2005－CVIM－150（10）
2005／9／5

光学的特徴の解析に基づく対話物体認識

Md. Altab Hossain, Rahmadi Kurnia, 久野 義徳

埼玉大学

高齢化社会を向かえ，福祉関連などのサービスロボットのニーズが高まってくると考えられる．このようなロボットでは，例えば頼まれたものを取ってくるというような作業を行う場合などにおいて物体認識が必要である．しかし，一般的な環境で確実に物体認識を行うことは難しい．そこで，そのような場合にユーザとの対話を通じて認識を行う方法を検討している．これまでは，物体のセグメンテーションは正しいとして，どれが対象物体かという問題を扱ってきたが，今回は，セグメンテーションの結果にも誤りの可能性があるとして，それを対話で修正する方法を提案する．オクルージョンや複数色からなる物体が存在する場合に，光学的特徴に応じて適切な質問をユーザにすることで，対象物体を認識する方法を示す．

Interactive Object Recognition Using Photometric Properties

Md. Altab Hossain, Rahmadi Kurnia, and Yoshinori Kuno
Department of Information and Computer Sciences, Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama-shi, Saitama 338-8570, Japan.
{hossain, kurnia, kuno}@cv.ics.saitama-u.ac.jp

Abstract An effective human-robot interaction is essential for wide penetration of service robots into the market. Such robot needs a vision system to recognize objects. It is, however, difficult to realize vision systems that can work in various conditions. More robust techniques of object recognition and image segmentation are essential. Thus, we have proposed to use the human user's assistance through speech. This paper presents a system that can recognize objects in occlusion and/or multicolor cases using geometric and photometric analysis of images. If the robot is not sure about segmentation results, it asks questions to the user by appropriate expressions. Through experiments on a real mobile robot, we have confirmed the usefulness of the system.

1. Introduction

Service robotics is an area in which technological progress leads to rapid development and continuous innovation. Many different research disciplines are involved in service robots, e.g. sensor design, control theory, manufacturing science, artificial intelligence, and also computer vision and speech understanding with dialogue. The latter two are especially important since service robots should serve as personal assistants. As a consequence, service robots differ from other mobile robotic systems mainly by their intensive interaction with people in natural environments. In typical environments for service robots, like hospitals or day care facilities for elderly people, the demands on the interface between robots and humans exceed the capabilities of standard robotic sensors, like sonar, laser, and infrared sensors. Thus, in many cases, computer vision as well as natural language dialogue components become essential parts of such a system.

Recently, helper robots or service robots which interact with humans in welfare domain have attracted much attention of researchers [1][2]. Multimodal interfaces [3][4][5] are considered strong candidates for user-friendly human-robot interactions. 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 question to the user so that the natural response by the user can give helpful information for its vision system.

In our previous work, however, 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 geometric and photometric analysis of images 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. Moreover, all of them consider relatively simpler scene containing single color objects without occlusion. In this research, however, we concentrate on computer vision issues in generating dialogs where the scene is complex. The scene may include multicolor or occluded objects.

2. Basic Framework

This section briefly describes our previous system since the basic framework is common to our system proposed in this paper.

We represent objects by their attributes such as color and shape. The vision system tries to detect regions with the attributes of the target object. For example, assuming that 'apple' is represented as a red round object. If the user asks the robot to get the apple, the robot initiates color segmentation and shape detection processes. If it can find a red round object, it asks the user for confirmation through speech. Otherwise, it explains the current vision results through speech, expecting that the user's reply may help to recognize the object.

In [6], we consider the cases where the robot has a priori knowledge about target objects and the failure of vision comes from the difference between the current object attributes and the stored knowledge. For example, in the apple's case mentioned above, the robot cannot detect an apple if the apple in the scene is a green apple. In this case, the robot tells the user that it cannot find a red object but has detected a green round object. From this, the user knows that the robot does not know the existence of green apples. We can expect him/her to say something about green color to the robot. In [8], we propose an object recognition method that learns appropriate vision processes depending on the environment through its use with interaction with the user. We also assume that the robot knows a priori knowledge about target objects.

