Use of Spatial Reference Systems in Interactive Object Recognition

Rahmadi Kurnia, Md. Altab Hossain, and Yoshinori Kuno
Saitama University
{kurnia, hossain, kuno}@cv.ics.saitama-u.ac.jp

Abstract

We are developing a helper robot that carries out tasks ordered by users through speech. The robot needs a vision system to recognize the objects appearing in the orders. It is, however, difficult to realize vision systems that can work in various conditions. They may find many objects and cannot determine which is the target. We have proposed a method of using a conversation with the user to solve this problem. The robot asks questions which the user can easily answer and whose answer can efficiently reduce the number of candidate objects. In our previous system, however, we assumed that there was no occlusion among objects. This paper presents an extended system that can detect target objects in occlusion cases. It is difficult to obtain some features in occlusion cases. To compensate the system for this shortage of features, we propose to use reference systems to express the positional relationships of objects. Experimental results show that the robot can efficiently detect objects through user-friendly conversation.

1. Introduction

Our aging society will in the near future require a significant increase in health care services and facilities to provide assistance to people in their environments to maintain a reasonable quality of life. One of the possible solutions to this is the use of helper robots. Thus, helper robots or service robots in welfare domain have attracted much attention of researchers [1][2]. Multimodal interfaces [3][4][5][6] are considered good interface means for such robots. Some researchers have developed a helper robot that carries out tasks ordered by the user through voice and/or gestures [7][8][9]. In addition to gesture recognition, such robots need to have vision systems that can recognize the objects mentioned in speech.

It is, however, difficult to realize vision systems that can work in various conditions. We have proposed a method of using a conversation with the user to solve this problem. When the vision system cannot achieve a task, the robot makes a speech to the user so that the natural response by the user can give helpful information for its vision system. In [8][9][10], we implicitly assumed that the scene is relatively simple so that the vision system detects one or at most a few regions (objects) in the image. Thus, even though detecting the target object may be difficult and need the user's assistance, once the robot has detected an object, it can assume the object as the target. However, in actual complex scenes, the vision system may detect various objects. The robot must choose the target object among them, which is a hard problem especially if it does not have much a priori knowledge about the object.

We have started to tackle this problem and presented our initial work in [11]. There, we have proposed a system to generate a sequence of utterances that can lead to determine the object efficiently and user-friendly. It determines what and how to ask the user by considering the image processing results and the characteristics of object (image) attributes. However, in that work, we assumed that there was no occlusion among objects and that each object consisted of only a single color part. In this paper, we present an extended system that can deal with occlusion cases. We still hold the single-color object assumption. (We are also working on a system without this single-color assumption [12]. We are planning to integrate that result with the system presented in this paper.)

If there is any occlusion, we cannot obtain certain type of features such as shape and size because the whole objects cannot be seen. This reduces the number of features that we can use for object recognition. In this paper, we propose to use spatial reference systems to solve this problem. We describe the position of an object in a certain reference system. For example, we express the position of an object by the positional relation between the object and an object that can be

easily identified, such as "the object to the left of the red object." We use this positional relationship as an additional feature to compensate the system for the shortage of features in occlusion cases.

Recently, several researchers have proposed to use the human user's assistance through speech [13][14][15] [16][17][18]. These systems may be similar to ours from outlook. However, the information that they obtain from the user is mostly symbolic-level knowledge about the world for the robot to act there, whereas ours tries to obtain the information necessary for image understanding. Our system completes the visual processing of the robot through interaction with the user.

2. Reference systems

2.1. Object features

In this paper, object recognition means to detect the region in a given image corresponding to the target object in the user's order. We consider object recognition based on object features. Thus, our object recognition procedure is to extract regions from a given image by color segmentation and to find regions that satisfy the features of the target. There are many object features such as color, size, position, shape, distance, texture and so on. In the current implementation, we use four features: color, size, position, and shape. In this problem setting, object recognition can proceed by asking the user about the features of the target object and by detecting the regions satisfying the features. In [11], we have proposed a system to generate a sequence of questions that can lead to determine the object efficiently and user-friendly. It determines what and how to ask the user by considering the image processing results and the characteristics of object (image) attributes.

In [11], we examine object features from four characteristics: vocabulary, distribution, uniqueness and absoluteness/relativeness. If we can represent a particular feature easily by word for any given object such as color and position, we call it a vocabulary(-rich) feature. If an object feature is difficult to express when several objects exist close together, we call the feature distribution(-dependent) feature. If the value of a particular feature is different for each object, we call it a unique feature. Position can be a unique feature since no multiple objects share the same position. If we can describe a particular feature by word even if only an object exists such as color and shape, we call it an absolute feature. Otherwise, we call it a relative feature.

