Recovery from Segmentation Failures Using Photometric Invariance in an Interactive Object Recognition System *

Md. Altab Hossain, Rahmadi Kurnia, and Yoshinori Kuno
Department of Information and Computer Sciences
Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama-shi, Saitama, Japan.
{hossain, kurnia, kuno}@cv.ics.saitama-u.ac.jp

Akio Nakamura
Department of Machinery System Engineering
Tokyo Denki University
2-2, Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, Japan.
E-mail: nkmr-a@cck.dendai.ac.jp

***Abstract* – We are developing a helper robot that carries out tasks ordered by the user 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. Thus, we have proposed to use the human user's assistance through speech. 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. Our previous system assumes that it can segment images without failure. However, if there are occluded objects and/or objects composed of multicolor parts, segmentation failures cannot be avoided. This paper presents an extended system that tries to recover from segmentation failures using photometric invariance. If the system is not sure about segmentation results, the system asks the user by appropriate expressions depending on the invariant values.**

Index Terms – Image Processing, Image Segmentation, Machine Vision, Object Recognition.

I. INTRODUCTION

Service robotics is an area in which technological progress leads to rapid development and continuous innovation. Developing robot companions that support a natural interaction with a human user is a challenging research topic. The basic idea of a robot companion is that it is used as a personal robot which a user shares his private home with. Thus, the interaction interface has to match all requirements for an easy usability, so that even naive users are able to interact with the robot without an extensive training phase and engineering knowledge.

Recently, helper robots or service robots in welfare domain have attracted much attention of researchers for the coming aged society [1][2]. Such robots need user-friendly human-robot interfaces. Multimodal interfaces [3][4][5] are considered strong candidates. Thus, we have been developing a helper robot that carries out tasks ordered by the user through voice and/or gestures [6][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. Thus, we have proposed to use the human user's assistance through speech [6][7][8][9]. 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 the initial stage of research [6][7][8], we assumed that the scene was 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 tackled this problem in [9]. The robot determines the target through a conversation with the user. We have presented a method of generating 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.

In our previous work, however, we still simplified the problem. We assumed that we could obtain perfect image-segmentation results. Each segmented region in images corresponds to an object in the scene. However, we cannot always expect this one-to-one correspondence in the real world. Segmentation failures are inevitable even by a state-of-the-art method. In this paper, we address this problem. Although segmentation fails due to various reasons, we consider two most typical cases here: occlusion and multi-color objects. If a part of an object is occluded by another object, these two objects might be merged into one region in an image. If an object is composed of multiple color parts, each part might be segmented as a separate region. We propose to solve this problem by combining a vision process with photometric invariance and interaction with the user.

There has been a great deal of research on robot systems understanding the scene or their tasks through interaction with the user [10][11][12][13][14][15][16]. These conventional systems mainly consider dialog generation at the language level. In this research, however, we concentrate on computer vision issues in generating dialogs where the

* This work is supported in part by the Ministry of Education, Culture, Sports, Science and Technology under the Grant-in-Aid for Scientific Research (KAKENHI 14350127).

scene is relatively complex. The scene may include multicolor and / or partially occluded objects.

II. PROBLEMS OF SEGMENTATION

The basic framework of the proposed system is the same as that of our previous system [9]. The system first carries out image segmentation. We have developed an object segmentation method based on the mean shift algorithm and HSI (Hue, Saturation, and Intensity) color space. Although the mean shift algorithm and the HSI color space have been separately used for color image segmentation, conventional methods using one of them fail to segment an image when the illumination condition will change. To solve this problem, we use the mean shift algorithm as an image pre-processing tool. This reduces the number of colors in the image and divides it into several regions.

Once the process using the mean shift algorithm is completed, the merging process of adjacent regions begins. The objective of this step is to find regions that can reasonably be assumed to belong to a single object. We use the Hue, Saturation and Intensity components of the HSI color space to merge the homogeneous regions which likely come from a single object. For homogeneous regions, we use threshold values for each component of HSI. We use the histograms of each component to select the appropriate threshold. The threshold values are selected dynamically based on the illumination condition of the image and thereby efficiently segment out specific color regions in different illumination conditions. Fig. 1 shows an example of image segmentation.