In [9], we dealt with objects with no a priori knowledge. The user may say object names that the robot does not know what they are, or he/she just mentions them using deictic words such as 'that' [17]. We would like to enable the robot to work in such situations. In addition, more importantly, we consider actual complex situations where it is difficult to choose a target object among many objects. In our previous work, we inexplicitly assumed that the scene was simple so that the vision system detected one or at most a few regions (objects) in the image. Thus, even though detecting the target object may be difficult, once something has been detected, the system can assume it as the target. However, in actual complex scenes, the vision system may detect various objects, especially if it does not have a priori knowledge about the object.

As mentioned earlier, we represent an object as a set of attributes and recognize it by finding a region with the attributes. Thus, if the user gives the robot the information about some attributes of the target object, it can remove the objects that do not satisfy the attributes, reducing the number of candidates for the target object. In other words, the robot can identify the target object by asking the user for the attributes of the target object. However, if it asks him/her all the information at once, he/she may find it difficult to answer. It is easy for humans to answer to short simple questions. On the other hand, it is not good for users if the robot needs too many questions, even if each is simple, to identify the target object. The point is, therefore, how to generate a sequence of utterances leading to identify the target objects efficiently and user-friendly. We have tackled this problem in [9].

What question that the robot should ask depends on the current vision results and the characteristics of attributes. If all the detected regions are different in a particular attribute, asking the attribute may help much to determine the target. For example, if all the regions in the initial segmentation result are different in color, it may be appropriate to ask, "What color is it?" However, even if all the objects are different in shape, if they are of irregular shape, it is not good to ask, "What shape is it?" The user finds it difficult to answer to such a question by speech. We need to consider such characteristics of features in generating utterances. We consider the characteristics of features to determine which feature the robot uses and how to use it from four viewpoints: vocabulary, distribution, uniqueness, and relativity. We make a binary decision from each viewpoint for each feature. We use four features: color, size, position, and shape. Table 1 summarizes the characteristics of the features.

Table 1: Features and their characteristics.

Characteristic	Color	Size	Position	Shape
Vocabulary	√	-	√	-
Distribution	-	-	√	-
Uniqueness	-	-	√	-
Absoluteness	√	Relative	Relative	√

Humans can easily describe some features by word but cannot do so for other features. If we can represent a particular feature easily by word for any given object, we call it a vocabulary-rich feature. The

robot can ask relatively complex questions such as 'what-type' questions since we can easily find an appropriate word for answer. For example, the robot can ask, "What is the color of the target object?" since color is a vocabulary-rich feature. If we can describe a particular feature by word even if only an object exists, we call it an absolute feature. Otherwise, we call it a relative feature. Color and shape are absolute features in general. Size and position are relative features. The robot prefers to use absolute features when the number of objects is large. If the feature is not a vocabulary-rich feature, the robot uses multiple-choice questions or yes-no type questions. For more details including 'distribution' and 'uniqueness', see [9].

However, our previous 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. If a part of an object is occluded by another object, it is not clear that the regions are from the same object or from different objects. Segmentation failure means failure of object recognition, because recognition is carried out based on the segmentation result. In this paper, we solve this problem by using geometric and photometric analysis of the image in the interaction framework.

3. Reflectance Ratio for Photometric Analysis

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 colored 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].

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

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. 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.

4. Intensity Profile for Geometric Shape Continuity Analysis

So far, we have discussed the shape compatibility between two adjacent regions using a measure based on the intensity values of border pixels. Now, we concentrate on the compatibility of the shape of adjacent regions by analyzing the intensity values within the two adjacent regions. Actually, if two regions are part of the same object, then the surface form of two regions must have a continuous profile. Thus we should represent the surface profile of two regions and compare them in the matter of compatibility or non-compatibility of their form. The intensity value of pixels within the regions gives a good indication of the form of region surfaces. As a matter of fact, we have allowed all variations of pixel intensity values within the regions by solely using the chromatic components of HSI space to perform the segmentation. These variations (or intensity profiles) represent the shape of regions in the image. In general, the intensity profile of regions in the image form 3D patches and their analysis and modeling which are a challenging task are out of the subject of this work.