In occlusion cases, however, these features may not show the correct properties of objects. For example, if a part of an object is occluded by another object, the shape feature obtained in a given image does not show the exact shape of the object. On the other hand, the color of an object does not change even in occlusion cases. We need to consider this characteristic in addition to the four characteristics. We call this partial observability. If we can obtain a correct feature value, even though we can see a part of the object, we call the feature a partial-observable feature. Table 1 summarizes the characteristics of the features used in the proposed system. Color is a partial-observable feature and can be used in occlusion cases. We need to be careful in using size and shape in occlusion cases. We consider position as a partial-observable feature. If we use a centroid of the region to represent the position of an object in an image, the coordinates of the centroid are affected if the object is occluded. However, we use position to identify objects and we do not need correct values. In this sense, position can be considered as a partial-observable feature.

Table 1. Feature characteristics

Characteristic	Color	Size	Position	Shape
Vocabulary	√		√	
Distribution			√	
Uniqueness			√	
Absoluteness	√	Relative	Relative	√
Partial Observability	√		√	

2.2. Reference systems

As described above, the number of features that can be used for object recognition decreases in occlusion cases. This makes object recognition difficult. We propose to use reference systems to describe positional relationships of objects to solve this problem.

Humans use reference systems to describe object positions. The relation between human's spatial cognition and language expression has been well studied in the field of psychology, linguistics, and other related fields. Levinson [19] has proposed that humans use three kinds of reference systems: intrinsic, relative, and absolute. In intrinsic reference systems,

the relative position of one object (the referent) to another (the relatum) is described by referring to the relatum's intrinsic properties such as front or back. For example, the expression such as "the book in front of you" is good enough to describe the position of the book since the front of a human body is intrinsically determined. On the other hand, in relative reference systems, we use a position of a third entity as origin instead of referring to inbuilt features of the relatum. An example is "viewed from the cup, the pen is to the left of the box." In absolute systems, neither a third entity nor intrinsic features are used for reference. Instead, we use some absolute direction specification terms, for example, such as north and south.

In addition to the above reference systems, Tenbrink and Moratz [20][21] have proposed group-based reference systems. When there are multiple same or similar objects in the scene, humans consider them as a group, describing the position of an object in the group by the spatial relation between the object and the total group. Group-based reference systems are considered to be relative reference systems using the group as a relatum.

In our research, we mainly use Levinson's relative reference systems. In some cases, we use reference systems based on the idea of the group-based reference system. When we use relative reference systems, we fix the viewpoint as that of the user. Thus, this part is omitted in utterances. We assume that the user and the robot are close together and are viewing in the same direction. The most important issue in using reference systems is to what object is used as a relatum. We have conjectured preferable conditions as a relatum as follows.

-Placed in the cluster: When some objects occlude others, they make a cluster in the image. We would like to specify the positions of the objects in this cluster. Thus, relata should be in the cluster.
-Placed around the center of the cluster: The approach of our method is to find the target object by removing unnecessary objects based on the user's answer. The robot can remove many unnecessary objects from the user's answer if the position of the relatum is around the center of the cluster.
-Single color: If an object is composed of a single color part, it is easy to specify the object by mentioning the color. This is a preferable condition for relata. In the current system, we assume that all objects are single-color objects. Thus, we do not need to mention this condition. However, even though the system can deal with multiple color objects, this condition is preferable.
-No other same color objects: In addition to the third condition, if there are no other same color objects in the cluster, the object is ideal for the relatum. Of course, we cannot always expect this case.
-Not too big: When an object is big and it covers a large part of the cluster, for example, more than a half of the cluster, it is difficult for the robot to divide the cluster into the left side and right side or up side and bottom side. It is better to avoid such objects as relata

Our basic idea is to use a relative reference system if we can find a relatum that satisfies most of the above conditions. If we cannot find an appropriate object as a relatum, we consider the cluster as a relatum following the group-based reference system proposed in [20][21].

2.3. Human observation experiments

We performed experiments to observe human behaviors before developing our dialog system. We prepared two scenes with occlusion cases as shown in Fig. 1**.** An experimenter chose an object as the target in mind. We asked ten participants to guess what was the target object by asking 'yes/no-type' and 'what-type' questions to the experimenter. We examined what features the participants used in such occlusion cases. The results of these experiments are shown in Fig. 2.