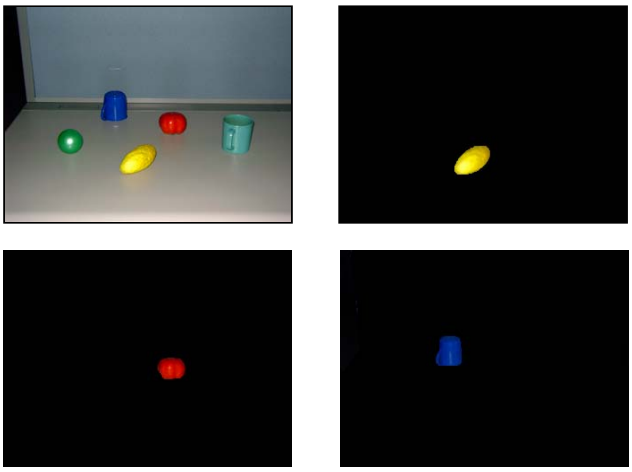

Figure 1. Original single color objects (upper left). Recognized target objects: a yellow one (upper right), a red one (lower left), and a blue one (lower right).

We can extract certain color objects by specifying their color. In [9], we have proposed a system that asks the user about the color, shape, size and position of the target object to identify it among the segmented objects (regions). The system determines what attribute it will ask depending on the segmentation result and the characteristics of image features. It also changes how to ask questions depending on the situation so that the user can easily answer the questions and the system can effectively identify the target.

The system can work as long as the segmentation results satisfy one-to-one correspondence, that is, each region in the image corresponds to a different object in the scene. However, we cannot always expect this in complex situations. Two most typical cases that break this assumption are occlusion and multi-color object situations. If an object is composed of multiple color parts, each part might be segmented as a separate region. Fig. 2 shows an example. The bottle is divided into two segments. If a part of an object is occluded by another object, these two objects might be merged into one region in an image. Fig. 3 shows an example. In this paper, we solve this problem by introducing photometric invariance in the interaction framework.

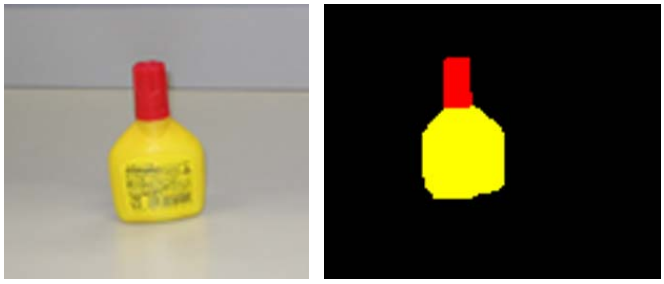

Figure 2. Multi-color object case. Left: Original image; Right: Segmentation result.

Figure 3. Occlusion case. Left: Original image; Right: Segmentation result.

III. REFLECTANCE RATIO TO MEASURE THE COMPARTIBILITY OF ADJACENT REGIONS

The reflectance ratio, a photometric invariant, represents a physical property that is invariant to illumination and imaging parameters. Nayar and Bolle [17] presented that reflectance ratio can be computed from the intensity values of nearby pixels to test shape compatibility at the border of adjacent regions. The principle underlying the reflectance ratio is that two nearby points in an image are likely to be nearby points in the scene. Consider two adjacent color regions r_1 and r_2. If r_1 and r_2 are parts of the same piece-wise uniform object and have a different color, then the discontinuity at the border must be due to a change in albedo, and this change must be constant along the border between the two regions. Furthermore, along the border, the two regions must share

similar shape and illumination. If r_1 and r_2 belong to different objects, then the shape and illumination do not have to be the same.

If the shape and illumination of two pixels p_1 and p_2 are similar, then the reflectance ratio, defined in Eq. (1), where I_1 and I_2 are the intensity values of pixels p_1 and p_2, reflects the change in albedo between the two pixels [17].

$$R = \left(\frac{I_1 - I_2}{I_1 + I_2} \right) \qquad (1)$$

For each border pixel p_{1i} in r_1 that borders on r_2, we find the nearest pixel p_{2i} in r_2. If the regions belong to the same object, the reflectance ratio should be the same for all pixel pairs (p_{1i}, p_{2i}) along the r_1 and r_2 border.