Rather than observing the intensity profile in 3D case, we abstract the problem to a simpler domain by analyzing it along the horizontal or vertical line crossing through both regions. In other words, we convert the pixels to a line profile that records the pixel intensity as a function of position. To obtain the line profile for region pair, we take into account pixel pair (p_{1i}, p_{2i}) along the middle of the adjacent region borders r_1 and r_2. We then fit a line passing these points and crossing both regions.

For a complex scene containing non-uniform 3D objects, intensity profiles may have any degree of complexity and their modeling is an elaborate task to do. However, for piece-wise uniform objects, we should be able to effectively represent the intensity

profiles by simple models. In this work, we present an approximate parametric approach for modeling the intensity profiles which are either straight-line segments or circular arcs. Our goal is to differentiate between these two cases and we are not searching for a precise modeling of each case (which is needed to consider highly order polynomials). We summarize this parametric modeling in details as;

(i) Straight line, $y = c$

(ii) Line with slope $y = bx + c$

(iii) Curve $y = ax^2 + bx + c$

Where c is the constant term, b is the linear term, and a is the quadratic term.

We computer the parameters as:

$$\text{c=mean(y)}, \begin{bmatrix} b \\ c \end{bmatrix} = \frac{X1}{Y}, \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \frac{X2}{Y} \qquad (2)$$

$$X1 = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ . & . \\ . & . \\ x_n & 1 \end{bmatrix}, \quad X2 = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ . & . & . \\ . & . & . \\ x_n^2 & x_n & 1 \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} \qquad (3)$$

Where, x_i is the position of pixel i related to the region border in the line profile, and y_i is the intensity value of pixel i in the line profile.

First, we calculate the parameters of each model for a given line profile and we then determine which model is a better match to the line profile using the minimum of mean absolute error between each model and the line profile. It is important to note that clearly a curved line can be arbitrarily close to a straight line and thus this distinction must be made by using a selected threshold. After our experiments, we found that a general threshold Tstraight/curved=0.015 is appropriate for this distinction if we use the mean absolute error as a measure of distinction. Furthermore, we do not process lines that are too small or long since they cannot be reliably modeled. Once the best model matches are determined for line profiles, we can examine the compatibility of models.

One drawback of this test tool could be that it cannot, in general, be used on small regions of an image because it violates basic assumptions necessary for the tool to function properly. However, in this situation, the problem can be solved with the user interaction using the reflectance ratio (Experiment 4).

5. Interactive Object Recognition

The system first carries out image segmentation. We use a robust approach of feature space method: the mean shift algorithm combined with HSI (Hue, Saturation, and Intensity) color space for color image segmentation [18].

Once the process of color segmentation 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. 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 parts 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. 1 shows the result.

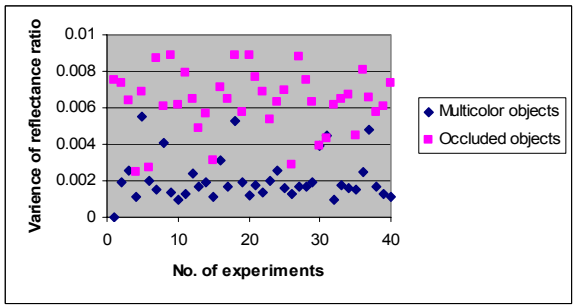

Fig. 1. Distribution of variances of reflectance ratio for multicolor and occluded objects.

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 objects.

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 object or different objects. The system must further investigate the image in such complex situation. We use intensity profile in addition to the reflectance ratio in this case. We use the quadratic regression to the intensity values along horizontal line for a straight lines or curves. Then, we check the continuity of the straight lines or curves of adjacent regions for their compatibility.