We have found the followings from the observation results.

1. Humans usually ask about color as the first question.

In the experimental scenes, there were only single-color objects because we consider only single-color objects in the current stage of research. Thus, it may be natural to ask about color at first. However, we think that we prefer to mention about color in other situations since it is easy for us to recognize color and we have rich vocabulary to express color.

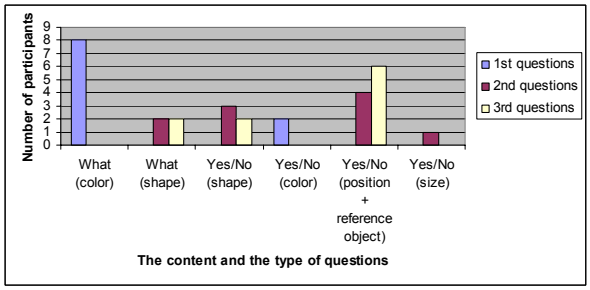
2. After asking color, humans often ask about shape to detect the object.

Humans can easily recognize the object shapes in occlusion cases although they can see only their parts. Humans can do this because they have much a priori knowledge about the objects. The robot cannot recognize the object shapes in occlusion cases as humans can because it has limited knowledge about object. As discussed in 2.1, it is better to avoid using shape in occlusion cases.

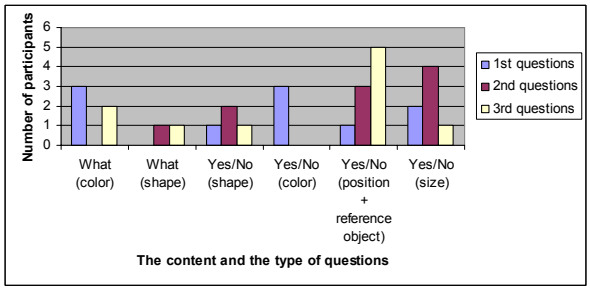
(a)

(b)

Figure 1. Experimental scenes for human behavior observation.



(a) Scene in Fig.1 (a).



(b) Scene in Fig. 1 (b).

Figure 2. Observation results of human behavior.

3. Humans often use reference systems.

The participants often chose an object as a relatum to specify the positions of other objects. We asked them about their reasons for selecting the object as a relatum. We allowed them multiple answers. The results are as follows.

- The object is unique. It looks different from the others (7 participants).
- The object is placed around the center of the cluster (5 participants).
- They know well about the object (4 participants).

We have confirmed through these experiments that there are cases that humans use reference systems in such situations and that our conjectures described in 2.2 are reasonable.

3. Dialog generation

The basic strategy for generating a dialog is *ask-and-remove*. The robot asks the user about a certain feature. Then, it removes unrelated objects from the detected objects using the information given by the user. It iterates this process until only an object remains. If there are both separated and cluster objects in the scene, the robot would like to know whether or not the target object is in the cluster. Thus, in such a condition, the first question is if the target object is in the cluster. If the target is not in the cluster, our previous system [11] can deal with the case. Otherwise, the robot tries to identify the target object in the cluster as follows. As in the same way for the separate object case [11], the robot divides the current situation into two cases according to the number of detected regions: the few-object case where the number of regions is equal to or less than three and the many-object case where that is greater than three.

3.1. Many-object case

The robot generates its utterances for dialog with the user as follows. First, it classifies the features of all regions into predetermined classes. For example, it assigns a color label to each region based on the average hue value of the region. For color, it classifies them into seven colors: blue, yellow, green, red, magenta, white, and black.

Then, the robot computes the percentage of the number of objects in each class to the total number of objects. It classifies the situation of each feature into two categories depending on the maximum percentage: the variation category and the concentration category. The variation category is the case where the maximum

percentage is less than one third. The concentration category is the case where that is more than one third for at least one class. These values are experimentally determined.

If the robot can obtain information about any feature that falls under the variation category, the information can reduce many unrelated objects among the regions (object candidates). Therefore, the first rule for determining what feature the robot chooses for its question to the user is to give a priority to the variation category features. If no such feature exists, the concentration category features are given the second priority.

1. Case with variation category features

If there is any feature that falls in the variation category, the robot asks the user about the feature. Actually, color is the only feature in the current implementation that the robot can ask directly of the user in occlusion cases. Thus, if color falls in this category, the robot asks, "What is the color of the target object?" The robot can ask the user 'what-type' questions because color is a vocabulary-rich feature [11].

