IEICE TRANS. INF. & SYST., VOL.E88–D, NO.11 NOVEMBER 2005

PAPER Special Section on Life-like Agent and its Communication

Interactive Object Recognition System for a Helper Robot Using Photometric Invariance

Md. Altab HOSSAIN[†a)], **Rahmadi KURNIA**[†], *Nonmembers*, **Akio NAKAMURA**[††], *and* **Yoshinori KUNO**[†], *Members*

SUMMARY 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 multi-color 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. Experimental results show the usefulness of the system.

key words: *segmentation, object recognition, human robot interaction, multimodal interface, interactive object recognition*

1. Introduction

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]–[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]–[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]–[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]–[8], we assumed that the scene was relatively 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 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 present 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]–[16]. These conventional systems mainly consider dialog generation at the language level, treating image features and attributes equally. Thus, as long as the certainty factors or any other values alike for features are the same, the systems act in the same way regardless what kind of features are involved. However, all image features and attributes are not the same in their characteristics. For example, humans can easily describe some features by word while not others. Some features should be treated differently depending on the existence of other features. Ours is different in that its main concern is the problems and issues of computer vision. The main purpose is to develop a vision system that can work in complex real world with the help of human interaction.

The paper is organized as follows: in Sect. 2, the basic framework of object recognition method is introduced; the problems of segmentation are presented in Sect. 3. Section 4 shows how reflectance ratio can be used to measure the compatibility of the adjacent regions. The proposed interactive

Manuscript received February 16, 2005.
Manuscript revised May 11, 2005.

[†]The authors are with the Department of Information and Computer Sciences, Saitama University, Saitama-shi, 338–8570 Japan.
[††]The author is with the Department of Machinery System Engineering, Tokyo Denki University, Tokyo, 101-8457 Japan.
a) E-mail: hossain@cv.ics.saitama-u.ac.jp

DOI: 10.1093/ietisy/e88–d.11.2500

object recognition method is given in Sect. 5. Section 6 includes experiments on real color images. The conclusion can be found in Sect. 7.

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.

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].

We give an example to show how the system works. In the case shown in Fig. 1, the user wanted the green ball in the scene. The dialog in this case was as follows.

Robot: What is the color of the target object?
User: Green. (Two objects remain as in Fig. 2 (left).)
Robot: Is the target object on the left side?
User: No.

Table 1 Features and their characteristics.

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

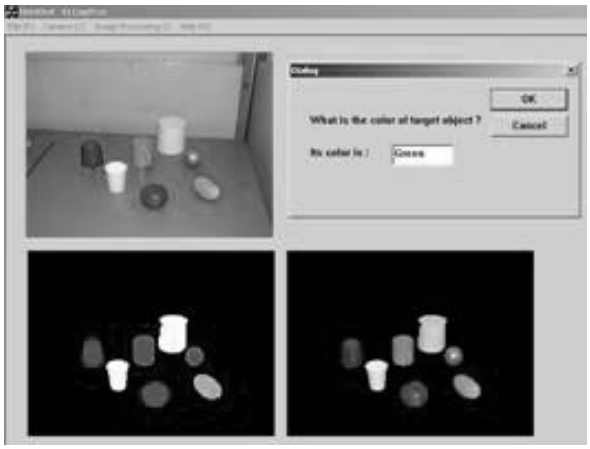


Fig. 1 Input image (top left), color segmentation (bottom left), and foreground objects (bottom right).

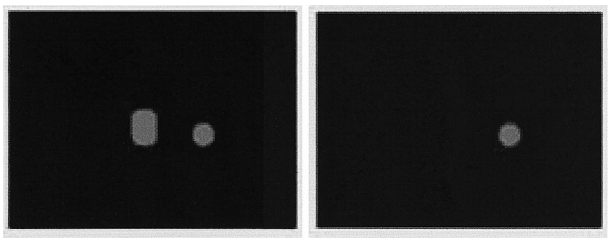


Fig. 2 Target objects recognition: after the first answer (left), final result (target object) (right).