6. Dialog Generation to Determine Multicolor or Occlusion

It is helpful for the robot to segment and recognize the target object from the scene if it knows the number of objects in the scene. Thus the robot asks to know the number of objects or to confirm its investigation result regarding the number of objects in the scene. This kind of question is plausible if the number of multicolor or occluded objects is not more than six in the scene by considering user's easiness. However, there may exit more than six objects in the scene. In such case, the robot asks the user to divide the scene into manageable size by using some objects as a reference. For example, the robot may ask, "Is the target object on the right of the blue object?" The robot can also simplify the scene by moving some objects by its hand. (This is not yet implemented.)

The robot's assumption based on reflectance ratio and intensity profile can be any of the four.

6.1 Robot's assumption is correct

In this case, the robot asks the user for confirmation. The question will be in the form "Are there A objects (B single color and C multicolor)?" where A, B, and C are numbers less than six and A= (B+C). The user's replies must be 'Yes'.

6.2 Robot's assumption is wrong

The first question will be the same as before. However, the user's reply will be 'No' in this case. The robot tries to know the number of objects in the scene. So, the next question asked by the robot is "How many objects?" The user replies the exact number of objects. As the robot's initial assumption is wrong and it now knows the number of objects, it should reinvestigate to adjust the result. Thus, all regions pairs that have lower reflectance ratios must undergo for further analysis because the regions with higher values are definitely from different objects.

6.3 Robot's assumption about the total number is correct but segmentation is not correct

The first question will be the same as 6.1 based on the scene analysis. However, the user's reply will be 'No'. So, the next question asked by the robot is "How many objects?" The user replies the same number of objects. As the numbers of objects are same, the robot comes up to know about segmentation failure. It will ask about the confused regions like "I am not sure about the red and yellow parts in front of the blue object." It will ask by referring a confirmed object near the regions. If it can find out the border of segmentation failure, it can ask about combination like:

Robot: Choose the combination:
A: One yellow and another multicolor.
B: One red and another multicolor.
C: Both multicolor.
User: B.

6. 4 Robot cannot conclude

In this case, the robot will ask about the confused regions like "I am not sure about the red and yellow parts in front of the blue object". It will ask by referring a confirmed object near the regions.

7. Experiments

We performed 80 experiments for various cases in different illumination conditions. Here, we show four typical example cases. They represent four different problems found in our experiments. We use Pioneer 2 by ActivMEDIA as a robot (Fig. 2) in our experimental purposes.

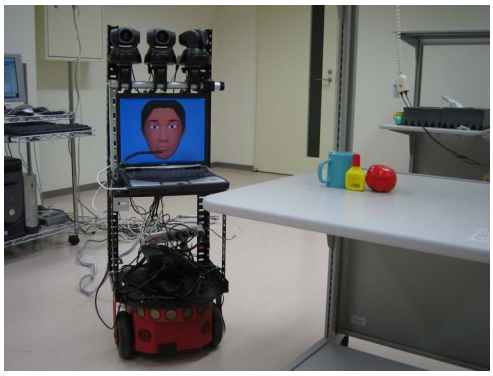

Fig. 2. Robot used in the experimental purposes.

Experiment 1: Robot's observation is the same as the user's

After applying the initial segmentation technique, the robot obtained four connected regions, R_1, R_2, R_3 and R_4. Fig. 3 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 than 0.0060 (case 3). Regions R_1 and R_4 are parts of the same object, because the value of the variance is less than 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.

From the value of the reflectance ratio, the robot assumes that these two regions may be parts of a

single object. However, after investigating the intensity profile along the horizontal line, it will sure that the regions are parts of different objects. So, the robot asks its user for confirmation.

Robot: Are there three objects (two single colors and one multicolor)?
User: Yes.