We use this reflectance ratio to determine whether or not geometrically adjacent regions in an image come from a single object. If the adjacent regions come from a single object, the variance of reflectance ratio should be small. Otherwise, large.

If there are n reflection ratios x_1, x_2,…., x_n in the border regions, the sample mean and variance are defined by

$$Mean = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (2)$$

$$Variance = \frac{1}{n} \sum_{i=1}^{n} (x_i - Mean)^2 \qquad (3)$$

In addition, we examine the reflectance ratio for isolated regions if their boundaries have discontinuous parts. If the ratio varies much along the line connecting the discontinuous points, multiple objects might form the region due to occlusion.

IV. INTERACTIVE OBJECT RECOGNITION

The system applies the initial segmentation method described in Section II to the input image to find uniform color regions in the image.

Then, the system examines one-to-one correspondence between a region and an object. A simple measure for this check is the variance of the reflectance ratio. If r_1 and r_2 are part of the same object, this variance should be small (some small changes must be tolerated due to noise in the image and small-scale texture in the scene). However, if r_1 and r_2 are not parts of the same object, the illumination and shape are not guaranteed to be similar for each pixel pair, violating the specified conditions for the characteristic. Differing shape and illumination should result in a larger variance in the reflectance ratio.

We performed experiments to examine the usefulness of this measure. We measured the variance of reflectance ratio from 80 test images that are taken in different illumination conditions. The images consist of 40 multicolor object cases and 40 occluded object cases. Fig. 4 shows the result.

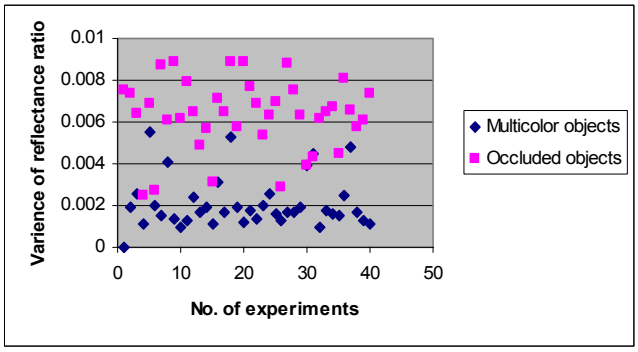

Figure 4. Distribution of variances of reflectance ratio for multicolor and occluded objects.

From this experimental result, we classify situations into the following three cases depending on the variance values of the reflectance ratio.

Case 1: If the value is from 0.0 to 0.0020, we confirm that the regions are from the same object.

Case 2: If the value is from 0.0021 to 0.0060, we consider the case as the confusion state.

Case 3: If the value is greater than 0.0060, we confirm that the regions are from different objects.

In cases 1 and 3, the system proceeds to the next step without any interaction with the user. In case 1, the system considers that the regions are from the same object, while in case 3, they are from different objects. In case 2, however, the system cannot be sure whether the regions are from the same objects or different objects. The system follows our basic framework in this situation. It asks questions of the user.

For simple and user friendly interaction with the user, we divide case 2 further into three categories. Different questions will be asked to the user, based on the value of the reflectance ratio.

Category A: If the value is from 0.0021 to 0.0030, the robot will ask, “Are those regions parts of the same object?” (Yes/No)

Category B: If the value is from 0.0031-0.0040, the robot will ask, “Are those regions parts of the same object or different objects?” (Same/Different)

Category C: If the value is from 0.0041-0.0060, the robot will ask, “Are those regions parts of a different object?” (Yes/No)

We assume that it is easy and convenient for the user to say ‘Yes’, because the reply ‘No’ sometimes may require some extra information to explain the justification of his/her answer.

V. EXPERIMENTS

We performed several experiments to examine the effectiveness of our approach. As mentioned in the introduction, we are developing a robot to get objects ordered by handicapped people. Main target objects are cups, cans, bottles, fruits, books, etc., on tables or shelves. We set up experimental scenes by considering this application.