The robot understood the object shown in Fig. 2 (right) was the target object.

Easiness for users to answer increases for what-type questions, multiple-choice questions, and yes-no type questions in this order. From the viewpoint of efficiency, the order is reverse. The robot generates questions by taking both points into account. The system proposed in this paper follows this framework.

3. Problems of Segmentation

The basic framework of the proposed system is the same as mentioned in Sect. 2. 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 preprocessing tool. This reduces the number of colors in the image and divides it into several regions. Then the Hue, Saturation and Intensity components of HSI color space are used for merging regions of similar colors. Figure 3 shows an example of image segmentation. 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, position of

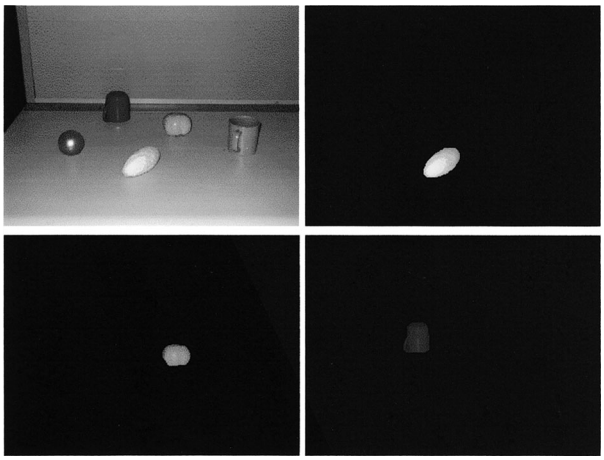


Fig. 3 Original single color objects (top left); Recognized target objects: a yellow one (top right), a red one (bottom left), and a blue one (bottom right).

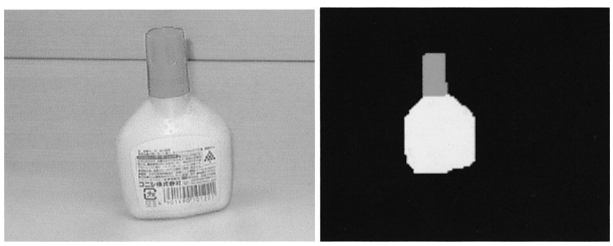


Fig. 4 Multi-color object case. Left: original image; Right: segmentation result.

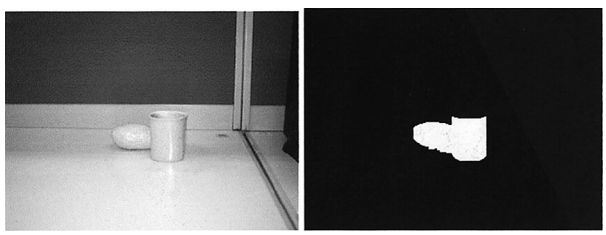


Fig. 5 Occlusion case. Left: original image; Right: segmentation result.

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 may 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. Figure 4 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. Figure 5 shows an example.

4. Reflectance Ratio to Measure the Compatibility of Adjacent Regions

The reflectance ratio, a photometric invariant, represents a physical property that is invariant to illumination and imaging parameters. Nayar and Bolle [18] 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 part of the same piecewise uniform object and have different colors, 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, 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, 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 [18].

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

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.

5. Interactive Object Recognition

We use the following steps to identify objects:

1. Apply the initial color segmentation method to an input image.
2. Use reflectance ratio to identify multicolor and occluded objects.
3. In the case of confusion, ask the user for assistance.

The system applies the initial segmentation method described in Sect. 3 to the input image to find regions in the image. The input image is first analyzed and segmented using the mean shift algorithm. The image may contain many colors and several regions. The algorithm significantly reduces the number of colors and regions. Thus, the output of the mean shift algorithm includes several regions with a fewer numbers of colors in comparison with those in the input image.

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.

Then, the system examines one-to-one correspondence between a region and an object. It checks all pairs of adjacent colored regions r_1 and r_2. 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.