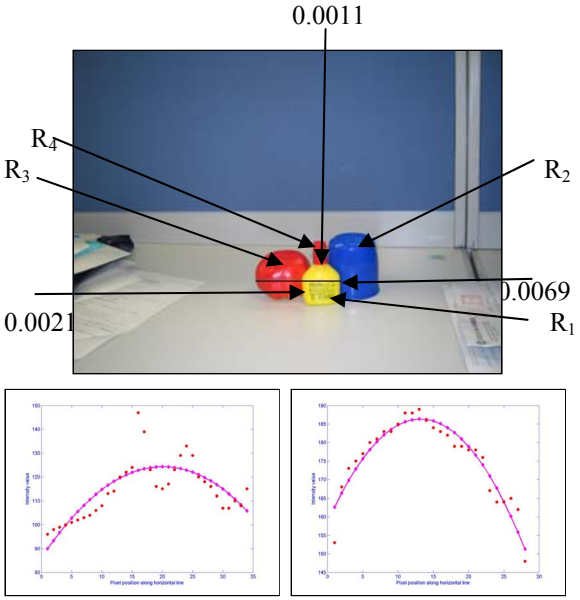

Fig. 3. Image containing single color, multicolor and occluded objects (top). Intensity profile of region R_3 and R_1 (bottom-left and bottom-right).

Experiment 2: Robot's observation is different from the user's

Fig. 4 shows six regions R_1, R_2, R_3, R_4, R_5, R_6 and the variances of reflectance ratios for the different adjacent region boundaries. According to the value of the reflectance, the robot concludes that region pairs (R_1, R_6), (R_2, R_6), (R_4, R_6) and (R_4, R_5) are parts of different objects, because the values of the variance are greater than 0.0060 (case 3). Regions R_1 and R2 are parts of the same object, because the value of the variance is less than 0.0020 (case 1). However, the robot is not sure about the region pairs (R_2, R_3), and (R_1, R_4) because the values of the variance are in the range of case 2.

However, after investigating the intensity profile along the horizontal line for the region pairs (R_2, R_3), and (R_1, R_4), the robot is sure that the regions are parts of different objects. So, the robot asks its user for confirmation.

Robot: Are there five objects (four single colors and one multicolor)?
User: No.
Robot: How many object in the scene?
User: Six.

As the robot's initial assumption is wrong, the robot should reinvestigate to adjust the result using reflectance ratio. Thus, all regions pairs that have lower values of reflectance ratio must undergo for further analysis. In this case, after investigating the intensity profile along the horizontal line for the region pair (R_1, R_2), it will sure that the regions are parts of different objects.

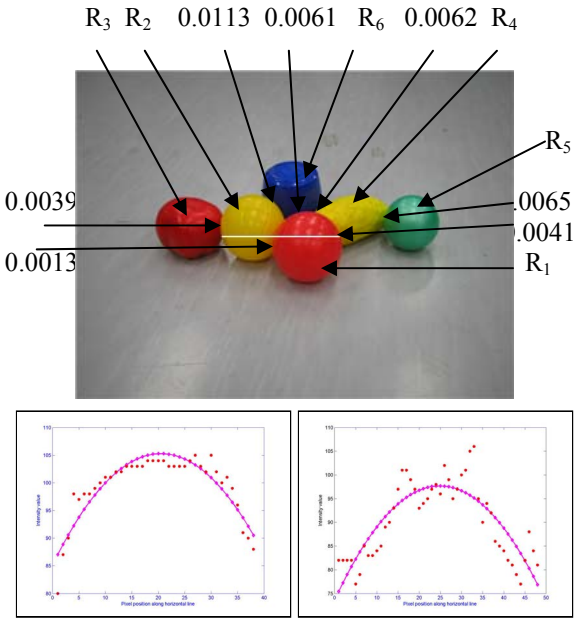

Fig. 4. Image containing single color occluded objects (top). Intensity profile of region R_2 and R_1 (bottom-left and bottom-right).

Experiment 3: Robot's observation is the same as the user's, but segmentation is not perfect

In the occluded object case shown in Fig. 5, two regions, yellow and red are found after initial segmentation. The variance of the reflectance ratio in the region boundary in this case is 0.0052. Since the situation is case 2, the robot needs further image investigation and also the user's assistance. From the analysis result of the intensity profile along the horizontal line path, the robot will come up into an assumption that there are two single color objects. However, to make sure it's assumption, the robot asks:

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

As the robots initial assumption is wrong, it will ask for the user's assistance by considering all possible combinations. Since the robot generates dialogue based on one-to-one correspondence between regions, it makes the following dialogue to disambiguate the scene.