A. Example Experiment Cases

We performed experiments for various cases in different illumination conditions. Here, we show four typical example cases.

Experiment 1: Multicolor object case

After the initial segmentation and merging regions based on the mean shift and HSI, two regions, yellow and red, are found (Fig. 5). The reflectance ratio in the region's boundary is 0.0011. Since the value falls in case 1, the system concludes that these two regions are parts of the same object.

Robot: Is there an object made of two colors yellow and red?
User: yes.

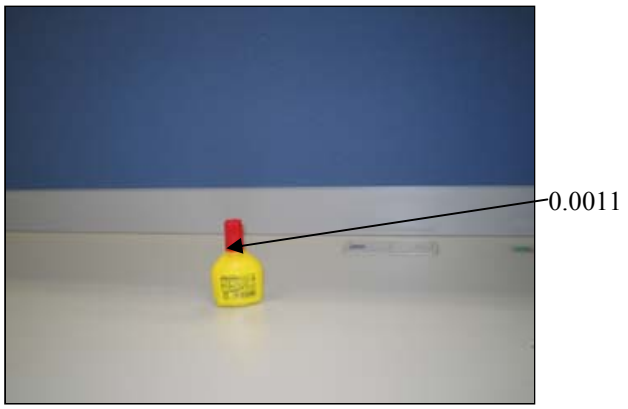

Figure 5. Multicolor object case.

Experiment 2: Occluded object case (1)

In the scene shown in Fig. 6, the segmentation process gives only a region. The system checks the reflectance ratio along the line segment connecting the points where the boundary is not continuous. Since the variance of reflectance ratio is 0.0061, the system judges that the region should be divided into two as shown in Fig.7.

Robot: Are there two partially occluded yellow color objects?
User: yes.

Figure 6. Occlusion case where only a region is detected in the segmentation result.

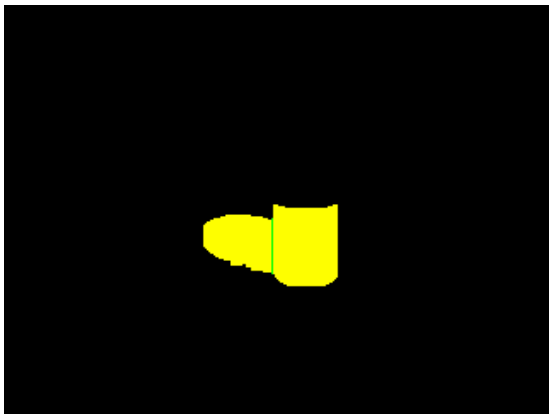

Figure 7. Final segmentation result.

Experiment 3: Occluded object case (2)

After initial segmentation, two regions, yellow and red similar to experiment 1, are found (Fig. 8). The variance of the reflectance ratio on the region boundary in this case is 0.0045. Since the situation is case 2, the robot needs the user's assistance. As the value falls in the range from 0.0041 to 0.0060, the robot asks the following question.

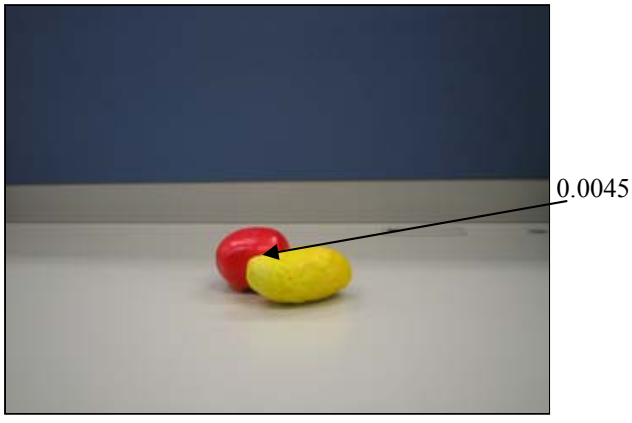

Figure 8. Occlusion case where two adjacent different color regions are detected in the segmentation result.

Robot: "Are there two different color objects partially occluded by the other?"
User: Yes.

Based on the user response, the robot confirms that the two regions are parts of different objects.