As to one-to-one correspondence, the system also needs to examine whether or not each region comes from a single object. A region may include multiple objects if an object is occluded by another object of the same color. We use the segmentation method proposed in [19] to find discontinuous points on the region boundary. If there are such points, the system examines the reflectance ratio along the line connecting the opposite discontinuous points. If the reflectance ratio varies much along the line, the system judges that multiple objects form the region due to occlusion.

We performed an experiment to examine the usefulness of this measure. We measured the variance of reflectance ratio for 19 multicolor object cases and 19 occluded object cases. Figure 6 shows the result. 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 0.0061, 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

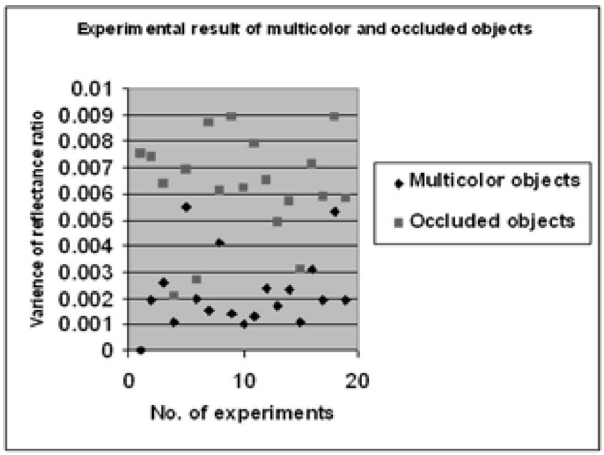


Fig. 6 Distribution of variances of reflectance ratio for multicolor and occluded objects.

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 to the user.

For simple and friendly interaction with the users, 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 different objects?" (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.

6. Experiments

6.1 Example Cases

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.

We use Pioneer 2 by ActivMEDIA as a robot (Fig. 7) in our experimental purposes. The current system does not have a robot arm. Thus, we consider it success if the robot finds and recognizes the object ordered by the user.

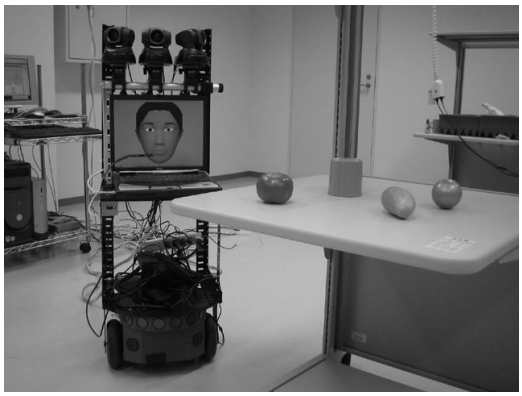


Fig. 7 Robot used in the experimental purposes.

Fig. 8 Multicolor object case.

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. 8). 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.

Experiment 2: Occluded object case (1)
In the scene shown in Fig. 9 (top), 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. 9 (bottom).

Experiment 3: Occluded object case (2)
After initial segmentation, two regions, yellow and red similar to experiment 1, are found (Fig. 10). The variance of the reflectance ratio in 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.

Robot: Are those regions parts of different objects?
User: Yes.

Based on the user response, the robot confirms that the two regions are parts of different objects. The robot comes up to know that there are two single color objects in the

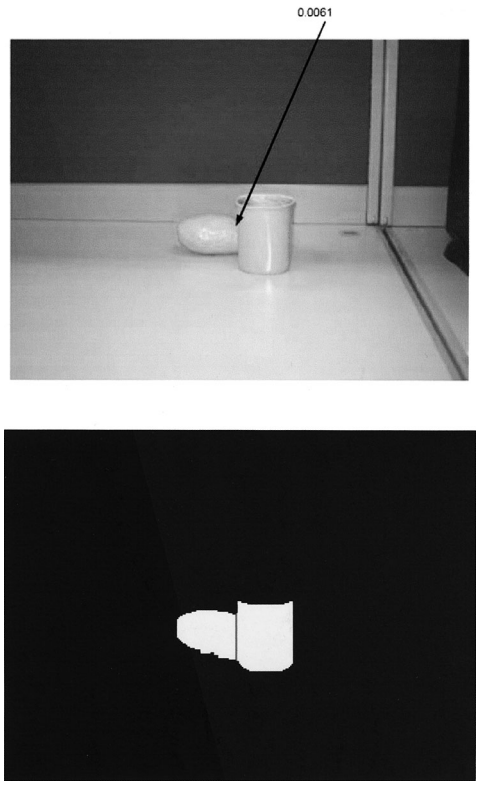


Fig. 9 Occlusion case and the final segmentation result. Only one region detected (top). Final segmentation result (bottom).