Robot: Choose the combination:

A: One yellow and another multicolor.

B: One red and another multicolor.

C: Both multicolor.

User: B

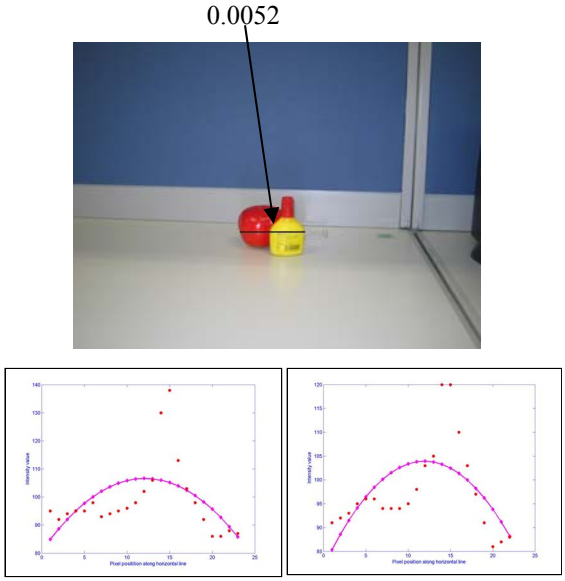

Fig. 5. Occlusion case where parts of two objects are merged into one region with intensity profile.

Now the robot can identify the target as multicolor one. Still it is difficult for the robot to segment out two objects, red one, and multicolor one containing red and yellow parts. So, the robot concentrates on the yellow part to pick the multicolor target.

Experiment 4: Robot's observation not mature without user's assistance

There are two objects, one single color and one multicolor object (Fig. 6). However, after applying the initial segmentation technique, the robot obtained four connected regions. 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.

Fig. 6 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 values of the reflectance, the robot concludes that both regions pairs R_1, R_3 and R_3, R_4 are parts of different objects, because the values of the variances are greater than 0.0060 (case 3). Regions R1 and R4 are parts of the same object, because the value of the variance is less than 0.0020 (case 1). However, the robot cannot confirm the situation about the regions R_1 and R_2, because the value of the variance is in the range of case 2. Since case 2, the robot needs further image investigation and also the user's assistance. However, the region R_1 is too small for analyzing the intensity profile along the horizontal line of the two regions R_1, R_2. So, this profile is not reliable. The robot interacts with the user in the following way,

Robot: Are red and yellow regions of the front-right side of the blue region parts of the same object?

User: Yes.

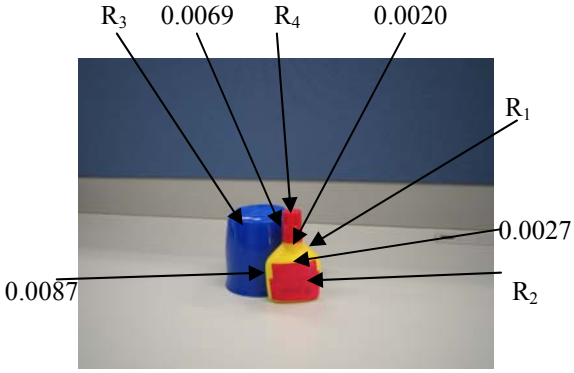

Fig. 6. Image containing single color, multicolor and occluded objects.

Then, the robot confirms that regions R_1 and R_2 are parts of the same object. Finally, the robot concludes that there are two objects; one is a multicolor object composed of regions R_1 (yellow), R_2 (red) and R_4 (red) and the other region R_3 (blue) is a single color object.