2. Case with concentration category features

This is the case that many regions are similar in several respects. Thus, the robot plans to use a reference system to obtain positional information. The case can be divided into the following three situations.

Situation 1. Use of a relative reference system

The robot first picks up a region around the center of the cluster as a candidate of the relatum if there is no large region covering more than a half of the cluster. If there are no other regions that have the same or similar color of the candidate, the robot uses the region as the relatum. For example, in the case shown in Fig.3, the robot asks the user, “Is the target object to the left side of the red object?"

If there are any other same or similar color objects in the cluster, the robot examines if it can easily describe the candidate by word. We use a simple rule to judge this. The candidate can be easily described by word if the number of the same or similar color objects is one or two, and if the candidate can be described using one of the following words: left, right, center, front, and back. Fig. 4 shows an example case. In this case, the robot asks the user, “Is the target object to the left side of the center yellow object?"

Figure 3. Situation 1 where no other objects have the same color as the center object.

Figure 4. Situation 1 where multiple objects have the same color as the center object.

Situation 2. Use of a group-based reference system

If the candidate cannot be easily described by word in the process described above, the robot adopts the group-based reference system. Fig. 5 shows an example case. In this case, the robot asks the user “Is the target object in the left side of the cluster?"

Figure 5. Situation 2 where a group-based reference system is used.

Situation 3. Use of size information

Fig. 6 shows an example case. In this situation, one big object covers a most part of the cluster. The positional relations between the big object and others are difficult to be specified. The robot solves this problem by separating the big object and others. The

robot first asks "Is the target object big?" If the answer is no, the robot proceeds to examine if the cluster composed of the remaining objects is in Situation 1 or 2.

Figure 6. Situation 3 where the system asks about size.

3.2. Few-object case

If the number of objects is at most three, the robot can ask about position. The robot asks a 'yes/no-type' question such as "Is the target object in the left side?"

4. Experimental results

We performed experiments for various cases. Here, we show two typical example cases. In the first example shown in Fig. 7, the user wanted the red ball in the scene. The dialog in this case was as follows.

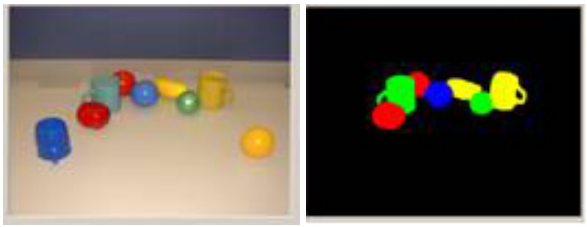
Robot : "*Is the target in the cluster?*"
User : "*Yes*". (The cluster objects remain as in Fig. 7 (b).)
Robot : "*Is the target object to the left side of the blue object?*" (The blue object is the relatum.)
User : "*Yes*".
Robot : "*Is the target on the right side?*"
User : "*Yes*".

The robot understood that the object shown in Fig. 7 (d) was the target object.

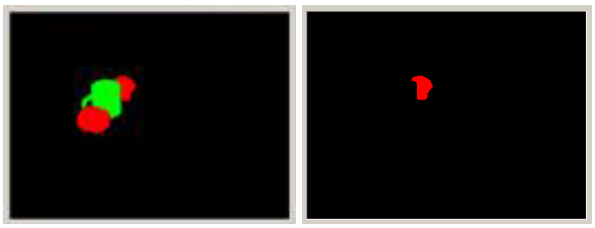
In this case, the robot found a relatum in the scene. Thus, the system asked the user about position using the relatum.

In the second example shown in Fig. 8, the user wanted the green cup. The dialog in this case was as follows.

Robot : "*Is the target in the cluster?*"
User : "*Yes*". (The objects in the cluster remain as in Fig. 8 (b).)
Robot : "*Is the target object big?*"



(a) Input image. (b) After the 1st answer.



(c) After the 2nd. (d) Final result (target object).

Fig. 7. Experimental result 1.

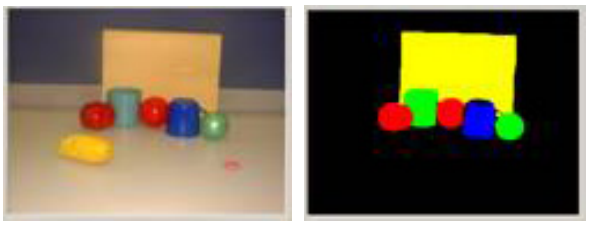
User : "*No*". (The system removes the big object as in Fig.8 (c).)
Robot : "*Is the color of the target object red?*"
User : "*No*" (The system removes the red objects as in Fig. 8 (d).)
Robot : "*Is the target object on the left side?*"
User : "*Yes*".