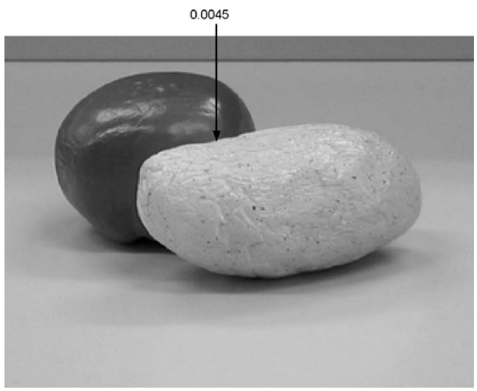


Fig. 10 Occlusion case where two adjacent different color regions are detected in the segmentation result.

scene and one is partially occluded by the other.

Experiment 4: Complex case (1)
There are two objects, one single color and one multicolor object. 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. Figure 11 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 both region pairs R_1, R_3

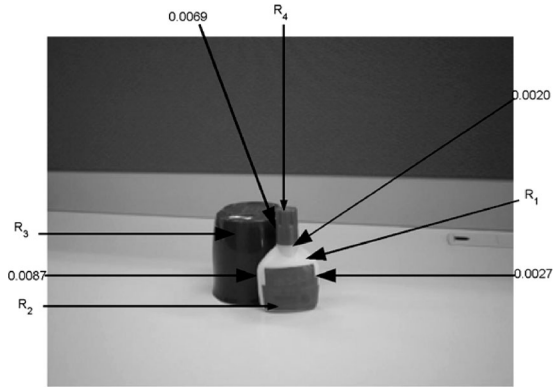


Fig. 11 Image containing single color, multicolor and occluded objects.

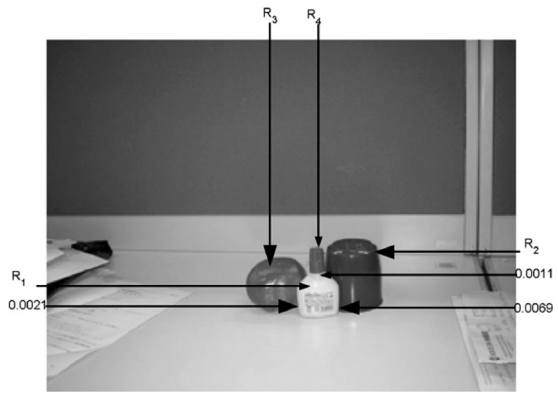


Fig. 12 Image containing single color, multicolor and occluded objects.

and R_3, R_4 are parts of different objects, because the variances are greater than 0.0061 (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 doubtful about the regions R_1 and R_2, 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: Yes.

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.

Experiment 5: Complex case (2)
In the scene shown in Fig. 12, 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 12 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. 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 is sure that regions R_1 and R_3 are parts of different objects. Finally, the robot concludes 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 two 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 objects 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.

6.2 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. 8).

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

Table 2 Comparison experiment.

Experiment No.	Number of objects (regions)	Number of required questions	
		Our method	Previous system
1	1 (2)	1	2
2	2 (1)	1	2
3	2 (2)	1	1
4	2 (4)	1	6
5	3 (4)	1	5
Other Experiments			
6	5 (7)	3	9
7	1 (3)	1	4
8	4 (6)	2	7
9	3 (5)	1	5
10	2 (2)	3	4

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 2 shows the results for the experimental cases 1–5 described in Sect. 6.1. It also shows the results for other five cases. The results confirm that our current system needs a smaller number of questions than the previous system.