Experiment 4: Complex case

In the scene shown in Fig. 9, there exist three objects: two single color objects and one multicolor object. Two objects are partially occluded by the third object. After applying the initial segmentation technique, the robot obtained four connected regions, R_1, R_2, R_3 and R_4. To confirm which regions are parts of the single or different objects, the robot examines the value of the reflectance ratio of the adjacent regions.

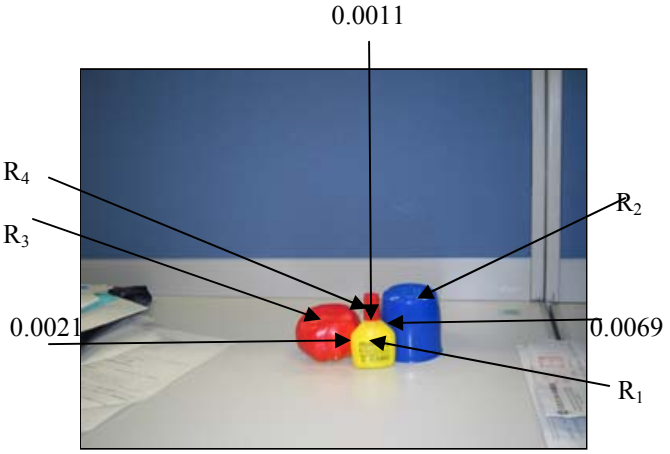

Figure 9. Image containing single color, multicolor and occluded objects.

Fig. 9 shows four regions R_1, R_2, R_3, R_4 and the variances of reflectance ratios for the different adjacent region boundaries. According to the value of the reflectance, the robot concludes that regions R_1 and R_2 are parts of different objects, because the value of the variance is greater 0.0060 (case 3). Regions R_1 and R_4 are parts of the same object, because the value of the variance is less 0.0020 (case 1). However, the robot is not sure about the regions R_1 and R_3, because the value of the variance is in the range of case 2. The robot needs the user's assistance. As the value is in the category A in case 2, the robot interacts with the user in the following way,

Robot: "Are those regions parts of the same object?"
User: No.

Then, the robot comes up to know that regions R_1 and R_3 are parts of different objects. Finally, the robot understands that there are three objects; one is a multicolor object composed of regions R_1 (yellow) and R_4 (red), and the other two regions R_2 (blue) and R_3 (red) are single color objects.

In complex cases like the above, the user may not know which part the robot is talking about. The robot should make this clear to the user. In the above experimental case, the user cannot understand what 'those regions' mean only from the robot's utterance. The system shows the regions of interest on the display screen to the user in the current implementation. We would like the robot to do this by speech and gesture as humans do. For example, the robot will point at the regions by its finger when they speak. And/or the robot will give more information by speech, such as saying, "I am talking about the objects besides the blue one," in the above case. The user now knows that the robot is talking about the red and yellow objects. These are left for future work.

B. Comparison Experiments

We performed experiments to examine how much the proposed system could reduce the user's burden. We modified our previous system [9] to compare with the current system. Our previous system assumed one-to-one correspondence. We have added a module to correct the segmentation result to satisfy the one-to-one correspondence through interaction with the user. The robot system tells the current segmentation result to the user and asks if this is correct. If the user's answer is negative, the robot asks the number of objects in the scene. If necessary, it asks which regions come from the same object or which region should be divided into multiple objects. Actually, this module has been developed for the current system so that the system can identify target objects when it cannot make decisions. We counted the number of questions necessary to identify target objects for this modified previous system and the proposed system.

For example, the modified previous system worked as follows in the case of experiment 1 (Fig. 5).

Robot: Are there two single color objects?
User: No.
Robot: How many objects in the scene?
User: One.

The numbers of required questions are two. However, using our method, the robot does not need any human assistance to know the number of objects in the scene. It needs only a question asking for confirmation.

Table 1 shows the results for the experimental cases 1-4 described in Section V (A). It also shows the results for other six cases. The results confirm that our current system needs a smaller number of questions than the previous system.