The robot understood that the object shown in Fig. 8 (e) was the target object.

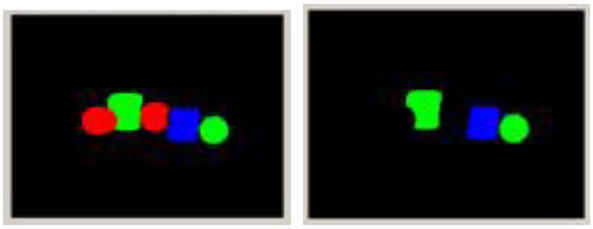
In this case, there was a big object that covered more than a half of the cluster. Thus, the system asks the user about size to separate the big object from the others.

To evaluate the user-friendliness of our system, we compared the system with humans by asking them some questionnaires. We showed seven scenes to fifteen participants. Two of the scenes are shown in Fig. 9. The experimenter chose an object as the target in each scene. We asked the participants to guess what was the target object by asking 'yes/no-type' and 'what-type' questions of the experimenter. Our system also generated a sequence of questions to ask the experimenter.

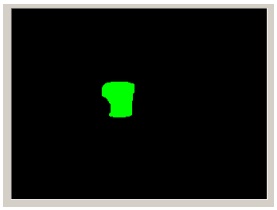
Fig. 10 shows the result. It shows the average, the maximum, and the minimum numbers of questions necessary for the human participants, and the number of questions that the system asked. Our system always needed equal or less number of questions compared with those by the humans. The result proves that the system can generate efficient questions.



(a) Input image. (b) After the 1st answer.



(c) After the 2nd. (d) After the 3rd.



(e) Final result (target object).

Figure 8. Experimental result 2.

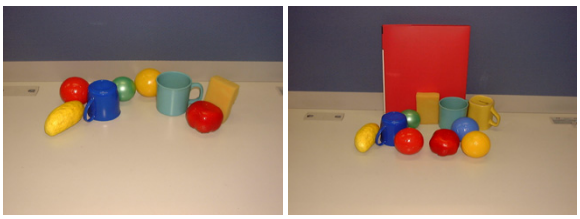


Figure 9. Experimental scenes. Left: Scene 1 Right: Scene 4 in Fig. 10.

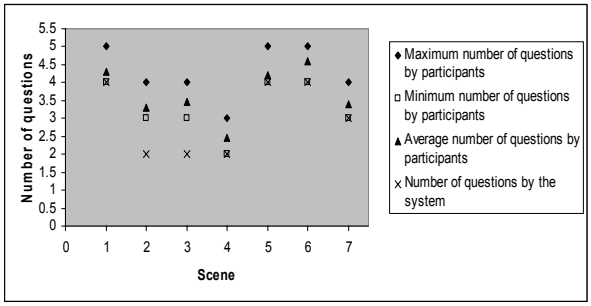


Figure 10. Number of questions necessary for the human participants and the system.

The main limiation of the current system is that it requires the assumption of one-to-one correspondence in image segmentation results. That is, each segmented region should correspond to a different object in the scene. Although this seems to be a strong limitation, there are many cases that satisfy this assumption if we ignore small regions. Still, it is true that there are many cases against this assumption. A simple example is the case with the existence of objects consisting of multiple different color parts. We are working on this problem and have proposed a system that can deal with such cases by examining photometric and geometric properties of region borders [12]. The system, however, has a problem in letting know the user the part currently attended by the system. The reference system method proposed in this paper can be a promising means to solve this problem. We are now integrating these two systems to realize a more capable system.

5. Conclusion

We have presented a system that can detect the target object in occlusion cases through the interaction with the user. It is difficult to obtain some features in occlusion cases. To compensate the system for this shortage of features, we propose to use reference systems to express the positional relationships of objects. Experimental results show that the robot can efficiently detect objects through user-friendly conversation. Although we introduce the reference systems for occlusion cases, the reference systems can be used in various other cases. We are now working on this issue.

Acknowledgements

This work was supported in part by the Ministry of Internal Affairs and Communications under the Strategic Information and Communications R&D Promotion Program, and by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (KAKENHI 14350127).

References

[1] M. Ehrenmann, R. Zollner, O. Rogalla, and R. Dillmann, "Programming Service Tasks in Household Environments by Human Demonstration," in *Proc.*

International Workshop on Robots and Human Interactive Communication, pp.460-467, 2002.