6.3 Discussion

We can expect that interactive vision systems can work in various conditions. However, if the situation is complex, they may need a great number of interactions and cannot be effective and user-friendly. In this paper, we show that we can improve the performance of an interactive vision system if we introduce a decision process with reasonable ability based on some image properties. We would like to consider 'reasonable ability' here.

The decision by our system is right or wrong. In addition, we allow no-decision judgment to our system. We denote the right decision probability by Pr, the wrong one by Pw, and no-decision one by Pn (Pr + Pw + Pn = 1). Given an image, we assume that one of the three cases happen according to these probabilities. If the decision is right, the system can reduce the number of interactions (questions) by Cr. In the no-decision case, the performance of the system is the same as that of the system without the decision capability. We consider this as the base performance without any advantage or disadvantage. (We need additional computing cost if we use the decision process. We assume that we can neglect it because we can have much computing power.) If the decision is wrong, the system needs additional interaction with the user to recover the error. We denote the number of interactions necessary for error recovery by Cw. Thus, whether or not the system is effective and worthy to be used depends on the following value Tc:

$$\mathrm{Tc} = \mathrm{Pr} * \mathrm{Cr} - \mathrm{Pw} * \mathrm{Cw}$$

The actual values of Cr and Cw depend much on the

situation. Thus, it is a good design policy to make Pr much larger than Pw. Generally, we can do this by setting decision parameters to allow large Pn. However, this means that Pr becomes small although Pr/Pw is large. To sum up, the effectiveness of introducing a new decision module depends on whether or not it can achieve Pr ≫ Pw while keeping Pn reasonably small.

In this paper, we have proposed to introduce a decision method based on the reflectance ratio into the interactive system. As shown in Table 2, the system can reduce the number of questions when the decision is right. The system did not give wrong decisions in these experiments. The current system asks the user for confirmation. Thus, even if it makes a mistake, the system can know the mistake soon by the user's negative response. In such cases the system goes back to the state without the decision module. Thus, Cw is considered not large. In the current implementation, we have determined the decision parameters based on the experimental results shown in Fig. 6 so that there are no decision errors among the measured cases. Pn is about 0.4 for the measured cases. Although we cannot tell the actual value of Pn, we can say from the experimental results that the current system achieves Pr ≫ Pw while keeping Pn reasonably small.

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. This is left for future work.

7. 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.

Interactive systems such as the one proposed here are good in that they can be expected to work under various conditions owing to human assistance. However, if they need too much human assistance, they cannot be accepted. This paper proposes to use photometric invariance to reduce segmentation failure cases. Although the system cannot recover from all segmentation failures, this kind of improvement can make the system more acceptable. When we add a new function like this, we need to design the function so that the user can easily help the system even if the function cannot work well. We would like to add such functions one by one to realize an actual working system that can be accepted by various people.

Acknowledgment

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," Proc. IEEE International Workshop on Robot and Human Interactive Communication, pp.460–467, Berlin, Germany, Sept. 2002.

[2] M. Hans, B. Graf, and R.D. Schraft, "Robotics home assistant care-o-bot: Past-present-future," Proc. IEEE International Workshop on Robot and Human Interactive Communication, pp.380–385, Berlin, Germany, Sept. 2002.

[3] G.A. Berry, V. Pavlovic, and T.S. Huang, "BattleView: A multimodal HCI research application," Proc. Workshop on Perceptual User Interfaces, pp.67–70, San Francisco, California, USA, Nov. 1998.

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

[5] R. Raisamo, "A multimodal user interface for public information kiosks," Workshop on Perceptual User Interfaces, pp.7–12, San Francisco, California, USA, Nov. 1998.

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

[7] M. Yoshizaki, Y. Kuno, and A. Nakamura, "Mutual assistance between speech and vision for human-robot interface," Proc. International Conference on Intelligent Robots and Systems, EPFL, pp.1308–1313, Lausanne, Switzerland, Sept.–Oct. 2002.

