descriptors in order to develop a vision system that can work in various situations.

VI. ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Internal Affairs and Communications under SCOPE-S and by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (KAKENHI 19300055).

REFERENCES

[1] A. Mansur, M.A. Hossain and Y. Kuno, "Integration of Multiple Methods for Class and Specific Object Recognition", *in International Symposium on Visual Computing*, Part I, 2006, pp. 841-849.

[2] A. Mansur and Y. Kuno, "Integration of Multiple Methods for Robust Object Recognition", *in SICE Anual Conference*, Kagawa, Japan, 2007.

[3] T. Winograd, *Understanding Natural Language*, Academic Press, New York; 1972.

[4] T. Kawaji, K. Okada, M. Inaba, H. Inoue, "Human robot interaction through integrating visual auditory information with relaxation method", *in Proc. IEEE International Conference on Multisensor Fusion on Integration for Intelligent Systems*, 2003, pp. 323-328.

[5] P. McGuire, J. Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human machine communication for instruction robot grasping task", *in Proc. IROS*, 2002, pp. 1082-1089.

[6] M. Takizawa, Y. Makihara, N. Shimada, J. Miura, and Y. Shirai, "A service robot with interactive vision- objects recognition using dialog with user", *in Proc. First International Workshop on Language Understanding and Agents for Real World Interaction*, 2003.

[7] M.A. Hossain, R. Kurnia, A. Nakamura, and Y. Kuno, Interactive Object Recognition through Hypothesis Generation and Confirmation, *IEICE Transactions on Information and Systems*, vol. E89-D, 2006, pp. 2197-2206.

[8] R. Kurnia, M.A. Hossain, A. Nakamura, and Y. Kuno, Generation of Efficient and User-friendly Queries for Helper Robots to Detect Target Objects, *Advanced Robotics*, vol. 20, 2006, pp. 499-517.

[9] D. Dunn and W.E. Higgins, Optimal Gabor Filters for Texture Segmentation, *IEEE Transactions on Image Processing*, vol.4, 1995, pp. 947-964.

[10] A.K. Jain and F. Farrokhnia, Unsupervised Texture Segmentation Using Gabor Filters, *Pattern Recognition*, vol. 24, 1991, pp. 1167-1186.

[11] B.S. Manjunath and W.Y. Ma, Texture Features for Browsing and Retrieval of Image Data, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, 1996, pp. 837-842.

[12] D. Lowe, Distinctive Image Features from Scale-invariant Keypoints, *International Journal of Computer Vision*, vol. 60, 2004, pp. 91-110.

[13] S.Z. Li, Q.D. Fu, L. Gu, B. Scholkopf, Y.M. Cheng, H.J. Zhang, "Kernel Machine Based Learning for MultiView Face Detection and Pose Estimation", *in Eighth International Conference on Computer Vision*, 2001, pp. 674-679.

[14] C. Liu, Gabor-Based Kernel PCA with Fractional Power Polynomial Models for Face Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, 2004, 572-581.

[15] B. Schölkopf, A.J. Smola and K.-R. Muller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, vol. 10, 1998, 1299-1319.

[16] Intelligent Robotics and Communication Laboratories, http://www.irc.atr.jp/index.html