We show that the decision based on the reflectance ratio is useful. However, there are cases that the system cannot determine where to check the ratio. For example, suppose that there is a small object in front of a large object and their colors are the same. If there are no discontinuous points on

the boundary, the system misjudges these objects as one object. In this case, the system can tell through interaction with the user that there are two objects. However, we need to improve image processing capability to detect these two objects separately such as by examining edges or slight color changes.

Table 1: Comparison experiment.

Experiment No.	Number of objects(regions)	Number of required questions	
		Our method	Our previous system
1	1(2)	1	2
2	2(1)	1	2
3	2(2)	1	1
4	3(4)	1	5
Other Experiments			
5	5(7)	3	9
6	1(3)	1	4
7	4(6)	2	7
9	3(5)	1	5
9	2(2)	3	4
10	2(4)	1	6

VI. CONCLUSION

The service robot that carries out tasks ordered by the user through speech needs a vision system to recognize the objects appearing in the orders. The target objects can be single or multicolor, and in real scenes, some objects may be occluded by others. The system should have a capability of dealing with all possible complexities of single color, multicolor and occluded objects. This paper proposes to use photometric invariance to reduce segmentation failure cases. Our proposed method using a photometric invariant with the help of the interaction with the user can efficiently and accurately identify single color, multicolor and occluded objects in different illumination conditions. Experimental results show the usefulness of the proposed method. Although the system cannot recover from all segmentation failures, this kind of improvement can make the system more acceptable.

REFERENCES

[1] M. Ehrenmann, R. Zollner, O. Rogalla, and R. Dillmann, "Programming service tasks in household environments by human demonstration," ROMAN 2002, pp.460-467.

[2] M. Hans, B. Graf, R.D. Schraft, "Robotics home assistant Care-O-bot: Past-present-future," ROMAN 2002, pp.380-385.

[3] G. A. Berry, V. Pavlovic, and T. S. Huang, "BattleView: A multimodal HCI research application," 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," Workshop on Perceptual User Interfaces, pp. 1-6, 1998.

[5] R. Raisamo. "A multimodal user interface for public information kiosks," Workshop on Perceptual User Interfaces, pp. 7-12, 1998.

[6] T. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-robot interface by verbal and nonverbal communication," IROS 1998, pp.924-929, 1998.

[7] M. Yoshizaki, Y. Kuno, and A.Nakamura, "Mutual assistance between speech and vision for human-robot interface," IROS 2002, pp.1308-1313, 2002.

[8] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Vision-speech system adapting to the user and environment for service robots," IROS2003, CD-ROM, 2003.

[9] Rahmadi Kurnia, Md. Altab Hossain, Akio Nakamura, and Yoshinori Kuno, "Object Recognition through Human-Robot Interaction by Speech," Proc. of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004), pp.619-624, Japan, September 20-22, 2004.

[10] M. Takizawa, Y. Makihara, N. Shimada, J. Miura, and Y. Shirai, "A Service Robot with Interactive Vision- Objects Recognition using Dialog with User," Proc. First International Workshop on Language Understanding and Agents for Real World Interaction, Hokkaido, (2003).

[11] T. Kawaji, K. Okada, M. Inaba, H. Inoue, "Human Robot Interaction through Integrating Visual Auditory Information with Relaxation Method," Proc. International Conference on Multisensor Fusion on Integration for Inteligent Systems, Tokyo, pp 323 – 328 (2003).

[12] 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," Proc. International Workshop on Robots and Human Interactive Communication, Berlin, pp. 1082-1089 (2002)

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

[14] Anita Cremers, "Object Reference in Task-Oriented Keyboard Dialogues, Multomodal Human-Computer Communication: System, techniques and experiments," Springer, pp. 279-293, (1998).

[15] T. Winograd, "Understanding Natural Language," New York: Academic Press (1972).

[16] D. Roy, B. Schiele, and A. Pentland, "Learning Audio-visual Associations using Mutual Information," Proc. International Conference on Computer Vision, Workshop on Integrating Speech and Image Understanding, Greece, (1999).

[17] S.K. Nayar and R.M. Bolle, "Reflectance based object recognition," Inter. Journal of Computer Vision, vol. 17, no. 3, pp. 219-240, 1996.