[2] M. Hans, B. Graf, R.D. Schraft, "Robotics Home Assistant Care-O-bot: Past-Present-Future," in *Proc. International Workshop on Robots and Human Interactive Communication*, pp.380-385, 2002.

[3] G. A. Berry, V. Pavlovic, and T. S. Huang, "Battle View: A Multimodal HCI Research Application," in *Proc. Workshop on Perceptual User Interfaces*, pp. 67-70, 1998.

[4] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward Natural Gesture/speech HCI: A Case Study of Weather Narration," in *Proc. Workshop on Perceptual User Interfaces*, pp. 1-6, 1998.

[5] R. Raisamo, "A Multimodal User Interface for Public Information Kiosks," in *Proc. Workshop on Perceptual User Interfaces*, pp. 7-12, 1998.

[6] P. McGuire, J.Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, H. Ritter, "Multi-modal Human Machine Communication for Instruction Robot Grasping Tasks," in *Proc. International Workshop on Robots and Human Interactive Communication*, pp. 1082-1089, 2002.

[7] T. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-robot Interface by Verbal and Nonverbal Communication," in *Proc. International Conference on Intelligent Robots and Systems*, pp.924-929, 1998.

[8] M. Yoshizaki, Y. Kuno, and A.Nakamura, "Mutual Assistance between Speech and Vision for Human-Robot Interface," in *Proc. International Conference on Intelligent Robots and Systems* , pp.1308-1313, 2002.

[9] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Vision-Speech System Adapting to the User and Environment for Service Robots," in *Proc. International Conference on Intelligent Robots and Systems*, pp.1290-1295, 2003.

[10] Z. M. Hanafiah, C. Yamazaki, A. Nakamura and Y. Kuno, "Human-robot Speech Interface Understanding Inexplicit Utterances using Vision," in *Extended Abstracts, Conference on Human Factors in Computing Systems (CHI 2004)*, pp.1321-1324, 2004.

[11] R. Kurnia, M.A. Hossain, A. Nakamura, and Y. Kuno, "Object Recognition through Human-Robot Interaction by Speech," in *Proc. 13th IEEE International Workshop on Robot and Human Interactive Communication*, pp.619-624, 2004.

[12] M.A. Hossain, R. Kurnia, A. Nakamura, and Y. Kuno, "Interactive Vision to Detect Target Objects for Helper Robots," in *Proc. 7th International Conference on Multimodal Interfaces (ICMI)*, pp.293-300, 2005.

[13] 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.

[14] T. Kawaji, K. Okada, M. Inaba, H. Inoue, "Human Robot Interaction through Integrating Visual Auditory Information with Relaxation Method," in *Proc. International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp 323-328, 2003.

[15] P. McGuire, J.Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, H. Ritter, "Multi-modal Human Machine Communication for Instruction Robot Grasping Tasks," in *Proc. International Workshop on Robots and Human Interactive Communication*, pp. 1082-1089, 2002.

[16] K. Komatani, T. Kawahara, R. Ito and H.G. Okuno, "Efficient Dialogue Strategy to Find User's Intended Items from Information Query Results," in *Proc. 19th International Conference on Computational Linguistics*, pp. 481-487, 2002.

[17] Y. Yamakata, T. Kawahara and H.G. Okuno, "Belief Network Based Disambiguation of Object Reference in Spoken Dialogue System for Robot," *in Proc. ISCA Workshop on Multi-modal Dialogue in Mobile Environment*, 2002.

[18] T. Inamura, M. Inaba, and H. Inoue, "Dialogue Control for Task Achievement based on Evaluation of Situational Vagueness and Stochastic Representation of Experiences," in *Proc. International Conference on Intelligent Robots and Systems,* pp. 2861-2866, 2004.

[19] S. C Levinson, "Frames of Reference and Molyneux's Question: Crosslinguistic Evidence," in P. Bloom, M. Peterson, L. Nadel, and M. Garrett, editors, *Language and Space*, pp.109-169, MIT Press, Cambridge, MA., 1996.

[20] T. Tenbrink and R. Moratz, "Group-based Spatial Reference in Linguistic Human-Robot Interaction," in *Proc. European Cognitive Science Conference*, 2003.

[21] R. Moratz, K. Fischer, and T.Tenbrink, "Cognitive Modelling of Spatial Reference for Human-Robot Interaction," *International Journal on Artificial Intelligence Tools, 10:4*, World Scientific Publishing, Singapore, 2001.