The proposed method is expected to reduce the user's verbal interaction through the analysis of image properties. We have examined our experimental results for 80 cases from this point. We used single and multicolor objects to set up the experimental scenes. Different numbers of objects and combinations were used for different cases. Table 2 shows the result. There were 335 adjacent regions, 81% of which were correctly judged by the method. The robot needed the user's assistance for 19% cases. This result confirms the usefulness of the method in terms of the reduction of user's burden.

Table 2. Experimental results.

Total Experiments	80
Single and multicolor objects used	17
Adjacent regions in experiments	335
Automatic regions merging/splitting	81%
User assistance needed for merging/ splitting	19%

However, in some more complicated cases like similar color objects occluded by each other or more than two object regions merge into one region, it is difficult to find out the border for the robot. Since this is an interactive system, it can obtain useful information from the user through interaction and it

can then use this information to tackle the situation. The robot may also use its arm to move some objects to disambiguate the scene.

8. 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. 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.

Acknowledgments

This work was supported in part 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 Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Berlin, Germany, pp.460-467, September 2002.

[2] M. Hans, B. Graf, R.D. Schraft, "Robotics home assistant Care-O-bot: Past-present-future," In Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Berlin, Germany, pp.380-385, September 2002.

[3] G. A. Berry, V. Pavlovic, and T. S. Huang, "BattleView: A multimodal HCI research application," In Proceedings of the Workshop on Perceptual User Interfaces, San Francisco, California, USA, pp. 67-70, November 1998.

[4] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward natural gesture/speech HCI: A case study of weather narration," In Proceedings of the Workshop on Perceptual User Interfaces, San Francisco, California, USA, pp. 1-6, November 1998.

[5] R. Raisamo. "A multimodal user interface for public information kiosks," In Proceedings of the Workshop on Perceptual User Interfaces, San Francisco, California, USA, pp. 7-12, November 1998.

[6] T. Takahashi, S. Nakanishi, Y. Kuno, and Y. Shirai, "Human-robot interface by verbal and nonverbal communication," In Proceedings of the International Conference on Intelligent Robots and Systems, Victoria, Canada, pp.924-929, October 1998

[7] M. Yoshizaki, Y. Kuno, and A.Nakamura, "Mutual assistance between speech and vision for human-robot interface," In Proceedings of the International Conference on Intelligent Robots and Systems, EPFL Lausanne, Switzerland, pp.1308-1313, Septem-October 2002.

[8] M. Yoshizaki, A. Nakamura, and Y. Kuno, "Vision-speech system adapting to the user and environment for service robots," In Proceedings of the International Conference on Intelligent Robots and Systems, Las Vegas, Nevada, USA, pp. 1290-1295, October 2003.

[9] R. Kurnia, M. A. Hossain, A. Nakamura, and Y. Kuno, "Object Recognition through Human-Robot Interaction by Speech," In Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Kurashiki, Okayama, Japan, pp.619-62, September 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," In Proceedings of the First International Workshop on Language Understanding and Agents for Real World Interaction, Hokkaido, Japan, 2003.

[11] T. Kawaji, K. Okada, M. Inaba, H. Inoue, "Human Robot Interaction through Integrating Visual Auditory Information with Relaxation Method," In Proceedings of the International Conference on Multisensor Fusion on Integration for Inteligent Systems, Tokyo, Japan, 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," In Proceedings of the IEEE International Workshop on Robots and Human Interactive Communication, Berlin, Germany, pp. 1082-1089, September 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," In Proceedings of the International Conference on Intelligent Robots and Systems, Sendai, Japan, pp. 2861-2866, 2004.

[14] A. Cremers, "Object Reference in Task-Oriented Keyboard Dialogues, Multomodal Human-Computer Communication: System, techniques and experiments," Springer-verlag, 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," In Proceedings of the 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," International Journal of Computer Vision, vol. 17, no. 3, pp. 219-240, 1996.

[18] M. A. Hossain, R. Kurnia, A. Nakamura, Y. Kuno, "Color objects segmentation for helper robot," In Proceedings of the International Conference on Electrical and Computer Engineering, Dhaka, Bangladesh, pp. 206-209, December 2004.