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

[9] R. Kurnia, M.A. Hossain, A. Nakamura, and Y. Kuno, "Object recognition through human-robot interaction by speech," Proc. IEEE International Workshop on Robot and Human Interactive Communication, pp.619–624, Kurashiki, Okayama, Japan, Sept. 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, pp.16–23, Hokkaido, Japan, 2003.

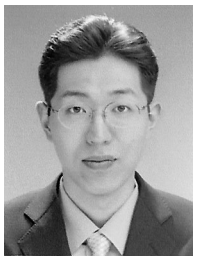
[11] T. Kawaji, K. Okada, M. Inaba, and H. Inoue, "Human robot interaction through integrating visual auditory information with relaxation method," Proc. International Conference on Multisensor Fusion on Integration for Intelligent Systems, pp.323–328, Tokyo, Japan, 2003.

[12] P. McGuire, J. Fritsch, J.J. Steil, F. Roothling, G.A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human ma-

chine communication for instruction robot grasping tasks," Proc. International Workshop on Robots and Human Interactive Communication, pp.1082–1089, Berlin, Germany, 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, pp.2861–2866, Sendai, Japan, 2004.
[14] A. Cremers, Object Reference in Task-Oriented Keyboard Dialogues, Multomodal Human-Computer Communication: System, Techniques and Experiments, pp.279–293, Springer-Verlag, 1998.
[15] T. Winograd, Understanding Natural Language, Academic Press, New York, 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, pp.147–163, Greece, Sept. 1999.
[17] Z.M. Hanafiah, C. Yamazaki, A. Nakamura, and Y. Kuno, "Human-robot speech interface understanding inexplicit utterances using vision," Proc. Conference on Human Factors in Computing Systems, pp.1321–1324, Vienna, Austria, April 2004.
[18] S.K. Nayar and R.M. Bolle, "Reflectance based object recognition," Int. J. Comput. Vis., vol.17, no.3, pp.219–240, 1996.
[19] D. Trytten and M. Tuceryan, "Segmentation and grouping of object boundaries using energy minimization," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.730–731, Hawaii, USA, June 1991.

Md. Altab Hossain received the B.Sc. and M.Sc. degrees in Computer Science and Technology from the University of Rajshahi, Rajshahi, Bangladesh in 1996 and 1997 respectively. He is a lecturer of the same department and university from 1st September 2001 to continuing. His research interest includes Robotics, Human Computer Interaction, Computer Vision, and Object Recognition. Specifically, he is now researching on object recognition for service robots. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).

Rahmadi Kurnia received the B.Sc. and M.Sc. degrees in Telecommunication Engineering from the University of Indonesia, in 1995 and 1998 respectively. From 1st October 1998, he is a lecturer of the department of Electrical Engineering, Andalas University, West Sumatra, Indonesia. Since October 2001, he has been a Ph.D. student in the graduate school of Science and Engineering, Saitama University, Japan. His research interest includes Robotics, Human Computer Interaction, Computer Vision, and Object Recognition. Specifically, he is now researching on object recognition for service robots.



Akio Nakamura received the M. Eng. and Dr. Eng. degrees from the University of Tokyo in 1998 and in 2001 respectively. In 2001, he became a research associate of the Department of Information and Computer Sciences, Saitama University. Since 2005, he has been an associate professor in the Department of Machinery System Engineering, Tokyo Denki University. He is engaged in education of computer science engineering and research on robotics, especially computer vision and man-machine interface systems. He is a member of the RSJ, and IEEE Robotics and Automation Society.

Yoshinori Kuno received the B.S. degree, the M.S. degree, and the Ph.D. degree in 1977, 1979, and 1982, respectively, all in electrical and electronics engineering from the University of Tokyo. In 1982, he joined Toshiba Corporation. From 1987 to 1988, he was a Visiting Scientist at Carnegie Mellon University. In 1993, he moved to Osaka University as an associate professor in the Department of Computer-Controlled Mechanical Systems. Since 2000, he has been a professor in the Department of Information and Computer Sciences, Saitama University.